Yan Li
Hualiang Shi *Editors*

Advanced Driver Assistance Systems and Autonomous Vehicles

From Fundamentals to Applications

Springer

Advanced Driver Assistance Systems and Autonomous Vehicles

Yan Li · Hualiang Shi
Editors

Advanced Driver Assistance Systems and Autonomous Vehicles

From Fundamentals to Applications

Springer

Editors
Yan Li
Intel Corporation
Chandler, AZ, USA

Hualiang Shi
Lyft
Palo Alto, CA, USA

ISBN 978-981-19-5052-0 ISBN 978-981-19-5053-7 (eBook)
https://doi.org/10.1007/978-981-19-5053-7

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Contents

Introduction

Hualiang Shi and Yan Li

Abstract Advanced driver-assistance systems (ADAS) and autonomous vehicles (AV) have the potential to reshape transportation, by reducing risky driver behaviors, traffic jams, carbon emission, and cost of transportation, as well as improving road safety, independence of seniors and people with disabilities, and human productivity. Although advanced driver-assistance systems and autonomous driving functions are promising, there are many challenges, including new technologies, requalification of non-auto-grade components, and new mission profiles for existing auto-grade components. The chapter reviews various challenges in ADAS and AV, as well as contents of other chapters in the book.

1 Reshape the Future of Transportation

Figure 1 shows the daily travel hour per licensed driver, reported by the U.S. DOT Volpe Center. On average, American drivers spent just under one hour daily behind the wheel [1]. Why do people waste time driving? Why don't they use this amount of time to recharge themselves (i.e., watch movies, read books, have a nap), bond themselves with families or friends (i.e., chat by facetime, play game), or complete work (i.e., write emails, analyze data, attend Zoom meetings)?

Figure 2 shows the number of fatal crashes in the U.S., reported by NHTSA [2]. Part of the crashes was induced by human's risky driving behaviors, such as drunk driving, drug-impaired driving, distracted driving, and speeding. If a human driver is pulled out of the driver seat, will these risky driving behaviors be reduced?

Greenhouse gases generated by human activities are changing the global climate. Many countries and organizations are making plans to achieve net-zero emission by 2050. Figure 3 shows the total greenhouse gas emission in the U.S., reported by U.S.

H. Shi (✉)
Palo Alto, San Jose, USA
e-mail: hualiang.shi@gmail.com

Y. Li
Intel Corporation, Chandler, Arizona, USA
e-mail: yan.a.li@intel.com

Y. Li and H. Shi (eds.), *Advanced Driver Assistance Systems and Autonomous Vehicles*,
https://doi.org/10.1007/978-981-19-5053-7_1

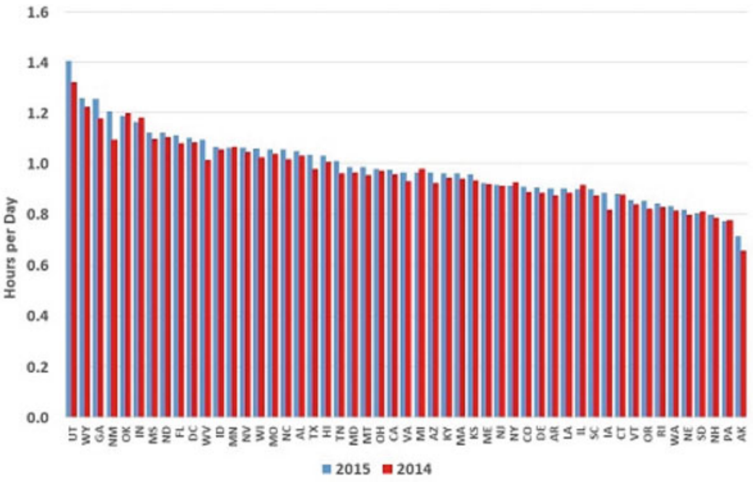
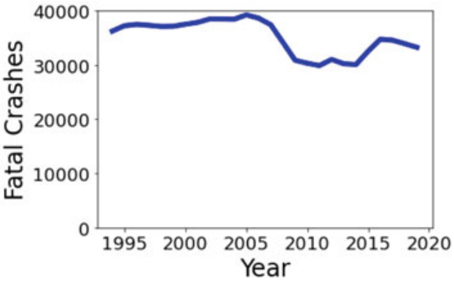


Fig. 1 Daily auto-travel per licensed driver. (Courtesy of U.S. DOT Volpe Center [1])

Fig. 2 Motor vehicle traffic fatal crashes from NHTSA



EPA [3]. As shown in Fig. 4, the transportation sector has the largest contribution to the U.S. greenhouse gas emission in 2019. The amount of greenhouse gases generated by cars depends on many factors [4, 5], such as driving behavior, traffic jams, efficiency of route, number of cars, weight of cars, type of cars, and health of cars. AAA reported that "More than 86% of U.S. households have at least one car for every driver in the home and 28% report having more cars than drivers." [6]. Will the carbon emission be reduced if the number of cars is reduced, the route is optimized, or driving behavior is smoother or eco-friendly (less braking and re-acceleration)?

Advanced driver-assistance systems (ADAS) and autonomous driving functions have the potential to reshape transportation and address the issues above. They can reduce risky driver behaviors, traffic jams, carbon emission, cost of transportation, and improve road safety, independence of seniors and people with disabilities, and human productivity.

In the past decades, advanced driver-assistance systems and autonomous driving functions have generated extensive research and development interests in academia

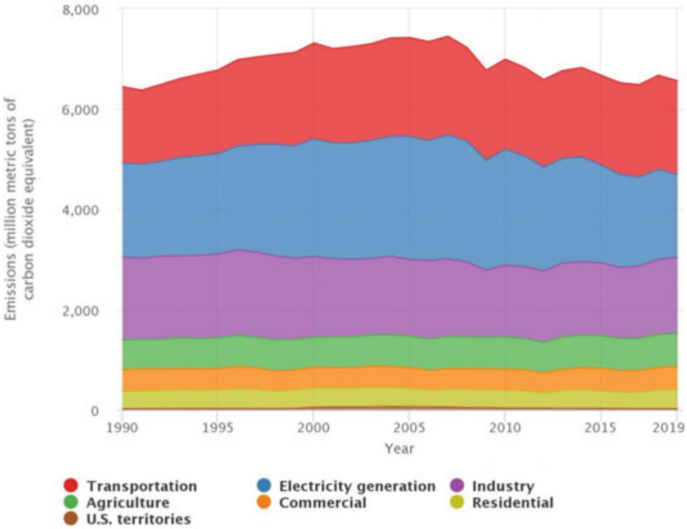


Fig. 3 U.S. greenhouse gas emission by Economic sector from U.S. EPA [3]

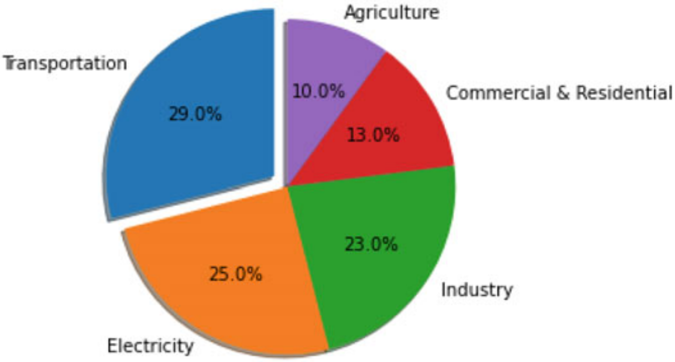


Fig. 4 U.S. greenhouse gas emission in 2019 [3]

and industry [7–30]. Some companies worked with original equipment manufacturer (OEM) to develop advanced driver-assistance systems and autonomous driving functions. Because this approach is modifying or optimizing existing vehicle platforms from OEMs, the technical challenge might be less, but the performance and reliability might be also limited by vehicle mass, power, and components which are qualified for traditional vehicles. Other companies develop new vehicle platforms with advanced driver-assistance systems and autonomous driving functions. Because the vehicle platform is built from scratch, the technical challenge is more, but the performance and reliability can also be improved by system and component optimizations. Since 2016, Lyft has been working hard to build a first-of-its-kind self-driving rideshare

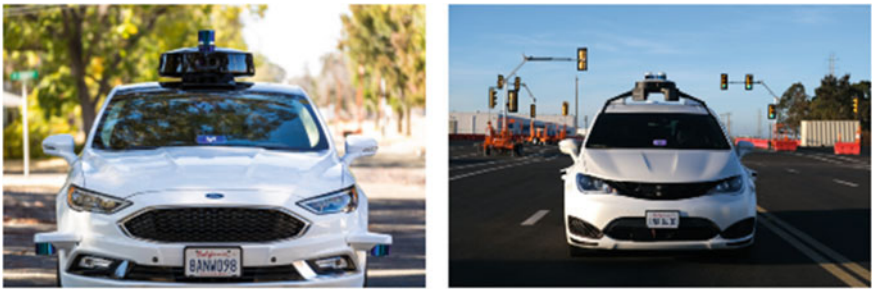


Fig. 5 Self-driving car developed by Lyft

program with some of the world's top autonomous partners [31]. By February 11, 2020, this largest public self-driving commercial platform in the U.S. has more than 100,000 paid rider trips [32]. In parallel, Lyft also developed four generations of self-driving vehicle platforms for employee pilots internally. Figure 5 presented two self-driving cars developed by the Lyft level-5 self-driving division, which were designed at the top of Ford Fusion and FCA Pacifica.

2 Challenges

Although advanced driver-assistance systems and autonomous driving functions are promising, there are many challenges.

2.1 New Technologies

Table 1 shows the six-level driving automation defined by SAE J3016 [33]. "DDT" means dynamic driving task. "ODD" means operational design domain. "OEDR" means object and event detection and response.

With the increase of the level of driving automation, more advanced technologies are needed, such as the perception subsystem, planning subsystem, and control subsystem.

The perception subsystem uses sensors to detect objects outside the ego-vehicle and localize ego-vehicles in the environment. Typical sensors include camera, GPS, IMU, Lidar, Radar, and others. Based on sensor data and various machine learning algorithms, objects in the environment will be detected, classified, and tracked. Camera, Lidar, and Radar will be discussed in detail in the following chapters. Table 2 is an overview about these three types of sensors, made by Schlager et al. [34]. Due to the cons and pros behind each type of sensor, it is not uncommon that various

Table 1 Levels of driving automation from SAE J3016 [33]

Level		DDT sustained lateral and longitudinal vehicle motion control	DDT OEDR	DDT fallback	ODD
0	**No driving automation** The performance by the driver of the entire DDT, even when enhanced by active safety systems	Driver	Driver	Driver	N/A
1	**Driver assistance** The sustained and ODD-specific execution by a driving automation system of either the lateral or the longitudinal vehicle motion control subtask of the DDT (but not both simultaneously) with the expectation that the driver performs the remainder of the DDT	Driver and system	Driver	Driver	Limited
2	**Partial driving automation** The sustained and ODD-specific execution by a driving automation system of both the lateral and longitudinal vehicle motion control subtasks of the DDT with the expectation that the driver completes the OEDR subtask and supervises the driving automation system	System	Driver	Driver	Limited

(continued)

Table 1 (continued)

Level		DDT sustained lateral and longitudinal vehicle motion control	DDT OEDR	DDT fallback	ODD
3	**Conditional driving automation** The sustained and ODD-specific performance by an ADS of the entire DDT with the expectation that the DDT fallback-ready user is receptive to ADS- issued requests to intervene, as well as to DDT performance-relevant system failures in other vehicle systems, and will respond appropriately	System	System	Fallback ready user (becomes the driver during fallback)	Limited
4	**High driving automation** The sustained and ODD-specific performance by an ADS of the entire DDT and DDT fallback without any expectation that a user will need to intervene	System	System	System	Limited
5	**Full driving automation** The sustained and unconditional (i.e., not ODD-specific) performance by an ADS of the entire DDT and DDT fallback without any expectation that a user will need to intervene	System	System	System	Unlimited

Table 2 Overview of sensor properties of Radar, Lidar, and camera from Schlager et al. [34]

	Radar	Lidar	Camera
Active/passive sensor	Active	Active	Passive
Spectral components	Millimeter waves	Near-infrared light	Visible light and near-infrared light
Distance capturing	Time of flight (TOF)	TOF	Stereo camera
Velocity capturing	Doppler principle	Track points over time	Optical flow
Strengths	Works well under different light and weather conditions	Very accurate distance measurements	Dense measurements, object classification works well
Weaknesses	Sensible to target reflectivity, bad angular resolution	Sparse measurement, affected by diverse weather conditions	Affected by diverse weather conditions

types of sensors are combined together for perception. The position and quantity of sensors depend on both vehicle dimension and sensor capability.

The information from the perception subsystem is relayed into the planning subsystem. Planning subsystem will generate a series of projected waypoints. Each waypoint has a specific target location and velocity.

Based on the information from the planning subsystem, the control subsystem will send acceleration messages, braking messages, or steering messages to vehicles.

These autonomy subsystems require strong computing capability by using CPUs and GPUs. Different architectures coexist in the market. Some companies use centralized architecture, while others use distributed architecture. Centralized architecture means data from various sensors are sent to a single or central computing/storage location. Distributed architecture means that an individual component has its own computing/storage capability. Due to the massive computing activities, thermal management is crucial for advanced driver-assistance systems and autonomous driving functions. At the top of HVAC air cooling, a liquid cooling subsystem is introduced. Cold plates are specially designed and attached to CPUs, GPUs, and power systems. Liquid coolant circulates through cold plates and carries heat away from CPUs, GPUs, and power systems. These cold plates are custom-designed and use new brazing material and process. Several companies encountered similar technical issues, such as cold plate buckling or deformation, breezing residue, or particles. Figure 6a shows the cold plate buckling reported by Shi et al. [35], and Fig. 6b shows the cold plate deformation reported by Chen [36]. Cold plate buckling/deformation can degrade thermal performance and induce potential electrical short and fire hazard. Breezing residue can clog the radiator and pull down the coolant pump or fan. These issues need massive design optimization, material optimization and process optimization and will be discussed in detail in the following chapters.

Corner cases in the field are potential concern for the implementation of autonomous vehicles. To mitigate this long tail effect, tons of autonomous vehicles have been dispatched for road testing. The perception data collected on the road are used for offline machine learning model training. Due to the large amount of

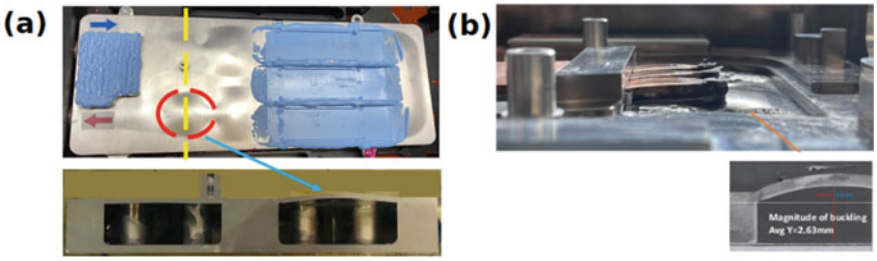
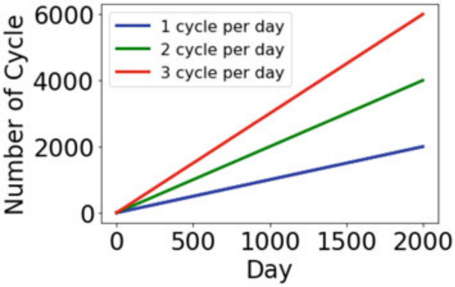


Fig. 6 **a** Cold plate buckling reported by Shi et al. [35] and **b** cold plate deformation reported by Chen [36]

Fig. 7 SSD mating and unmating cycling



perception data and the speed limitation of data transfer over the air, many companies are using solid-state drives (SSD) to store the perception data during road tests. During lunchtime or dinner time, operators unplug SSDs from the car and transfer the data to a data center which uses hard disk drives (HDD). Once data are transferred, SSDs are inserted back to the car. During the SSD mating and unmating (insertion and ejection) inside the car and inside the data center, the metal surface finishes of the socket on PCB boards and metal pads on SSDs will be scratched and might wear out eventually. If this happens, data loss might occur. Figure 7 presents the number of mating/unmating cycles as function of number of days and cycling frequency. This kind of cycling is far beyond SSD and PCB manufacturing specs and needs gold plating thickness optimization. NAND SSD and HDD will be discussed in detail in the following chapters.

2.2 Requalification of Non-Auto-Grade Components

To save time to market, some non-auto-grade components will be used for advanced driver-assistance systems and autonomous driving functions. Jung et al. [37] reviewed the role of DRAM memory. Due to the design limitations of these non-auto-grade components, many challenges surfaced during the requalification process.

Figure 8 is a DC-to-DC power supply unit failure reported by Shi et al. [35]. Inside the computing subsystem, DC-to-DC power supply units are used to convert 48 V input to 12 V output. This off-the-shelf (OTS) DC-to-DC power supply unit is a well-known server grade component. As indicated in Fig. 8a, after a vibration requalification test, this power supply unit failed at 48 V input side during power-on. Figure 8b shows the leading hypothesis—during vibration, the protruded pins on PCB and metal clips on chassis abrade the insulation layer between PCB board and metal chassis; when the thickness of insulation layer is reduced enough, dielectric breakdown happens due to the voltage between PCB and metal chassis. This kind of phenomena would not be an issue for servers due to the lack of high vibration.

Figure 9 is a coolant leakage failure reported by Shi et al. [35]. When several GPUs are bundled in parallel to enhance the computing capability, manifolds can be used to integrate individual water blocks together and simplify the coolant loop design. Manifold is sitting at the top of the water blocks. EPDM gasket is sandwiched between manifold and water blocks. When the screw is tightened, the EPDM gasket is compressed which generates a repulsive force on the manifold/screw. This off-the-shelf (OTS) water block/EPDM gasket/manifold system is a well-known consumer-grade product. As indicated in Fig. 9a, during the temperature cycle requalification test, coolant leakage surrounding the joints between manifold and water blocks was observed. Based on the fishbone diagram, the leading hypothesis is that the EPDM gasket had compression set and permanent plastic deformation at high temperature. This kind of phenomena would not be an issue for consumer electronics due to the lower ambient temperature. This failure mode will be discussed in detail in the following chapters.

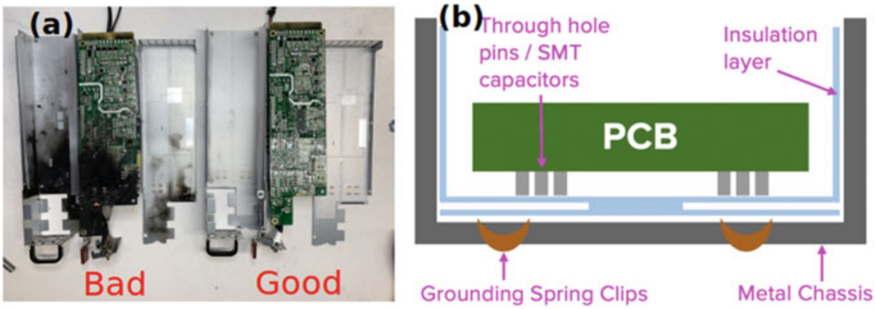


Fig. 8 Power supply unit failure

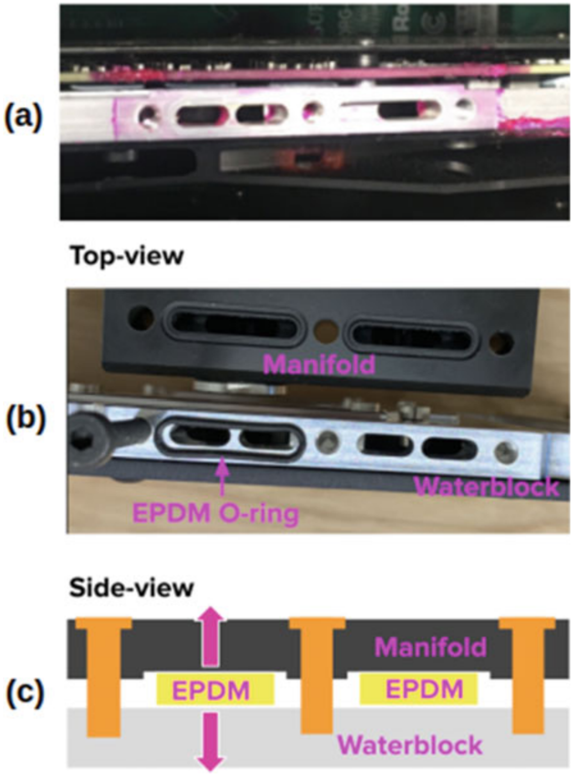


Fig. 9 **a** Coolant leakage, **b** top-view, **c** side-view

2.3 *New Mission Profiles for Existing Auto-Grade Components*

Use cases of autonomous vehicles can be very different with traditional cars. As discussed in Subsection 1.1, American drivers drive just under one hour daily on average. Currently, many automobile original equipment manufacturing (OEM) and component vendors are using 10–15 years, 10,000–15,000 h, and 100,000–150,000 miles as target lifetimes for component and system qualifications [38, 39]. However, for autonomous vehicles such as robo-taxi, the daily operation hour will be very long. As indicated in Fig. 10a, with the increase of daily operation hours, it will take less number of years for robo-taxi to reach 10,000 operation hours. For example, if a robo-taxi drives 11 h per day, it will take ~ 2.5 years to achieve 10,000 operation hours. By assuming vehicle speed 35 mile per hour (MPH), Fig. 10b shows that with the increase of daily operation hour, it will take less number of years for robo-taxi to reach 100,000 miles. For example, if a robo-taxi drives 11 h per day, it will take ~0.7 year to achieve 100,000 miles. Beyond 10,000 h or 100,000 miles, components or vehicles might have high probability for failure. This simple math tells us that the lifetime of a robo-taxi will be much shorter than a traditional car from the viewpoint

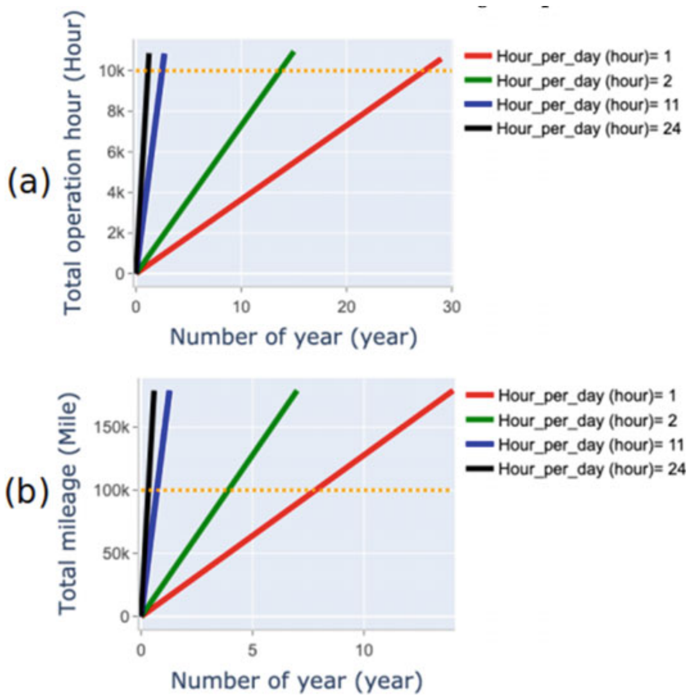


Fig. 10 Mission profile

of "number of years." This conclusion is consistent with what Ford presented before, "The thing that worries me least in this world is decreasing demand for cars. We will exhaust and crush a car every four years in this business" [40]. Test plan customization based on mission profile will be discussed in detail in the following chapters.

3 Overview of Chapters

Most of the publications about advanced driver-assistance systems and autonomous driving functions are related to software algorithm and system architecture. A very limited number of publications are discussing the design, fabrication, testing, and reliability analysis related to hardware subsystem, module, and component [41]. In this book, we try to deep dive into some of these areas by using the state-of-the-art information. The following is a brief summary about other chapters.

Chapter "The applications of Artificial Intelligence in Advanced Driver-Assistance Systems and Autonomous Vehicles" reviews various AI methods (supervised learning, unsupervised learning, reinforcement learning, deep learning), safety standards, and methodologies, challenges (edge cases and heavy tail distribution), publicly available training and testing datasets, open-source simulators, infrastructures for AI systems, validation, testing, and implementation.

Chapter "Computing Technology in Autonomous Vehicle" reviews ADAS compute technology (levels of autonomous driving, platform for autonomous driving, perception and localization, prediction, planning, and control, functional safety), advanced centralized computing system (architectures, environment perception sensors, system on chip, memory, storage, network, connectors, real-time operating system, management, failure detection and diagnostics, security, and middleware), electrical test and reliability validation (automotive-level electrical functional tests, reliability validation tests based on mission profile, EMC/ESD validation), challenges to safe deployment at scale (artificial intelligence: perception and prediction, power consumption, thermal management, manufacturing, assembly and quality control, size and cost, quality and reliability, security and safety).

Chapter "Overview of Packaging Technologies and Cooling Solutions in ADAS Market" reviews various packaging technologies employed to meet the demands of an automotive life and the associated quality and reliability requirements, as well as the system packaging and thermal management strategies.

Chapter "Flash Memory and NAND" reviews the fundamentals of Flash memory, and NAND in particular. It discusses the basic floating gate memory cell structure, the historical evolution of NAND Flash, the basic operations (Read, Write, and Program), the memory architecture, manufacturing processes, the unique technology and design challenges of 3D NAND, various reliability issues (write error, disturb error, and data retention error), and the 3D NAND future outlook.

Chapter "Interconnect" reviews various interconnects and solder joint technologies for applications under the hood, including nanoparticle sintering method, transient liquid-phase bonding technology, low melting point solders, Cu–Cu bonding by surface-activated bonding process, Cu–Cu bonding by chemical pretreatment, Cu–Cu bonding by thermal compressive bonding, low-temperature Cu–Cu bonding by (111) nanotwinned structure, and low-temperature Cu-to-Cu bonding with Ag passivation under atmosphere.

Chapter "Cameras in ADAS/AD Vehicles" reviews camera system hardware (image sensor, lens, optical filter, emitter, EEPROM, OTP, SerDes, image signal processor), image processing pipeline (black-level subtraction, decompanding, defective pixel correction, noise reduction, lens shading correction, chromatic aberration correction, demosaic, autowhite balance, autoexposure, color correction matrix, tone mapping, gamma correction, color space conversion, sharpening, color enhancement, chroma noise reduction, distortion correction, temporal noise filter, video codec), calibration, and camera product development cycle. Various architectures, design concepts, defect types, and reliability concerns are also covered.

Chapter "LiDAR technology" reviews the state of the art of Lidar sensors for autonomous driving or ADAS applications. The important metrics for Lidar sensor performance are discussed, including detection range, field of view (FOV), angular resolution, frame rate, and eye safety. Distance calculation methods are covered, including Time of Flight (TOF) and frequency-modulated continuous wave (FMCW)

Lidars. Different Lidar mapping methods are presented, including mechanical spinning scanner, Opto-mechanical scanning, MEMS scanning, Flash, optical-phased array (OPA).

Chapter "Radar Technology" reviews Radar architecture (RF transceiver, antenna, signal processing, data processing, and system management), Radar categories (monostatic Radars, bistatic or multistatic Radars, RX-MIMO, Radar networks, digital beamforming), waveform design (pulse Radar, pulse-coded Radar, FMCW Radar), link budget analysis for FMCW Radar, and challenges and solutions (interference, under- and over-clustering, classification, lack of resolution, data fusion, and Radar integration). Simplified examples are used to help readers understand the topics.

Chapter "Electrochemical Power Systems for Advanced Driver-Assistance Vehicles" reviews the status and challenges of electrochemical batteries, fuel cells, and capacitors. The type, chemistry, structure, and process of battery cells are discussed in detail. Battery management systems are also covered. Failure mode and effect analysis are provided as reference. Various industry standards for battery testing (IEEE 1625, IEEE 1725, SAE J2464, UL 1642, UL 2054, UL2271, UL 2580, and IEC 62,660–2) are compared.

Chapter "In-Vehicle Display Technology" reviews various in-vehicle display technologies and architectures (LCD, TFT LCD, OLED, LED, Mini-/Micro-LED, head-up display, flexible and free-form, touchscreen), requirements (optical performance, appearance, integration, fabrication, performance characterization, Mura, defect, inspection and Demura, visibility in bright light and complete darkness, improvement of image and touch quality, reliability, durability, functional safety), challenges (specification, functionality, quality, reliability, validation, EMC/EMI, ESD, and high transient voltage), and common failure modes and effects case studies (FOS spotlighting failure, BLU film buckling/waving/wrinkle failure, metal oxide TFT panel-level VGH and VGL, LCD panel UV aging, polarizer bleaching failure, freefall object impact test and LCD glass crack failure, LED luminance degradation).

Chapter "Disk Drive for Data Center Storage" reviews the current status and challenges of hard disk drives (HDD) used by data centers. Components and materials inside the hard disk drive are explained in detail. Various solutions to achieve higher areal data density are summarized, including microwave-assisted magnetic recording and heat-assisted magnetic recording. Both working principles and reliability issues are discussed.

Chapter "Role and Responsibility of Hardware Reliability Engineer" covers several key roles and responsibilities of hardware reliability engineers across product life cycle, including risk assessment methodologies (failure mode and effect analysis, fault tree analysis, and stress-strength analysis), accelerated life testing and highly accelerated life testing, reliability statistics (sample size calculation, life distribution analysis by linear least square regression and maximum likelihood estimation, confidence interval calculation, hypothesis tests for mean and variance), failure analysis and corrective/preventive actions, system reliability metrics, reliability block diagram methods, and repairable system. Cameras, cold plates, dash mount audio device, LED display, Lidar bracket, magnetic sensor, network and multimedia PCB boards, power supplies, Radar, and waterblock are used to illustrate these ideas.

Chapter "Failure Analysis in Advanced Driver-Assistance Systems" reviews failure analysis flow of system/board/package/device, and various electrical failure analysis techniques, physical failure analysis approaches, material analysis methods, and non-destructive imaging techniques. It covers I-V curve tracing, time-domain reflectometry (TDR), electro-optic terahertz pulse reflectometry (EOTPR), lock-in thermography (LIT), magnetic field imaging (MFI), magnetic current imaging (MCI), infrared emission microscope (IREM), photon emission microscopy (PEM), thermally induced voltage alteration (TIVA), Seebeck effect imaging (SEI), nanoprobing and E-beam imaging, E-beam probing, mechanical polishing, laser ablation techniques, plasma-FIB, broad-beam ion milling, scanning electron microscopy (SEM), transmission electron microscopy (TEM), energy disersive X-ray spectroscopy (EDX), Fourier transform infrared spectroscopy (FTIR), atomic force microscopy-based infrared spectroscopy (AFM-IR), X-ray photoelectron spectroscopy (XPS), Time-of-Flight secondary ion mass spectrometry (TOF–SIMS), electron backscatter diffraction (EBSD), optical and infrared (IR) imaging, scanning acoustic microscopy (SAM), 2D X-ray radiography, and 3D X-ray computed tomography (CT).

Chapter "Corrosion Mechanisms of Copper and Gold Ball Bonds in Semiconductor Packages" reviews the corrosion mechanism that causes reliability failures of Cu and Au ball bonds, by unifying approaches based on microstructure characterization and electrochemical investigation.

4 Summary

In this chapter, the advantages and challenges of advanced driver-assistance systems (ADAS) and autonomous driving functions are discussed briefly. The contents of other book chapters are also reviewed. The topics covered in this book can benefit students, researchers, and engineers who are working on automobiles, robotics, augmented reality, virtual reality, data center, desktop, laptop, tablet, smart phone, smartwatch, etc. Due to the time limit, many topics cannot be covered by this edition of the book. We hope more topics can be added in the future edition of this book.

Acknowledgements The authors would like to thank Fen Chen from Cruise to share his experimental data. Special thanks are to Angel Shi and Charlotte Shi for the grammar check.

References

1. How much time do americans spend behind the wheel?, U. S. Department of Transportation (DOT), December 11, 2017.
2. https://www-fars.nhtsa.dot.gov/Main/index.aspx, U.S. Department of Transportation, Federal Highway Administration.
3. https://www.epa.gov/ghgemissions/inventory-us-greenhouse-gas-emissions-and-sinks, United States Environmental Protection Agency.
4. A. Brown, B. Repac, J. Gonder, Autonomous Vehicles Have a Wide Range of Possible Energy Impacts, Workshop on Road Vehicle Automation, July 16, 2013.
5. M. Massar, I. Reza, S. Masiur Rahman, S. Muhammad Habib Abdullah, A. Jamal, and F. Saleh Al-Ismail, Impacts of Autonomous Vehicles on Greenhouse Gas Emissions—Positive or Negative?, Int. J. Environ. Res. Public Health 2021, 18, 5567.
6. Americans Spend an Average of 17,600 Minutes Driving Each Year, A. Gross, AAA Public Relations, September 8, 2016.
7. H. Waschl, I. Kolmanovsky, F. Willems, Control Strategies for Advanced Driver Assistance Systems and Autonomous Driving Functions, Springer, ISBN: 978–3–319–91569–2, 2019.
8. M. Hasenjäger; H.Wersing, Personalization in advanced driver assistance systems and autonomous vehicles: A review, 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC).
9. J. Piao, M. McDonald, Advanced Driver Assistance Systems from Autonomous to Cooperative Approach, Transport Reviews, Volume 28, 2008 - Issue 5.
10. V. Kumar Kukkala, J. Tunnell, S. Pasricha, T. Bradley, Advanced Driver-Assistance Systems: A Path Toward Autonomous Vehicles, IEEE Consumer Electronics Magazine,Volume 7, Issue 5, Sept. 2018.
11. R. Erik Haas, S. Bhattacharjee, D. P. F. Möller, Advanced Driver Assistance Systems, Smart Technologies pp 345–371, 2019.
12. Lyft self-driving safety report, Lyft, 2020.
13. Waymo safety report, Waymo, September, 2020.
14. Self-driving safety report, General Motor, 2018.
15. A matter of trust, Ford's approach to developing self-driving vehicles, Ford, August 16, 2018.
16. Delivering safety: Nuro's approach, Nuro, September 13, 2018.
17. Safety innovation at Zoox, Zoox, December 12, 2018.
18. The new era of mobility, Aurora, April 29, 2019.
19. Self-driving vehicles in logistics, a DHL perspective on implications and use cases for the logistics industry, DHL Trend Research, 2014.
20. Methodology report: puclic perception of self-driving technology for long-haul trucking and last-mile delivery, RARC-WP-17–011-B, Office of Inspector General, United States Postal Service, September 5, 2017.
21. A plan to develop safe autonomous vehicles. And prove it., Intel, 2017.
22. Self-driving safety report, Nvidia, 2018.
23. Automated Vehicles Comprehensive Plan, U. S. Department of Transportation, January 11, 2021.
24. Self-driving cars: mapping access to a technology revolution, National Council on Disability, November 2, 2015.
25. B. Richardson, Power train and automotive, IEEE ECTC 2017.
26. V. Venky Sundaram, Packaging for autonomous vehicle electronics. Application and market projections, IEEE ECTC 2017.
27. D. Xie, Data processing by DRIVE PX2, IEEE ECTC 2017.
28. T. A. Tran, Autonomous driving packaging, IEEE ECTC 2019.
29. N. Brese, AHEAD™ technology for automotive electronics, IEEE ECTC 2019.
30. S. Sun, A. P. Petropulu, and H. Vincent Poor, MIMO Radar for Advanced Driver-Assistance Systems and Autonomous Driving Advantages and challenges, IEEE SIGNAL PROCESSING MAGAZINE, July 2020, p98–117.

31. Moving to the next phase of autonomous vehicles on the Lyft network, Lyft, Deceomber 15, 2020.
32. Lessons after 100,000 self-driving rides, powered by Aptiv technology, Lyft, February 11, 2020.
33. (R) Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles, SAE J3016, April 2021.
34. B. Schlager, S. Muckenhuber, S. Schmidt, H. Holzer, R. Rott, F. Michael Maier, K. Saad, M. Kirchengast, G. Stettinger, D. Watzenig, and J. Ruebsam, State-of-the-Art Sensor Models for Virtual Testing of Advanced Driver Assistance Systems/Autonomous Driving Functions, SAE Int. J. of CAV 3(3):233–261, 2020, doi:https://doi.org/10.4271/12-03-03-0018.
35. H. Shi, H, Talisse, S. Khau, M. Marroquín, Hardware reliability in robo-taxi, IEEE 71th Electronic Components and Technology Conference (ECTC), June 1 – July 4, 2021.
36. F. Chen, Hardware Reliability Qualification of Robo-Taxi: Environmental Stress and Failure Modes for Autonomous Vehicle Modules/Components, IEEE-Electronics Packaging Society/SCV, Aug, 2021.
37. S. Jung, S. A. McKee, C. Sudarshan, C. Dropmann, C. Weis, N. Wehn, Driving Into the Memory Wall The Role of Memory for Advanced Driver Assistance Systems and Autonomous Driving, MEMSYS, October 1–4, 2018.
38. General specification for electrical / electronic components – environmental/durability, GM worldwide engineering standards GMW3172, August 2008.
39. S. Buntz, T. Hogenmuller, S. Korzin, K. Matheus, M. Mehnert, T. Streichert, M. Tazebay, J. Wuelfing, H. Zinner, Tutorial for lifetime requirements and physical testing of automotive electronic control units (ECUs), DGS-EC/EHM3-Mrt, June 25th, 2012.
40. Self-driving cars will only last four years, Ford says, Telepgraph, August 25th, 2019.
41. H. Winner, S. Hakuli, F. Lotz, C. Singer, Handbook of Driver Assistance Systems Basic Information, Components and Systems for Active Safety and Comfort, Springer, ISBN: 978–3–319–12352–3, 2016.

Basics and Applications of AI in ADAS and Autonomous Vehicles

Yan Li and Zhiheng Huang

Abstract Life-saving advanced driver-assistance systems (ADASs) and autonomous vehicles (AVs) are the fastest growing technology segment in the automotive market. Artificial intelligence (AI) is one of the most critical components in ADAS and AV. Machine learning (ML), deep Learning (DL), simulators, cloud computing, and embedded hardware platforms are entering the equation of ADAS and AV innovation, especially at level four and level five automation, where the classic rule-based ADAS functions reach their limits. This chapter reviews the basic concepts and recent applications of AI in ADAS and AV, including supervised learning, unsupervised learning, reinforcement learning, DL architectures in AVs, mostly used DL algorithms, edge cases and safety, training datasets, simulators, and infrastructures.

1 Introduction

1.1 Advanced Driver-Assistance Systems (ADASs)

Safety has been one of the top priorities in automotive industry. Statistic studies indicate that around 90% of road crash accidents can be attributed to human errors, for example, overlooked risks or low reaction time from human drivers [1]. Passive safety measures, such as shatter-resistant glass, three-point seatbelts, and airbags, which are designed to diminish damage during an accident have been the focus of automotive safety improvements in the past [2]. The newly developed advanced driver-assistance systems (ADASs) technology can actively help human drivers avoid crashing accidents in the first place and has become indispensable for modern cars.

Y. Li
Intel Corporation, Chandler, AZ, USA
e-mail: yan.a.li@intel.com

Z. Huang (✉)
The Key Laboratory of Low-Carbon Chemistry and Energy Conservation of Guangdong Province, and School of Materials Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China
e-mail: hzh29@mail.sysu.edu.cn

Y. Li and H. Shi (eds.), *Advanced Driver Assistance Systems and Autonomous Vehicles*,
https://doi.org/10.1007/978-981-19-5053-7_2

For example, *blind spot detection* alerts a driver when the car is moving into an occupied lane; *Lane departure warning* and *lane keep aid* alert actively steers the car back into its original lane during its lane drifting; *Pedestrian detection* notifies the driver when pedestrians are in front or behind the car; *Automatic emergency braking* employs the brake to avoid an accident or pedestrian injury [2].

Figure 1a illustrates four main elements of a current ADAS architecture: (1) longitudinal control, (2) lateral control, (3) driving vigilance monitoring system, and (4) parking assistance. The longitudinal control shown in Fig. 1a has been first introduced in 1995, using embedded sensors inside cars to measure the distance between two vehicles on the same lane [3]. Current mid-range vehicles [4] are providing adaptive cruise control [5] and collision avoidance system [6]. The two systems can inform the human driver about risks coming from the lane, help control the vehicle's speed, and offer evasive steering actions of the vehicle.

The lateral control illustrated in Fig. 1a is similar with that of longitudinal control, which can provide danger detection on the lateral lanes. Change lane and lane keeping are two systems assisting planned lane change action and warning unintended lane change. They are fulfilled by detecting incoming vehicles or other dangerous situations during planned lane change and recognizing sudden drifting of the vehicle's path from the road lane [7, 8].

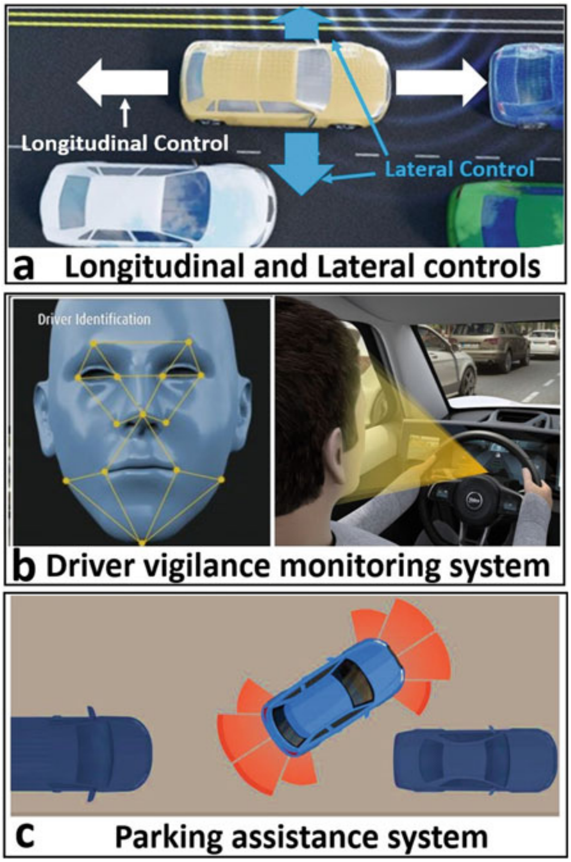
Distraction during driving is one of the leading causes of accident, especially for young drivers [9]. As illustrated in Fig. 1b, driver vigilance monitoring system can evaluate driver stress, fatigue, and anger levels through cameras and sensors embedded in the cockpit, monitoring face and expression from the driver in real time. The system can be utilized to reveal potential safety issues from unhealthy emotions of human drivers. The system is current available in some high-end cars, for example, the Mercedes-Benz attention assist system [10].

Parking-assistance system presented in Fig. 1c is currently the most advanced ADAS, as it could provide auto-parking without the steering action or presence of the driver. Cameras in front and at the back of the car, together with other lateral and longitudinal sensors, provide the perception or input layer for the car. An intelligent system employed with machine learning algorithms analyzes the input data, provides correct decisions, and automatically parks the car to the proposed location [11]. The auto-parking system is currently available in high-end cars, such as BMW X6 and Audi A8 [1].

1.2 Autonomous Vehicles (AVs) and Automation Levels

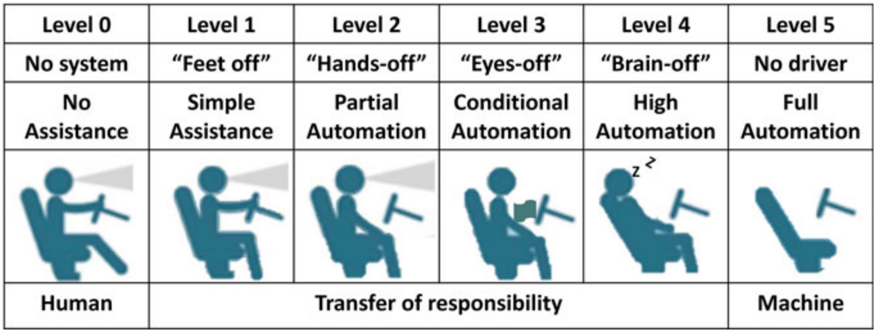
An autonomous vehicle (AV) is a vehicle that can drive itself without input from a human driver. A classification system describing the automation levels of AVs, ranging from fully automated to fully manual, has been published in 2014 by the Society of Automotive Engineers (SAE) international, and adopted by the U.S. Department of Transportation. As illustrated in Fig. 2, Level 0 to Level 2 require a human driver to always monitor the driving conditions. Level 0 means little or no assistance; the

Fig. 1 Main components of ADAS. Adapted from Ref. [1]


Longitudinal Control
Lateral Control
a Longitudinal and Lateral controls
Driver Identification
b Driver vigilance monitoring system
c Parking assistance system

human driver is completely in charge of the car, while Level 1 provides simple assistance, such as speed control; the human driver could be "Feet off" and starts transfer some driving responsibility to the machine. At level 2 or partial automation stage, both lateral and longitudinal control could be provided by the machine in particular circumstances. The human driver could be "Hand-off" but is required to constantly oversee the traffic and be ready to take over the vehicle control immediately [12].

The most important transition is between partial automation (Level 2) and conditional automation (Level 3). The vehicle takes charge of lateral and longitudinal control in many situations at level 3, while the human driver could be "Eyes-off." However, the driver needs to be ready to take over the vehicle control as a backup, whenever notified by the machine in a timely manner. As displayed in Fig. 2, the main difference between Level 4 (high automation) and Level 5 (full automation) is the system's capability to handle specific restricted driving modes versus all driving modes. At Level 4, the human driver could be "Brain-off," but still needs to take the



Level 0	Level 1	Level 2	Level 3	Level 4	Level 5
No system	"Feet off"	"Hands-off"	"Eyes-off"	"Brain-off"	No driver
No Assistance	Simple Assistance	Partial Automation	Conditional Automation	High Automation	Full Automation
Human	Transfer of responsibility				Machine

Fig. 2 Vehicle automation levels. Adapted from Ref. [12]

driving responsibility at certain defined driving modes. At Level 5 or full automation stage, machine takes full charge of driving; no human driver is needed.

Highway driving assist, installed in Genesis, Hyundai, and Kia vehicles, and Autopilot in Tesla are good examples of Level 2 automation which has been realized successfully in current ADAS [13, 14]. The recently released Honda Legend and Audi Traffic Jam Pilot demonstrate ADAS with Level 3 automation [15, 16]. The quantum leap is between Level 3 and Level 4 in terms of system reliability. Level 3 is also called conditional automation; a human driver is required to be ready to take control of the vehicle within a couple of seconds. While at Level 4, the system is managing specified traffic conditions without any intervention of human drivers. In case of unexpected events, it can reach a safety fallback state [12]. Waymo driver [17] and newly announced product-level Mobileye Drive™Self-Driving System [18] are examples of Level 4 automation.

However, the development from Level 3 to Level 4 automation is not easy. Traffic situations are highly dynamic and complex except confined spaces such as highways. Classic rule-based ADAS functions have been extended to their limits in Level 3 demands. Every viable use condition or combination of them in any given traffic situation need to be calculated in the conventional linear "if then" computer programming. This is essentially impossible for Level 4 and Level 5 automation in urban environments, requiring complicated scene interpretation, behavior prediction, and trajectory planning. Self-learning systems based on artificial intelligence (AI), especially deep learning (DL), which can mock human decisions, are becoming the key technology enablers for the success of Level 4 and Level 5 autonomous driving [12].

1.3 Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL)

Figure 3 illustrates the relationship of AI, ML, and DL. AI starts from 1950s and is a broader term that describes computer programs capable of sensing, reasoning, acting, and adapting. ML is a subfield of AI and initiates from 1970s. It involves algorithms whose performance improves with more data. Figure 4 depicts schematically the difference between traditional programming and ML [19]. Input and the algorithm are known in the traditional software development approac; the computation produces results as the output. However, in the ML approach, input and desired result are known, the algorithm which predicts the desired result, or the "program" is the output of computation. ML is capable of learning without being explicitly programmed and has wide applications in detection, prediction, and generation of data [19].

DL emerges in 2010s and is a subset of ML in which multilayered neural networks learn from vast amounts of data. Figure 5 illustrates schematically the difference between ML and DL [20]. One of the advantages of DL over ML is the needlessness of *"feature extraction"*. ML algorithms typically cannot be applied directly to the raw data, such as images and text. A process termed "feature extraction" is required to provide an abstract representation of the given raw data. For example, the classification of the data into several categories or classes. Feature extraction is usually complex and needs detailed knowledge of the problem domain. It is essential to have it tested and refined over a couple of iterations for optimized results.

DL does not require the feature extraction step because of the multi-layer artificial neural networks (ANNs). The layers can learn a valid representation in the raw data

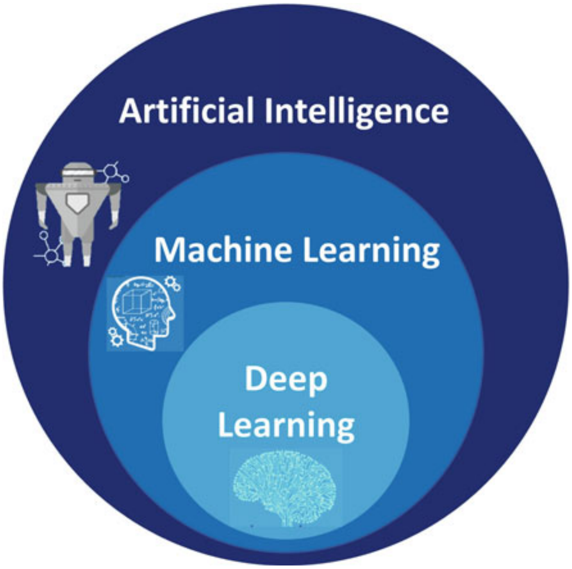


Fig. 3 Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL). Adapted from Ref. [20]

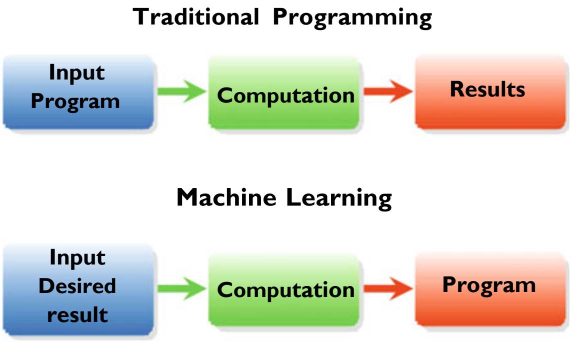


Fig. 4 Traditional programming versus machine learning. Adapted from Ref. [19]

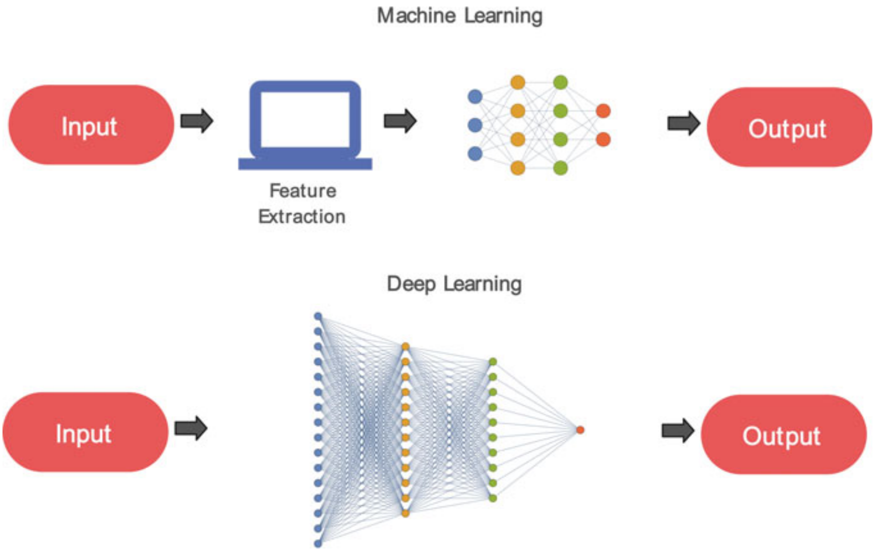


Fig. 5 Machine learning versus deep learning. Adapted from Ref. [20]

directly on their own. In other words, the feature extraction step is already a part of the process that happens in a multi-layer artificial neural network thus requires little to no manual effort to perform and optimize the feature extraction process [20].

It should be noted that, however, AI systems are nowhere near advanced enough to replace humans in many tasks, e.g., reasoning. They are showing human-level competence in low-level pattern recognition skills, but at the cognitive level, they are merely imitating human intelligence, not engaging deeply and creatively [21, 22].

2 Applications of AI in ADAS

2.1 Supervised Learning

Supervised learning is something like searching for a key under the streetlight [23]. It is the most common form of ML, and its task is to learn a mapping from inputs $\mathbf{x} \in \mathcal{X}$ to produce $\mathbf{y} \in \mathcal{Y}$. The experience is given in the form of a set of N, called the sample size, input–output pairs $\mathcal{D} = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^{N}$, known as the training dataset. The performance relies on the type of output that the model predicts [24].

Just as a teacher supervising students' learning, the supervised learning algorithm learns the process from the training dataset. Before accomplishing an acceptable level of performance, the algorithm iteratively predicts results on the training data and is corrected by the human "teacher." Supervised learning techniques are used for classification of discrete inputs and regression of continuous inputs [1].

Supervised learning has been used in ADAS to predict driving context, recognize driving events, lane changing assistance, and lane changing intent prediction [25–29]. Input data from accelerometer, GPS, and many other several sensors are preprocessed and then sent to the learning algorithms. Final output is generated when the algorithms have acceptable performance. Various supervised learning techniques, including decision trees, neural networks, Bayesian networks, Naive Bayes, nearest neighbors or instance-based classifiers, and support vector machines (SVM) have been employed for classification or regression in ADAS applications [25]. The performance of each algorithm in % accuracy is listed and compared; the best model is then chosen for the desired application [25].

For lane changing, one of the critical functions in ADAS, the accuracy for non-merge events is far more important than that of merge events. Misclassifying of a non-merge event as a merge event can result in a car accident, while misclassifying a merge event as a non-merge event would only lead to merge opportunity loss. A good balance of extremely high accuracy for non-merge event and a good accuracy for merge event is needed to ensure safety as well as effectiveness. It is found that the best prediction accuracy for lane changing is from combining two supervised learning techniques (Bayes and Decision Tree) into a single classifier using a majority voting principle [27].

2.2 Unsupervised Learning

Unlike supervised learning, unsupervised learning lets the data talk [23]. In supervised learning, each input $\mathbf{x}$ in the training set has an associated output target $\mathbf{y}$, and the goal is to learning the input–output mapping. Thus, supervised learning is essentially just "glorified curve fitting" [30]. However, unsupervised learning is to try to "make sense of" data instead of just learning a mapping. That is to say "inputs" $\mathcal{D} = \{\mathbf{x}_n : n = 1 : N\}$ can be observed in unsupervised learning but without any

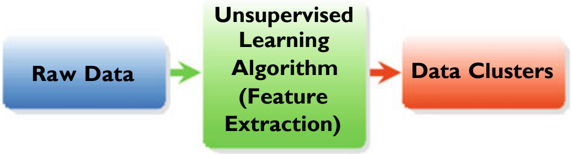


Fig. 6 Unsupervised learning algorithm. Adapted from ref. [31]

corresponding "outputs" $\mathbf{y}_n$ [24]. As illustrated schematically in Fig. 6, besides clustering data into groups, the algorithm can also extract the hidden features in the data clusters [31]. The "feature extraction" step is performed manually in supervised learning, consuming considerable amount of skilled human labor in assigning and verifying labels. In unsupervised learning, the training dataset is clustered and labeled by the algorithm, saving enormous human labor. It can also exploit previously undetected patterns from the raw data, not visible to human "experts." The benefit of unsupervised learning is more pronounced in highly dimensional or large datasets. However, compared with supervised learning models, developing unsupervised learning models typically needs more time for acceptable performance and requires a larger amount of training data. During the exploratory stage, it is common to have elevated storage and computational demands. In the supervised learning algorithm, anomalies or artifacts in the training data, which are obviously erroneous or irrelevant, could be allocated with undue significance [29, 32].

The era of *big data* for transportation arises in 2010s, due to the traffic data exploding along with the wide application of emerging traffic sensor technologies [33, 34]. Big data are typically known to have the four versus — volume, variety, velocity, and veracity. As a result of big data, feature extraction has turned into an essential step in ML applications for ADAS. Unsupervised learning becomes the key technique in the preprocessing phase of other ML techniques, especially in feature extraction, as it can deal with enormous amount of raw data. For example, in the ML applications of lateral and longitudinal control systems, unsupervised learning is successfully applied to manage all the sensors input dataset [1, 35–37].

2.3 Reinforcement Learning

Reinforcement learning (RL) is the closest to the learning process of humans and other animals. Many core algorithms of reinforcement learning are originally inspired by biological learning systems. Figure 7 illustrates schematically the key components of reinforcement learning. The learning agent, which is the RL model, takes actions in an environment, indicating the physical world in which the agent operates. The action to the environment is interpreted into a reward and a representation of the state and are fed back to the agent. The agent takes the feedbacks, adjusts its action to the environment to maximize the cumulative rewards. The learning agent interacts

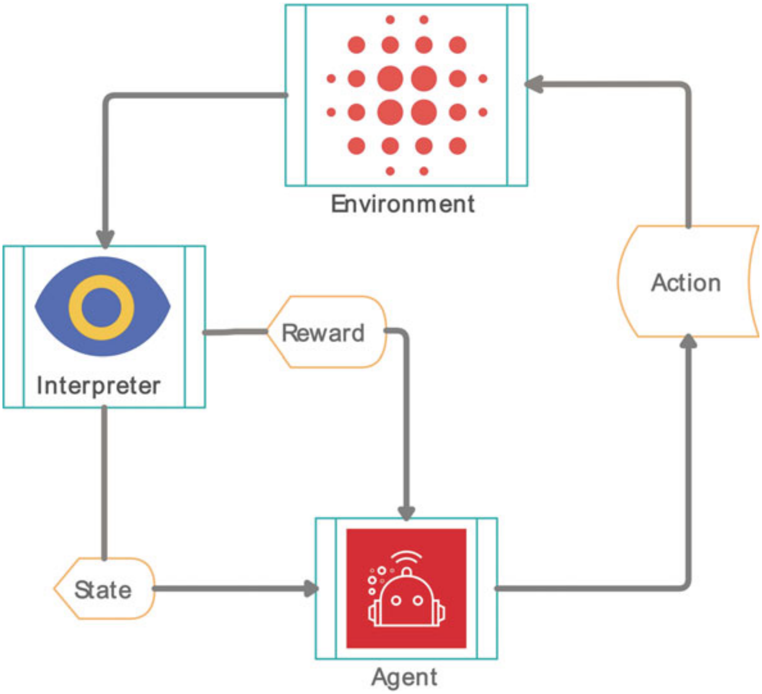


Fig. 7 Reinforcement learning. Adapted from Ref. [39]

with its environment to achieve a goal: maximizing the reward signal. It learns from interaction with the environment and can learn from its own experience [38–40].

Reinforcement learning aims to solve sequential decision-making problems in uncertain environments and is different from supervised and unsupervised learning. Unsupervised learning intends to find hidden structure in a large amount of raw data, which could not address the goal of maximizing a reward signal. Supervised learning learns from a training set with labeled examples provided by experts or supervisors. The objective of supervised learning is to extrapolate the developed model correctly in the situations not presented in the training set. However, it is not adequate for interaction problems. Because it is impractical to have examples of desired behavior that are both correct and representative of all the situations in the environment. Being able to learn from its own experience is crucial for ML algorithms to be most beneficial in uncharted territories [38].

One of the challenges emerge in reinforcement learning is the compromise between exploration of unknown domain and exploitation of present knowledge. To achieve maximum reward, a reinforcement learning agent prefers effective actions that it has tried in the past. However, to invent such actions, it needs to try actions not being selected before. On a random task, each action needs to be tested many times to obtain a reliable predict of the expected reward. The exploration–exploitation dilemma has been exhaustively studied for decades by mathematicians to reach a good balance [38].

2.3.1 Markov Decision Processes (MDPs)

MDPs consist a mathematical framework for modeling an environment in reinforcement learning [38, 39]. It is made up with:

- A set of finite environment states, S.
- A group of possible actions A in each state.
- A transition model $P_a(s, s') = \Pr(s_{t+1} = s'|s_t = a,\ a_t = a)$, describing the probability of transition at time t from state s to state s' under action a.
- A real valued reward function $R_a(s, s')$, indicating the instant reward after transition from to with action a.
- A discount factor γ, accounting for the time dimension and determining the importance of future rewards.

At time t, the environment is in state s_t, an agent takes action a_t and receives a reward $r(s_t, a_t)$. A new state s_{t+1} occurs after action a_t is taken. A state s_t in an MDP has the Markov property [23]. A future state s_{t+1} depends only on the current state s_t but not on any previous states. The current state contains all the useful information to predict the future state [23].

2.3.2 Model-Based and Model-Free RL

Model-based RL assumes that the agent knows the dynamics of the environment in the form of a model. There is only planning involved and no learning because the agent knows the transition probabilities precisely from one state to another given this history of states and actions [23]. The agent can use dynamic programming or search algorithms to find the optimal policy [23]. Model-free RL assumes that the agent does not know the model. The agent interacts with the environment but doesn't know the intrinsic dynamic that affects it. The agent has to learn and predict the rewards from her actions from experience and develop the optimal behavior. The experience can be online (live), stored in memory, or retrieved from human experts, or simulated experiences [23].

2.3.3 The Goal of an RL Agent

The goal of an agent wants to maximize his cumulative lifetime rewards. The agent has a policy π, a state-value function V, and a state-action value function Q. The policy π maps actions and current state to new states and defines the transition probability $P_\pi(s'|s)$ from state s to state s'. A state-value $V_\pi(s)$ represents the expected cumulative sum of rewards that the agent will receive if the policy π is followed starting from state s. A state-action value $Q_\pi(s, a)$ represents the expected cumulative sum of rewards that the agent will receive if the policy π is followed starting from state s and action a. The policy π can be deterministic; for each state policy s, it assigns an action $a = \pi(s)$, or it can be stochastic policy with a conditional probability distribution $\pi(a|s)$ [23].

$E[\cdot]$ is the expectation operator. Here, the expectation is given the policy π and current state s_t. The policy defines the actions $a_t, \ldots, a_T$ [23].

$$V_\pi(s_t) = E_\pi\big[r(s_t, a_t) + \gamma r(s_{t+1}, a_{t+1}) + \cdots + \gamma^T r(s_T, a_T)\big] \tag{1}$$

where $a_{t+i} = \pi(s_{t+i}),\ i = 0, \ldots, T - t$. We also have that [23]

$$Q_\pi(s_t, a_t) = E_\pi\big[r(s_t, a_t) + \gamma r(s_{t+1}, a_{t+1}) + \cdots + \gamma^T r(s_T, a_T)\big]. \tag{2}$$

2.3.4 *Q*-Learning

Real-world environments typically lack prior understanding of environment dynamics. Model-free reinforcement learning techniques would be more convenient. Q-learning is a frequently applied model-free reinforcement learning algorithm. It operates around the concept of updating Q values, which represents value of performing action a in state s. The following value-update rule is the core of the *on-policy temporal-difference* (TD) algorithm [23].

$$Q^{\text{new}}(s_t, a_t) = (1 - \alpha) \cdot Q(s_t, a_t) + \alpha \cdot \big[r_{t+1} + \gamma \cdot \max_a Q(s_{t+1}, a)\big] \tag{3}$$

Where α is the learning rate ($0 < \alpha \leq 1$), defining the extent of newly acquired information overriding old information. $Q^{\text{new}}(s_t, a_t)$ is the sum of three factors [23]:

- $(1 - \alpha) \cdot Q(s_t, a_t)$: the current value weighted by the learning rate. The larger the learning rate, the faster to Q value change.
- αr_{t+1}: the obtained reward at action a_t when in state s_t, weighted by learning rate.
- $\alpha\gamma \cdot \max_a Q(s_{t+1}, a)$: the maximum reward that can be obtained from state s_{t+1}, weighted by learning rate and discount factor γ, determining the importance of future rewards.

An episode of the algorithm ends when state is a final or terminal state. However, Q-learning can also learn in non-episodic tasks, because of the property of convergent infinite series. If the discount factor is lower than 1, the action values are finite even if the problem can contain infinite loops [23, 38–40].

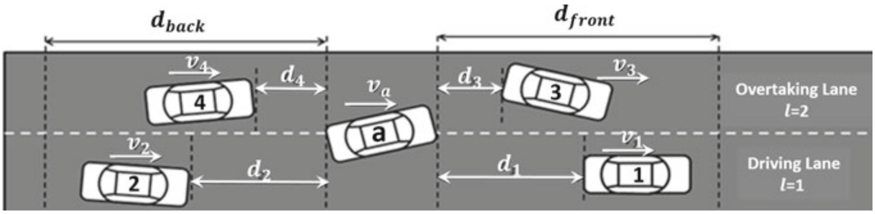


Fig. 8 Relationship between the autonomous vehicle (vehicle a) and its surrounding vehicles (vehicles 1–4). Adapted from Ref. [43]

2.3.5 Applications

RL has been applied in many disciplines, including game theory, control theory, operations research, information theory, simulation-based optimization, multi-agent systems, swarm intelligence, and statistics. Being good at making decisions continuously in unknown environment, reinforcement learning has numerous applications in ADAS, such as parking-assistance system and evasive steering actions. It has been employed to the decision-making system for complex traffic condition design and overtaking decision-making for highway autonomous driving [41–43].

A vehicle model with 14 degrees of freedom is adapted to the highway environment [42, 43]. Q-learning algorithm of reinforcement learning is used to obtain an optimized decision-making policy for autonomous car overtaking via numerous simulated driving scenarios. Besides the autonomous vehicle (AV), other vehicle in the traffic can also overtake or keep in land. Figure 8 displays the relationship between the autonomous vehicle and its surrounding vehicles during the simulated overtaking. It is assumed that all the vehicles drive on a two-lane highway, containing a driving lane ($l = 1$) and overtaking lane ($l = 2$). To make the overtake decision, only nearest vehicles, named vehicle 1, 2, 3, and 4, surrounding the AV (vehicle "a" in Fig. 8) is taken into consideration. As demonstrated in Fig. 8, d_1, d_2, d_3, and d_4 are used to represent the related locations between the corresponding vehicles. Only, the longitudinal distances are considered. d_{front} and d_{back} stand for the maximum forward sight range and maximum backward sight range, respectively. A simple expert system to achieve the overtaking decision-making is proposed and defined in Table 1 [43].

Two main indicators are selected to evaluate the overtaking policy performance for the autonomous vehicles driving in the simulated traffic flow with multiple other vehicles. (1) the average velocity of the AV; (2) the minimum distance between the autonomous vehicle and other vehicles while driving. The following equation shows the reward function used in the study, having five different aspects for giving rewards. Human driving experiences are also considered in designing the reward function [43].

Table 1 The rules of a simple expert system used in the reinforcement learning of overtaking decision-making

The current lane	Action	Condition
1	Move to Lane2	$(d_1 < 50\text{ m} \mid v_1 - v_f < -3\text{ m/s})\ \&\ d_4 > 80\text{ m}\ \&\ (d_3 - d_1 > 10\text{ m})\ \&\ (v_3 - v_1 > 1\text{ m/s})$
	Keep in Lane1	Else
2	Move to Lane2	$((d_1 > 150\text{ m}\ \&\ v_1 - v_f > 1\text{ m/s}) \mid v_1 > v_3)\ \&\ d_2 > 60\text{ m}$
	Keep in Lane1	Else

Adapted from Ref. [43]

$$r^{(k)} = \begin{cases} -300 & c = 1 \quad (4a)\\ -150 & c = 0 \text{ and } v_d^{(k)} > 10\text{ m/s} \quad (4b)\\ -0.1v_d^{(k)} & \text{else and } v_a > v_f \quad (4c)\\ v_a^{(k)} - v_f^{(k)} - 0.1v_d^{(k)} & \text{else and } l = 2 \quad (4d)\\ v_a^{(k)} - v_f^{(k)} - 0.1v_d^{(k)} + (d_1 - v_1) & \text{else and } l = 1 \quad (4e) \end{cases}$$

where $r^{(k)}$ is the reward given at k step, l is the current line number, $v_a^{(k)}$ is the AV speed at k step, $v_d^{(k)}$ is the speed difference between final moment and current one, $v_f^{(k)}$ is the expected upper limit of $v_a^{(k)}$, while d_1 and v_1 are distance between vehicle 1 and the AV and speed of vehicle 1, respectively. c is the detector of collision.

The five aspects are considered in the reward function [43]:

- Collisions must be avoided, indicated in Eq. 4a, the lowest reward is assigned when a collision happens ($c = 1$).
- Safety is the top priority: velocity change during overtaking should be minimized, illustrated in Eq. 4b. If it changes too fast: $v_d^{(k)} > 10$ m/s, an extreme low reward (half as that of a collision) is granted.
- When the current velocity of the AV is higher than its speed limit, $v_a > v_f$, a negative reward value is given to encourage deceleration, displayed in Eq. 4c.
- Vehicle should not occupy in the overtaking lane for a long time, the reward value in the overtaking lane is always lower than that in the driving lane. Reward value in Eq. 4d is lower than that of Eq. 4e.
- When overtaking a vehicle in front in the driving lane, d_1 should be kept approximately equal to the value of v_a, described in Eq. 4e.

The Q-learning algorithm illustrated in Eqs. 4a–4e are employed during the simulation. The learning rate α defining the level of repealing old information is set to 1. The discount factor γ setting up the significance of future rewards is put as 0.95 [43].

The overtaking policies developed by reinforcement learning are evaluated and validated in various traffic conditions. Simulation results indicate that the developed approach through reinforcement learning is promising and better than non-ML methods [43].

2.4 Deep Learning (DL)

As illustrated in Fig. 5, DL is based on multi-layer artificial neural networks (ANNs), inspired by information processing and distributed communication nodes in biological systems. The word "deep" originates from the fact of having "multiple" hidden layers, besides the input and output layers of ANN. DL strategies can be supervised learning, unsupervised learning, and reinforcement learning [44–46]. Recent prompt development in DL leads to many breakthroughs in computer vision, robotics, natural language processing, and Level 4–5 autonomous driving [47]. Product-level Level 4 autonomous driving system based on DL technique is commercially available in 2021 [18].

2.4.1 AI System Architectures

Autonomous driving needs decision-making systems to make real-time driving decisions based on tremendous data from GPS and on-board sensors, like cameras, radars, light detection and ranging (LiDAR) devices, and ultrasonic sensors. The driving decisions are computed either in a modular Perception-Planning-Action pipeline, illustrated schematically in Fig. 9a, or in an end-to-end learning model, presented in Fig. 9b. The modular pipeline in Fig. 9a has four sequential components, which can be designed using either DL and ML approaches, or classical methods:

- Perception and Localization
- High-Level Path Planning
- Behavior Arbitration, or Low-Level Path Planning
- Motion Controllers.

A safety monitor is designed to ensure the safety for each module. For a given itinerary, the first mission of an autonomous car is to sense and identify itself in the surrounding environment. Based on the information, a continuous path is planned, and the following detailed actions of the car are defined by the behavior arbitration system. At the end, the motion control system reactively corrects errors during the execution of the planned motion and assures that the vehicle is kept on the planned trajectory. The hierarchical process listed in Fig. 9a(a) can be encoded into a single deep learning architecture, named end-to-end learning system, which instantly links percipient information to control outputs. A safety monitor is programmed to assure its safety [47].

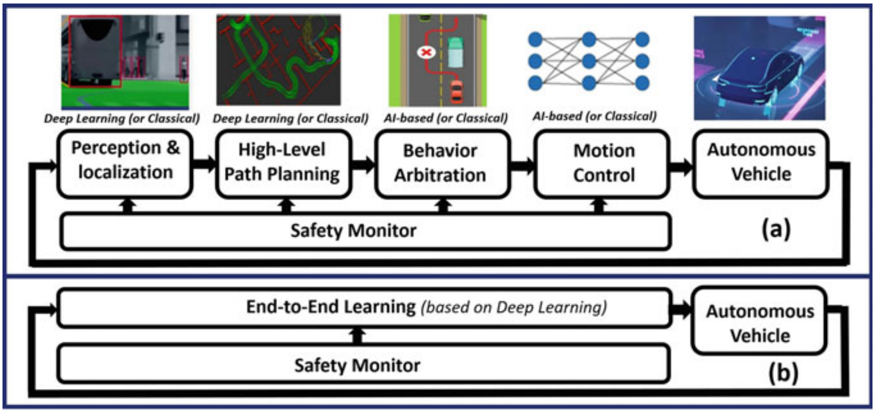


Fig. 9 Deep Learning-based autonomous car architectures: **a** Modular perception-planning-action pipeline. **b** End-to-end learning system. Adapted from Ref. [47]

The mostly used DL methodologies in autonomous driving are deep convolutional neural networks (DCNNs), recurrent neural networks (RNNs), and deep reinforcement learning (DRL). Basic concepts and applications of each approach are discussed below [47].

2.4.2 The Functions of Deep Learning

Before diving into different deep learning techniques, the functions of deep learning are first introduced in this section following the ideas from Strang [48, 49]. In fact, the constructive approximation of functions is a beautiful subject, and approximation theory is an established field [50]. While understanding on its underlying mathematics is still ongoing, deep convolutional neural networks turn out to be the best approximation of functions.

Constructing a function F to correctly classify the training data and make it also working for unseen test data is the ultimate goal of deep learning. Vectors, matrices, or sometimes tensors are the inputs to the function F. Each training sample forms an input vector v. The computed classifications, $w = F(v)$, are the outputs. The easiest *learning function* that could be thought of would be a linear form, i.e., $w = Av + b$, where b is a bias vector. The elements of the matrix A contain the weights to be optimized during the learning process. However, a linear function form is limited in many ways, and therefore, the logistic sigmoid function with an S-shaped curve turned out to be a better choice to construct F. Substantial progress had been made by using those nonlinear sigmoid functions between matrices A and B to construct the form of $A(S(Bv))$. Further research revealed that a simple ramp function R could replace the smooth function S. The function R is now named ReLU and defined as $\mathbf{ReLU}(x) = \mathbf{max}(0, x)$.

The functions behind deep learning are of the form $F(v) = L(R(L(R(\cdots(Lv)))))$. A composition of the affine functions $Lv = Av + b$ with nonlinear functions R is constructed, and this function composition acts on each component of Lv. The components in the matrices A and the bias vectors b form the weights to be learnt by using the training samples to make sure that the outputs $F(v)$ work properly for the majority of the data. After this stage, F may then be applied to new samples from the same population. If the weights are optimized well, the outputs $F(v)$ for untested data should be accurate. Intentionally, constructing more layers into the function F will usually generate more accuracy. As such, the additionally introduced hidden layers make the depth of the network deeper. It is exactly this depth in the construction of the composite function F makes deep learning such a big success. As a matter of fact, the number of weights, i.e., the components in A_{ij} and b_j in the neural network, is typically larger than the number of inputs from the training samples v. This is exactly the origin of "the curse of dimensionality" of machine learning, remaining as yet a challenging task to be solved.

2.4.3 Recurrent Neural Networks (RNNs)

Following a blog on "Understanding LSTM Networks" written by Colah [51], a brief introduction to RNNs is given here. Rather than thinking from scratch all the time, human beings typically think and understand things based on previously accumulated experience and knowledge. In other words, thoughts from humans have both memories and persistence. However, such characteristics have not been automatically built into common neural networks. RNNs are therefore designed with loops and built to address this shortcoming, as illustrated in Fig. 10. The neural network, denoted here as "NN", processes some input $x(t)$ and outputs a value $h(t)$. It is the loop built for "NN" that makes RNNs special. A careful examination on the structure of Fig. 10a pinpoints the key fact that a RNN can be thought of as multiple copies of the same network, each copy passing an information to the next one, as clearly demonstrated in the unfolded version of the network shown in Fig. 10 after the "=" sign. The chain-like nature reveals that RNNs are intimately related to temporal sequence data such as texts and videos, for which RNNs are indeed the natural architecture to choose for processing.

There is yet one more problem remains to be solved before RNNs become the working horse for processing sequence data. In some cases, only, very recent information is needed to complete the prediction for the present task, i.e., the distance between relevant information and the place where the prediction is needed is small, and RNNs can learn from the recent information and do the job well. However, RNNs become unable to learn and connect the information as the distance grows larger and larger. Hochreiter and Bengio [52] studied this fundamental problem and uncovered the reasons behind it. Based upon this progress, Hochreiter and Schmidhuber [52] further introduced a special kind of RNN named long short-term memory networks (LSTMs) to address the long-term dependencies. LSTMs perform well on a large variety of problems and are now widely used.

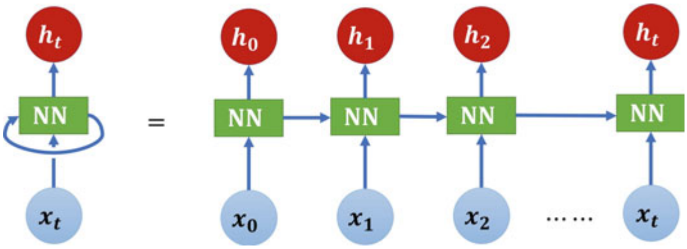


Fig. 10 A folded and unfolded recurrent neural network. Adapted from Ref. [51]

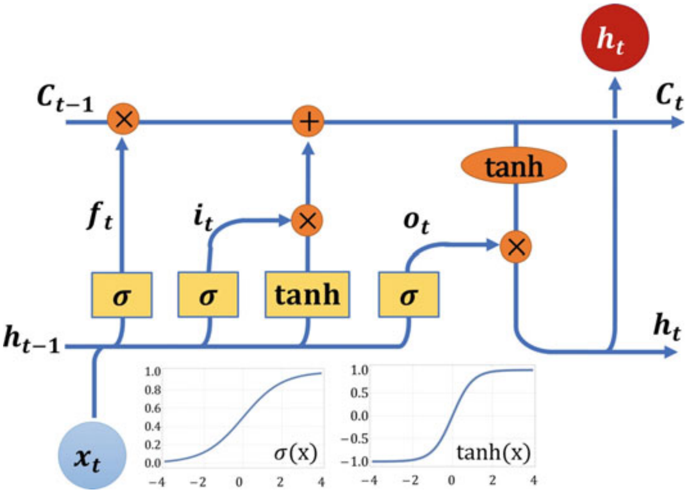


Fig. 11 The structure of the repeating module in an LSTM containing four interacting layers. The inserts below plot the logistic sigmoid and the tanh functions. Adapted from Ref. [51]

Figure 11 details the structure inside the repeating module of LSTMs. In contrast to a single layer in a standard RNN, there are four interacting neural network layers in the repeating module. The following details the working mechanism behind LSTMs.

The first layer is the “forget gate layer” denoted as f_t, which is implemented by a logistic sigmoid function $\sigma(x)$. It evaluates h_{t-1} and x_t and output a number between 0 and 1 for each number in the cell state C_{t-1}. $f_t = 1$ represents “keep the information completely” while $f_t = 0$ represents “forget the information completely”. The operations above can be written in mathematical languages as:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right), \tag{5}$$

where W_f and b_f are weights associated with the “forget gate layer”, and the other three layers will also introduce similar weights. Following the “forget gate layer” is to decide what new information needs to be stored in the cell state. An “input gate layer” enabled also by the sigmoid function decides which values to be updated,

$$i_t = \sigma\big(W_i \cdot [h_{t-1}, x_t] + b_i\big), \tag{6}$$

followed by a tanh layer creating a vector of new candidate values, $\tilde{C}_t$,

$$\tilde{C}_t = \tanh\big(W_C \cdot [h_{t-1}, x_t] + b_C\big), \tag{7}$$

that could be added to the state. W_i, b_i, W_C, and b_C are all weights introduced into the network. Next, those two operations combine to create an update to the state,

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t. \tag{8}$$

Finally, the output is constructed as follows. A sigmoid layer first decides what parts of the cell state should be output. Then, the cell state passes through a tanh layer to transform the values to be between -1 and 1 and multiply it by the output of the sigmoid gate:

$$o_t = \sigma\big(W_o \cdot [h_{t-1}, x_t] + b_o\big), \tag{9}$$

$$h_t = o_t * \tanh(C_t). \tag{10}$$

2.4.4 Deep Reinforcement Learning (DRL)

For autonomous driving tasks, the DRL typically takes the form of the partially observable Markov decision process (POMDP). The agent, which is the autonomous car, senses the environment with observation $\mathbf{I}^{\langle t\rangle}$, take an action $\mathbf{a}^{\langle t\rangle}$ in state $\mathbf{s}^{\langle t\rangle}$, interacts with its environment through an obtained reward $R^{\langle t+1\rangle}$, and moves to the next state $\mathbf{s}^{\langle t+1\rangle}$ following a transition function $T^{\mathbf{s}^{\langle t+1\rangle}}_{\mathbf{s}^{\langle t\rangle},\mathbf{a}^{\langle t\rangle}}$. Similar with the RL in classical ML approaches, the goal of a DRL is to learn an optimal driving policy navigating from state $s^{\langle t\rangle}_{\text{start}}$ to the destination state $s^{\langle t+k\rangle}_{\text{dest.}}$ by maximizing the rewards. k is the number of time steps required to reach the destination state [47]. However, the standard RL method is not feasible in high-dimensional state spaces of autonomous driving applications, with tremendous data from cameras, radar, LiDAR, and other sensors.

A nonlinear approximator termed as deep Q-network (DQN) is developed to estimate the approximate action-value function [53]. DQN is using a variant of the Q-learning algorithm to train a neural network. In DQN, an agent interacts with an environment, observes a current state, takes a discrete (legal) action that leads to a new state, and a reward. The agent wants to maximize the sum of future rewards. The simplest reward can be just 1 it wins and 0 if it loses, or it could be a total game score. DQN is a model-free off-policy reinforcement learning algorithm. It uses only samples of the environment and does not attempt to model it. It learns about the optimal strategy by mixing a greedy strategy with random exploratory strategies. In this environment, there is an optimal action-value function $Q^*(s, a)$ which is the maximum amount of rewards after taking action a in state s. To obtain the optimal strategy, a Bellman equation that relates current $Q^*(s, a)$ value and the current reward and future $Q^*(s', a')$ values is solved by iteration [23, 47].

$$Q^*(s, a) = E_{s'}\left[r + \gamma \max_{a'} Q^*(s', a')|s, a\right] \tag{11}$$

A function approximator using parameter θ is used in practice to estimate $Q^*(s, a)$:

$$Q(s, a; \theta) = Q^*(s, a) \tag{12}$$

$Q(s, a; \theta)$ is the Q-network. The parameter θ is estimated by minimizing a sequence of loss functions by SGD. There are several enhancements to the DQN algorithms. Among them are double DQN [54], prioritized experience replay [55], and dueling DQN [56].

The main challenge in DRL autonomous driving applications is the training, as the agent needs to explore its environment, typically through learning from collisions. Models trained solely by simulators tend to learn a biased version of the driving environment. Imitation learning methods, such as inverse reinforcement learning (IRL), which can learn from human driving demonstrations without exploring unsafe actions, is desirable [57].

2.4.5 DL Applications in the Modular Architecture

Cameras, LiDAR devices, radar, and acoustic sensors provide tremendous and redundant information for vehicle perception and localization. DL-based perception, especially CNNs, are good at detecting and recognizing objects in 2D images and 3D point clouds obtained from video cameras and LiDAR sensors, respectively, [58–61]. Visual-based localization algorithms target at computing the pose (position and orientation) of the autonomous vehicle at it navigates. Deep learning approaches have been applied to improve the accuracy of the visual localization [62].

Making decisions for path planning and behavior arbitration, as illustrated in Fig. 9a, is not trivial for autonomous vehicles. It needs to consider all possible obstacles in the surrounding environment and compute a collision-free trajectory. Inverse reinforcement learning (IRL), also known as imitation learning (IL) [63, 64], and deep reinforcement learning (DRL) algorithms have achieved promising results [65]. IRL or IL learns the reward function from a human driver and generates human-like driving trajectories. It can get trained with data collected from the real world. However, the real-world training data are insufficient of corner driving cases, like driving off-lanes and vehicle crashes. As a result, the response is uncertain when confronted with unseen scenarios for the IRL models. DRL mainly learns driving trajectories in a simulator, where the real environmental model is abstracted and transferred into a virtual environment by a transfer model. It can explore all types of assumed driving situations within a digital world. However, the DRL models tend to have a biased behavior once applied in the real world owning to the difference between a simulated world and a real world [47].

Motion control module displayed in Fig. 9a is responsible for making the longitudinal and lateral steering commands of the autonomous vehicle. Traditional controllers with fixed model parameters cannot foresee every possible situation in the driving and need learning-based model for high-level automatic driving. Learning controllers, such as interactive learning control [66] and model predictive control [67], build upon a priori model with a dynamic DL model. They have been used to optimally combine the established traditional control with learning algorithms for desirable outputs.

2.4.6 DL Applications in the End-to-End Architecture

Unlike conventional processing pipelines, direct mapping from high-dimensional sensory data to detailed control commands is required for the end-to-end architecture demonstrated in Fig. 9b. End-to-end learning typically involves scaled-up complex ANN models. The technological development in computing hardware over the last couple of years has promoted the application of end-to-end learning models. Some examples of published end-to-end learning systems are listed in Table 2 [47].

The approach of NVIDIA PilotNet end-to-end learning model listed in Table 2 is to train a CNN which maps raw pixels from raw camera images directly to steering commands [68]. The training data contain images and steering commands collected in driving situations of a diverse set of lighting and weather conditions, as well as on different road styles. The data are also enriched by augmentation, adding artificial shifts and rotations to the original data. PilotNet has 250,000 parameters and approximately 27 million connections, implemented on parallel graphic processing units (GPUs).

Table 2 Some examples of the End-to-End learning system

Name	Function	ANN architecture	Sensor input
ALVINN	Road following	3-layer back-prop. network	Camera, laser range finder
NVIDIA PilotNet	Autonomous driving in real traffic situations	CNN	Raw camera images
Drive360	Steering angle and velocity control	CNN + Fully connected + LSTM	Surround-view cameras etc.
DeepPicar	Steering angle control	CNN	Camera images
TORCS E2E	Steering angle control in a simulated environment	CNN	TORCS simulator images
Agile autonomous driving	Steering angle and velocity control for aggressive driving	CNN	Raw camera images

Adapted from Ref. [47]

"TORCS E2E" in Table 2. demonstrates DRL, the other approach to design end-to-end driving system. It is mainly trained in simulators, where an autonomous agent can safely explore different driving strategies [69]. DRL-based learning model can include classical model-based techniques, similar with the motion controllers discussed in Sect. 2.4.5. The hard constraints of the model-based system can be transferred into the neural network policy. Such a DRL policy trained on real-world images is used in the "Agile Autonomous Driving" of Table 2 [70].

3 Safety in ADAS and AV Based on AI

As illustrated in Fig. 2, the transition from "eyes-on" (Levels 1–2) driver assistance to "eyes-off" (Levels 3–5) demands many changes to system safety approaches. For instance, a higher level of availability is required, because the system cannot be simply deactivated right after the detection of a component hardware failure during driving. At a system functional level, making decisions on the subsequent actions based on the interpreting of current driving situations including environmental conditions are essential and challenging for conventional ADAS methodologies. AI applications, especially DL models in ADAS and AV can tackle these challenges. However, AI-based autonomous driving needs to be acceptably safe before its releasing into the public domain [71].

3.1 Safety Standards and Methodologies

ISO 26262 is a globally recognized standard for functional safety of automotive electrical/electronic (E/E) systems. The standard recommends the adoption of a Hazard Analysis and Risk Assessment (HARA) method to reveal hazardous incidents in the system and to designate safety goals that reduce the hazards. According to the standard, a "hazard" is defined as "potential source of harm caused by a malfunctioning behavior, where harm is a physical injury or damage to the health of a person" [71]. However, DL systems can create new types of hazards, which can occur because humans assume that the automated driving-assistance system based on AI techniques is more reliable than it is. For example, in a DRL system with a faulty reward function, the automated vehicle can learn a dangerous way to avoid getting punished for driving too close to other vehicles: exploiting sensor vulnerabilities so that it cannot see how close it is getting to the other vehicle. Such a unique mistake could be fatal in the real world [47].

Unlike faults of a programmed component, specific failures of a DL component could come from unreliable or noisy sensor signals, ANN topology, learning algorithm, training data set, or unexpected changes in the environment. Demonstrating the safety of an AI system depends heavily on the type of technique and the application context. It requires [47]:

- Understanding the impact of the potential fails
- Understanding the context within the whole system
- Defining the assumption of the system context and its environment
- Defining safety standards, including non-functional constraints.

3.2 AI Safety Challenges: Edge Cases and Heavy Tail Distribution

An edge case is a situation that occurs only at extreme or unexpected conditions. An edge case can be expected or unexpected during the designing and development of an AI system. Nontrivial unexpected edge cases could result in severe accidents in autonomous driving AI applications [71–73].

As illustrated in Fig. 12, a heavy tail distribution of "Probability of Surprises" versus "Total Training & Test Time" has a tail heavier than an exponential distribution. The possibility of surprises goes to zero much slower in a heavy tail distribution, comparing with that in an exponential distribution. In other words, a heavy tail distribution tends to have more outliers or edge cases. Possibility of surprises in the real world follows the heavy tail distribution. The autonomous vehicles based on AI need to be robust when encountering novel but dangerous events in the real world. The training and testing data for the AI system, which enables it to handle the common things in the real world, are either general real-world road data or hypothesized extreme situations in simulators. However, the heavy tail distribution of surprises in the real-world suggests that the AI system will face new but vicious situations in the real world regardless of its prolonged training and testing time. For example, in the first fatal autonomous driving accident, the autonomous car driven by AutoPilot collides with a truck, despite the 130 million miles of testing and evaluation. The accident happens under an extremely rare condition, i.e., an edge case. The combination of following factors: (1) the height of the truck, (2) its white color under bright sky, (3) its position on road, leads to an object misclassification error in the AutoPilot system [74]. The misinterpretation directs the AV to drive into the truck and results in a fatal accident. Human drivers are good at dealing with edge cases in the real world, as humans have common sense learned from their daily life that autonomous cars, trained by statistical data, don't have [73].

3.3 Safety in AI System Design, Validation, Testing and Implementation

As illustrated in Fig. 9, in AI systems of autonomous vehicles, safety monitors are designed for each module of a modular architecture and the whole end-to-end learning system in an end-to-end architecture. The autonomous control software will

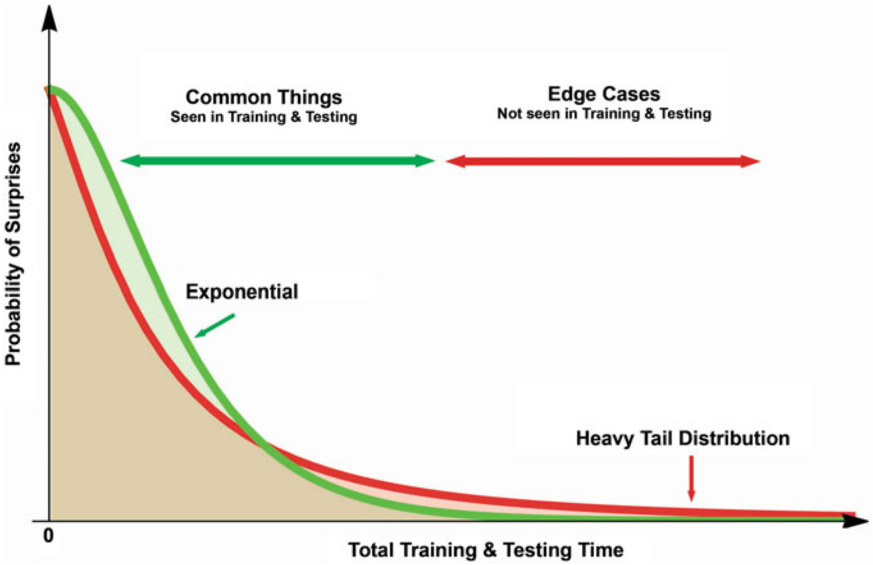


Fig. 12 Heavy tail distribution and edge cases. Adapted from Ref. [72]

be suspended if a failure is detected by the safety monitors. Specific fault type and failures have been cataloged for ANN, which facilitates the development of specific tools and techniques for fault identification. A white box technique is employed to inject faults onto an ANN by breaking the links or randomly changing the weights [47].

The training dataset plays a key role in the safety of an AI component. ISO 26262 standard states that the component behavior shall be fully specified, and each refinement shall be verified with respect to its specification. However, in a DL system, a training set is used instead of a specification. It is not distinct how to assure that the corresponding hazards are always mitigated. As detailed in Sect. 3.2, edge cases of the heavy tail distribution in the real world, which is not present in the training set, could lead to fatalities. Detailed requirements, which specifies the methodologies of training validation and testing sets, shall be formulated and traced to hazards. Current standards and regulations from the automotive industry cannot be fully applied to DL systems; the development of new safety standards targeting DL systems is still ongoing [47].

Collecting all the edge cases happened in the real world and including them in the training set is one of the approaches to improve the safety of an AI system. A large amount of fatal and non-fatal accident reports, as well as enormous examples of the tight, tricky interactions occurring between cars and other road users are gathered and codified. The collection of edge cases can be exposed to the AI system through training and testing [75].

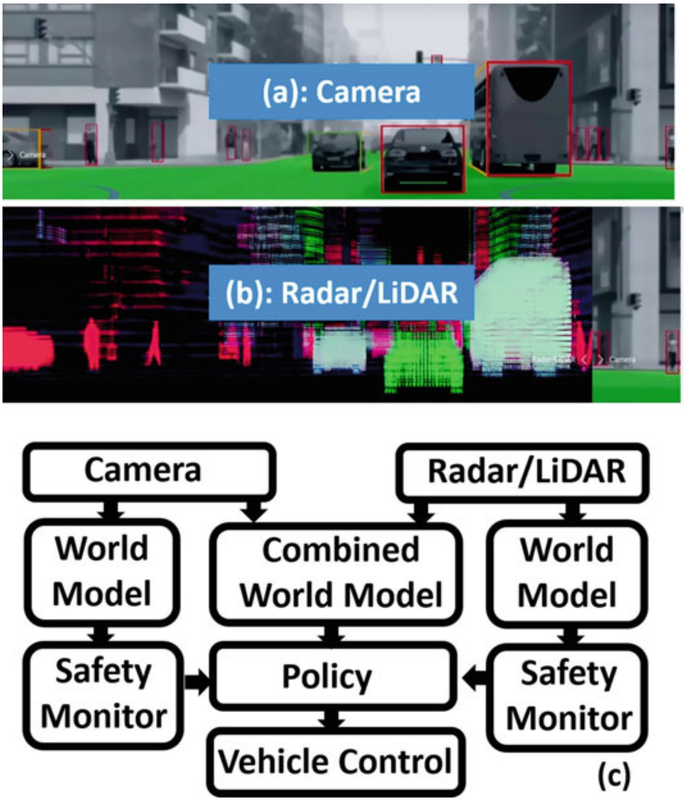


Fig. 13 The true redundancy system combining two independent perception sub-systems used in Mobile Drive. Adapted from Ref. [76]

Making the AI system more robust to surprises is an alternative approach for safety improvement. Testing the AI system limit by inject noise into sensor values, enabling the AI system's awareness of its own limits, and transition vehicle operation to safer modes when uncertainty increases have the potential to improve the robustness [72]. Additionally, design sensor redundancy can also significantly improve the robustness. As illustrated in Fig. 13, the true redundancy system utilized in Mobil Drive, the newly announced commercially available Level 4 autonomous driving system, has two independent perception sub-systems: the camera system and the radar-LiDAR system. Each system can drive the car alone, having independent safety monitor and serve as the backup for one another [76].

Progressive and rigorous AI system validation, testing, and implementation plans are proven to be successful in AI safety improvements. Each change of the AI system in Waymo Driver, the publicly available Level 4 autonomous driving, needs to go through:

- Simulation Testing
- Closed-Course Testing
- Real-World Driving.

In simulation testing, any changes or updates are rigorously tested. The most challenging situations observed in the real-world are turned into virtual scenarios for the testing. Data from crash databases are studied to reveal other possible collision situations for further improvement of the simulation test. After the simulation testing, the AI system is updated to a few vehicles for closed-course testing, performed on privately owned test track, designed to mimic the cities in the real word. Specific features within different operation design domains are tested in the private facility. Once the updated AI system gets validated in the closed-course testing, the new software is introduced to a small number of autonomous vehicles on public roads. It is then gradually applied to the entire fleet after more confidence in the real-world driving test is built up. The AI system is continuously refined and updated during the real-world driving to assure its safe drive at Level 4 automation [17].

4 Datasets, Simulators, and Infrastructures for AI Systems

4.1 Publicly Available Training and Testing Datasets

High amount of real-world driving data for training and testing is crucial to the success of the AI system development in AVs [47]. However, collecting representative real-world data is very resource intensive. To facilitate the research and development of AI applications in AVs, in recent years, more and more companies and research institutes have made their autonomous driving datasets open to the public. For example, Waymo Open Dataset, extracted from Waymo autonomous driving vehicles, provides well synchronized and calibrated high-quality annotated LiDAR and camera data obtained from a wide range of urban and suburban geographies [77]. It contains 1000 types of different segments where each segment has 20 s of continuous driving, corresponding to 200,000 frames at 10 Hz per sensor [78]. In addition to Waymo Open Dataset, other well-known publicly available datasets including [78]:

- A2D2 Datasets for Autonomous Driving—released by Audi
- ApolloScape Open Dataset for Autonomous Driving—Part of Apollo project
- Argoverse Dataset—modern AV dataset providing forward facing stereo imagery
- Berkeley DeepDrive Dataset
- CityScapes Dataset
- Comma2k19 Dataset
- Google-Landmarks Dataset
- KITTI Vision Benchmark Suite
- LeddarTech PixSet Dataset
- Level 5 Open Data—published by Lyft

- nuScenes Dataset—developed by Motional
- Oxford Radar RobotCar Dataset
- PandaSet
- Udacity Self-Driving Car Dataset.

4.2 Open-source Simulators

As detailed in Sects. 2 and 3, simulators play an important role in AI system development, especially during training, testing, and implementation. Various open-source simulators have been developed in the past few years to promote the research and development of AI systems in ADAS and AV. For example, CARLA, an open-source simulator, provides open-source code and protocols, as well as digital assets such as urban layouts, buildings, and vehicles. The free simulation platform facilitates flexible specification of sensor suites, environmental factors, as well as full control of map generation and actors. It can support both modular and end-to-end AI system development [79].

Simulator for Urban Driving in Massive Mixed Traffic (SUMMIT), built upon CARLA, is a high-fidelity simulator facilitating the development and testing of crowd-driving algorithms. It emulates unregulated and heavy urban traffic by leveraging the open-source OpenStreetMap map database and a heterogeneous multi-agent motion prediction model. Besides CARLA and SUMMIT, other top autonomous driving open-source simulators are [79]:

- Flow: for DRL and control experiment of traffic microsimulations
- PGDrive: built upon the Panda3d and Bullet engine
- Deepdrive: support for Linux and Windows
- AirSim: developed by Microsoft
- LGSVL Simulator: developed by LG Electronics
- Gym-Duckietown: simulator for the Duckietown Universe.

4.3 Infrastructures for AI Systems

Unlike conventional ML models, a DL system involves millions, instead of hundreds, of parameters and much larger datasets, such as video, image, or text data, for training. Training the DL algorithms demands extendable storage, distributed processing, computing capabilities, and accelerators. Cloud computing turns into an attractive platform for the development of end-to-end AI applications in ADAS and AV, by providing comprehensive services for data storage, processing, as well as backend services [80]. It offers:

- Data storage
- Platform-as-a-Service (PaaS)
- Software-as-a-Service (SaaS)
- Infrastructure-as-a-Service (IaaS).

For example, Amazon Web Service (AWS) provides the hardware necessary for DL algorithm training and implementation. Developers do not need to physically have a computing-intensive hardware system for computation. By using clusters of GPUs and central processing units (CPUs) for complex matrix computations, as well as leveraging distributed networks to ingest and manage large datasets for algorithm training, DL on the cloud is easier, faster, and with a lower cost [80].

Deploying DL algorithms on target edge devices, which are the autonomous vehicles for ADAS and AV applications, is not a trivial object. Embedded hardware platforms are required to be portable, versatile, and energy efficient and are crucial for the integration of trained AI algorithms inside vehicles. A systems-on-a-chip (SoC) system containing multiple cores and GPUs is proven to be capable for DL applications [47].

A field-programmable gate array (FPGA) is the other choice, showing great improvements in both power consumption and performance for DL applications. Studies show that FPGAs consume ten times less power comparing with GPUs when processing algorithms with the same computation complexity [47]. By reducing the latency in DL applications, FPGAs can provide additional raw computing power. The large amount of chip cache memory in FPGAs reduces or eliminates memory bottlenecks in current SOC systems. Additionally, FPGAs can support a full range of data types, alongside with custom user-defined types. FPGAs are built with multiple architectures, which are a mix of hardware programmable resources, digital signal processors, and processor block RAM (BRAM) components. The architecture flexibility fits well for deep and sparse ANNs, which currently have wide applications in ADAS and AVs. Connecting to various input and output peripheral devices, such as sensors, network elements, and storage device, is also possible due to its flexibility. FPGAs have been originally designed to meet the rigorous functional safety requirement for a wide range of industry applications. Comparing to GPUs, which are originally designed and built for graphics and high-performance computing systems without strict functional safety requirements, FPGAs provide a significant advantage in AI applications in ADAS and AV [47].

5 Summary

ADAS and AV, which can actively avoid human mistake-induced car accidents, are one of the focused areas in both academia and automotive industry. AI applications are essential for Level 4 and Level 5 autonomous driving, as the classical computation model cannot consider every possible use condition and combination of them in complex urban driving conditions. Learning-based algorithms that can mimic human decision-making processes are the solutions of complicated scene interpretation, behavior prediction, and trajectory planning.

ML techniques, including *supervised learning*, *unsupervised learning*, and *reinforcement learning*, have wide applications in longitudinal and lateral controls, driving vigilance monitoring, and parking-assistance systems. Thanks to recent technology breakthroughs, DL, which is based on ANNs having many hidden layers, becomes the key technological enabler for Level 4 and Level 5 autonomous driving. Basic concepts and applications of the mostly used DL methodologies in AVs, with either a hierarchical modular pipeline or an end-to-end architecture, are discussed.

The possibility of having *edge cases*, i.e., rare but dangerous events, in the real-world follows a heavy tail distribution. The AI system will face new but destructive conditions in the real world no matter how long its training and testing time is. Moreover, some unique mistakes in DL algorithm could be fatal if having wrong human assumptions. Safety monitors in AI system design, which can detect and diagnose failures, are critical to assure successful AI system implementations. Collecting all the edge cases happened in the real world and encoding them to training datasets is one of the approaches to enhance AI safety. Making the AI system more robust to surprise by testing the AI system limit, enabling the AI system's awareness of its own limit, and designing sensor redundancy is another methodology for safety improvement. Furthermore, comprehensive and strict AI system validation, testing, and implementation plans are proven to be useful for AI system safety assurance. ISO 26262 is a globally recognized standard for functional safety of automotive E/E systems. However, some of the regulations cannot directly apply to a DL system, whose performance relies on training datasets. The development of new safety standards specific to DL systems is still ongoing.

Large amount of real-world driving data collected exhaustively by various companies and research institutes are publicly available and critical for AI applications in AVs. Various open-source simulators haven been developed in the past few years and played key roles in AI model training, testing, and implementation. Cloud computing, which can provide all-around services, including data storage, processing, backend services, makes the training of DL algorithms easier, faster, and with a lower cost. SOC systems and FPGAs are proven to be appropriate embedded hardware platforms capable for DL applications inside AVs.

Acknowledgements The author (ZH) acknowledges financial support for his research by the National Natural Science Foundation of China (NSFC) grant number 51832002. The editors would like to thank Ms. Susan Li from Infineon for her dedicated review of this chapter.

References

1. A. Moujahid, M.E. Tantaoui, M.D. Hina, A. Soukane, A. Ortalda, A. ElKhadimi, A. Ramdane-Cherif, Machine learning techniques in ADAS: a review, in h2018 International Conference on Advances in Computing and Communication Engineering (ICACCE) (IEEE, 2018), pp. 235–242

2. G. Cooper. The evolution of deep learning for ADAS applications (2017). https://semiengineering.com/the-evolution-of-deep-learning-for-adas-applications. Accessed 30 January 2022
3. Mitsubishi motors develops "new driver support system" (1998). https://www.mitsubishi-motors.com/en/corporate/pressrelease/corporate/detail429.html. Accessed 30 January 2022
4. The SUV with cutting-edge technology and design 500 (2018).https://www.fiat.it/fiat-500x. Accessed 30 January 2022
5. BMW technology guide: Cruise control (2018).http://www.bmwesys.com/guides/DIY. Accessed 30 January 2022
6. Pre-collision braking system (2018). https://carmanuals2.com/get/subaru-forester-2014-pre-collision-brakingsystem-.html. Accessed 30 January 2022
7. R.O. Frankel, O. Gudmundsson, B. Miller, J. Potter, T. Sullivan, S. Syed, D. Hoang, J.M. John, K.S. Liao, P. Nahass, A. Schwab, J. Yuan, D. Stavens, C. Plagemann, C. Nass, S. Thrun, Assisted highway lane changing with RASCL, in AAAI Spring Symposium: Embedded Reasoning (2010)
8. Ford. Lane keeping technology: Helps drivers stay between the lines.https://media.ford.com/content/dam/fordmedia/North. Accessed 30 January 2022
9. E.S.. Associates. Teen drivers car accidents statistics (2017). https://www.edgarsnyder.com/car-accident/who-was-injured/teen/teen-driving-statistics.html. Accessed 30 January 2022
10. Mercedes-Benz of Greenwich. What is Mercedes-Benz attention assist®? (2019). https://www.mercedesbenzgreenwich.com/mercedes-benz-attention-assist/. Accessed 30 January 2022
11. G. Briochi, M. Colombetti, M.D. Hina, A. Soukane, A. Ramdane-Cherif, Techniques for cognition of driving context for safe driving application, in 2016 IEEE 15th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC) (IEEE, 2016), pp. 388–397
12. Deloitte. Autonomous driving moonshot project with quantum leap from hardware to software & AI focus (2019). https://www2.deloitte.com/content/dam/Deloitte/be/Documents/Deloitte_Autonomous-Driving.pdf. Accessed 30 January 2022
13. J.S. Choksey, C. Wardlaw. Levels of autonomous driving, explained (2021). https://www.jdpower.com/cars/shopping-guides/levels-of-autonomous-driving-explained. Accessed 30 January 2022
14. R. Baldwin. Tesla tells California DMV that FSD is not capable of autonomous driving (2021). https://www.caranddriver.com/news/a35785277/tesla-fsd-california-self-driving. Accessed 30 January 2022
15. The audi vision of autonomous driving (2021).https://media.audiusa.com/en-us/releases/184. Accessed 30 January 2022
16. L. Kim. The audi vision of autonomous driving (2021). https://www.jdpower.com/automotive-news/the-friday-five-the-new-outlander-prices-crumbling-u-s-infrastructure-self-driving-honda-edition?make=&model=. Accessed 30 January 2022
17. Waymo safety report (2021). https://downloads.ctfassets.net/sv23gofxcuiz/4gZ7ZUxd4SRj1D1W6z3rpR/2ea16814cdb42f9e8eb34cae4f30b35d/2021-03-waymo-safety-report.pdf. Accessed 30 January 2022
18. Presenting the mobileye drive™self-driving system (2021). https://www.mobileye.com/blog/mobileye-drive-self-driving-system. Accessed 30 January 2022
19. A gentle introduction to machine learning (2019). https://towardsdatascience.com/a-gentle-introduction-to-machine-learning-599210ec34ad. Accessed 30 January 2022
20. A. Oppermann. Artificial intelligence vs. machine learning vs. deep learning (2019). https://towardsdatascience.com/artificial-intelligence-vs-machine-learning-vs-deep-learning-2210ba8cc4ac. Accessed 30 January 2022
21. K. Pretz. Stop calling everything AI, machine-learning pioneer says (2021). https://spectrum.ieee.org/stop-calling-everything-ai-machinelearning-pioneer-says. Accessed 30 January 2022
22. L.B. Eliot, AI Self-driving Cars Divulgement: Practical Advances in Artificial Intelligence and Machine Learning (LBE Press Publishing, 2020)
23. M. Trinh, The AI Model Handbook: A Guide to the World of Artificial Intelligence Modeling (Rodeo Press, 2021).https://books.google.com/books?id=lIO3zgEACAAJ

24. K.P. Murphy, Probabilistic Machine Learning: An introduction (MIT Press, 2021)
25. P. Tchankue, J. Wesson, D. Vogts, Using machine learning to predict the driving context whilst driving, in Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference (2013), pp. 47–55
26. C. D'Agostino, A. Saidi, G. Scouarnec, L. Chen, Learning-based driving events classification, in 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013) (IEEE, 2013), pp. 1778–1783
27. Y. Hou, P. Edara, C. Sun, IEEE Transactions on Intelligent Transportation Systems **15**(2), 647 (2013)
28. C. D'Agostino, A. Saidi, G. Scouarnec, L. Chen, IEEE Transactions on Intelligent Transportation Systems **16**(4), 2155 (2015)
29. B. Morris, A. Doshi, M. Trivedi, Lane change intent prediction for driver assistance: On-road design and evaluation, in 2011 IEEE Intelligent Vehicles Symposium (IV) (IEEE, 2011), pp. 895–901
30. J. Pearl, arXiv preprint arXiv:1801.04016 (2018)
31. M. Jones. A gentle introduction to machine learning (2017). https://developer.ibm.com/articles/cc-unsupervised-learning-data-classification/. Accessed 30 January 2022
32. G.E. Hinton, T.J. Sejnowski, et al., Unsupervised Learning: Foundations of Neural Computation (MIT press, 1999)
33. V. Jha, Study of machine learning methods in intelligent transportation systems. Master's thesis, University of Nevada, Las Vegas, Nevada, USA (2015)
34. Y. Lv, Y. Duan, W. Kang, Z. Li, F.Y. Wang, IEEE Transactions on Intelligent Transportation Systems **16**(2), 865 (2015)
35. A. Jahangiri, H.A. Rakha, IEEE Transactions on Intelligent Transportation Systems **16**(5), 2406 (2015)
36. C. Miyajima, Y. Nishiwaki, K. Ozawa, T. Wakita, K. Itou, K. Takeda, F. Itakura, Proceedings of the IEEE **95**(2), 427 (2007)
37. R. Bhoraskar, N. Vankadhara, B. Raman, P. Kulkarni, Wolverine: Traffic and road condition estimation using smartphone sensors, in 2012 Fourth International Conference on Communication Systems and Networks (COMSNETS 2012) (IEEE, 2012), pp. 1–6
38. R.S. Sutton, A.G. Barto, Reinforcement learning: An introduction (MIT Press, 2018)
39. E. Gravelle. Shield AI fundamentals: On reinforcement learning (2018). https://shield.ai/content/2018/11/20/shield-ai-fundamentals-on-reinforcement-learning. Accessed 30 January 2022
40. S. Bhatt. Reinforcement learning 101 (2018). https://towardsdatascience.com/reinforcement-learning-101-e24b50e1d292. Accessed 30 January 2022
41. R. Zheng, C. Liu, Q. Guo, A decision-making method for autonomous vehicles based on simulation and reinforcement learning, in 2013 International Conference on Machine Learning and Cybernetics, vol. 1 (IEEE, 2013), vol. 1, pp. 362–369
42. M.G. Lagoudakis, R. Parr, The Journal of Machine Learning Research **4**, 1107 (2003)
43. X. Li, X. Xu, L. Zuo, Reinforcement learning based overtaking decision-making for highway autonomous driving, in 2015 Sixth International Conference on Intelligent Control and Information Processing (ICICIP) (IEEE, 2015), pp. 336–342
44. Y. Bengio, A. Courville, P. Vincent, IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(8), 1798 (2013)
45. J. Schmidhuber, Neural Networks **61**, 85 (2015)
46. Y. LeCun, Y. Bengio, G. Hinton, Nature **521**(7553), 436 (2015)
47. S. Grigorescu, B. Trasnea, T. Cocias, G. Macesanu, Journal of Field Robotics **37**(3), 362 (2020)
48. G. Strang, SIAM News **51**(10), 1 (2018)
49. G. Strang, Linear Algebra and Learning from Data (Wellesley-Cambridge Press Cambridge, 2019)
50. N. Trefethen, Approximation Theory and Approximation Practice, Extended Edition (SIAM, 2020)

51. C. Olah. Understanding LSTM networks (2015).https://colah.github.io/posts/2015-08-Understanding-LSTMs/. Accessed 28 March 2022
52. J. Schmidhuber. Sepp Hochreiter's fundamental deep learning problem (2013). https://people.idsia.ch/~juergen/fundamentaldeeplearningproblem.html. Accessed 28 March 2022
53. V. Mnih, K. Kavukcuoglu, D. Silver, A.A. Rusu, J. Veness, M.G. Bellemare, A. Graves, M. Riedmiller, A.K. Fidjeland, G. Ostrovski, et al., Nature **518**(7540), 529 (2015)
54. H. Van Hasselt, Y. Doron, F. Strub, M. Hessel, N. Sonnerat, J. Modayil, arXiv preprint arXiv:1812.02648 (2018)
55. T. Schaul, J. Quan, I. Antonoglou, D. Silver, arXiv preprint arXiv:1511.05952 (2015)
56. Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, N. Freitas, Dueling network architectures for deep reinforcement learning, in International Conference on Machine Learning (PMLR, 2016), pp. 1995–2003
57. M. Wulfmeier, D.Z. Wang, I. Posner, Watch this: Scalable cost-function learning for path planning in urban environments, in 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE, 2016), pp. 2089–2095
58. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., International Journal of Computer Vision **115**(3), 211 (2015)
59. J. Dai, Y. Li, K. He, J. Sun, R-FCN: Object detection via region-based fully convolutional networks, in Advances in Neural Information Processing Systems (2016), pp. 379–387
60. H. Zhao, X. Qi, X. Shen, J. Shi, J. Jia, ICNET for real-time semantic segmentation on high-resolution images, in Proceedings of the European Conference on Computer Vision (ECCV) (2018), pp. 405–420
61. C.R. Qi, W. Liu, C. Wu, H. Su, L.J. Guibas, Frustum PointNets for 3d object detection from RGB-d data, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018), pp. 918–927
62. D. Barnes, W. Maddern, G. Pascoe, I. Posner, Driven to distraction: Self-supervised distractor learning for robust monocular visual odometry in urban environments, in 2018 IEEE International Conference on Robotics and Automation (ICRA) (IEEE, 2018), pp. 1894–1900
63. L. Sun, C. Peng, W. Zhan, M. Tomizuka, A fast integrated planning and control framework for autonomous driving via imitation learning, in Dynamic Systems and Control Conference, vol. 51913 (American Society of Mechanical Engineers, 2018), vol. 51913, p. V003T37A012
64. S.M. Grigorescu, B. Trasnea, L. Marina, A. Vasilcoi, T. Cocias, IEEE Robotics and Automation Letters **4**(4), 3441 (2019)
65. L. Yu, X. Shao, Y. Wei, K. Zhou, Sensors **18**(9), 2905 (2018)
66. B. Panomruttanarug, International Journal of Automotive Technology **18**(6), 1099 (2017)
67. P. Drews, G. Williams, B. Goldfain, E.A. Theodorou, J.M. Rehg, Aggressive deep driving: Combining convolutional neural networks and model predictive control, in Conference on Robot Learning (PMLR, 2017), pp. 133–142
68. M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L.D. Jackel, M. Monfort, U. Muller, J. Zhang, et al., arXiv preprint arXiv:1604.07316 (2016)
69. E. Perot, M. Jaritz, M. Toromanoff, R. De Charette, End-to-end driving in a realistic racing game with deep reinforcement learning, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (2017), pp. 3–4
70. Y. Pan, C.A. Cheng, K. Saigol, K. Lee, X. Yan, E.A. Theodorou, B. Boots, Agile autonomous driving using end-to-end deep imitation learning, in Robotics: Science and Systems (2018)
71. S. Burton, L. Gauerhof, C. Heinzemann, Making the case for safety of machine learning in highly automated driving, in International Conference on Computer Safety, Reliability, and Security (Springer, 2017), pp. 5–16
72. P. Koopman, The heavy tail safety ceiling, in Automated and Connected Vehicle Systems Testing Symposium (SAE International, 2018)
73. C. Barnden. Is the vehicle automation numbering system useful? (2021). https://www.embedded.com/is-the-vehicle-automation-numbering-system-useful. Accessed 30 January 2022

74. S. Levin. Tesla fatal crash: ‘autopilot’ mode sped up car before driver killed, report finds (2018). https://www.theguardian.com/technology/2018/jun/07/tesla-fatal-crash-silicon-valley-autopilot-mode-report. Accessed 30 January 2022
75. dRISK. What are edge cases? (2022).https://drisk.ai/what-are-edge-cases. Accessed 30 January 2022
76. Mobileye. True redundancy™: The realistic path to deploying AVS at scale (2022).https://www.mobileye.com/true-redundancy. Accessed 30 January 2022
77. P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, et al., Scalability in perception for autonomous driving: Waymo open dataset, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020), pp. 2446–2454
78. C. Viernicke. 15 best open-source autonomous driving datasets (2021). https://www.siasearch.io/blog/best-open-source-autonomous-driving-datasets. Accessed 30 January 2022
79. A. Choudhury. Top 8 autonomous driving open source projects one must try hands-on (2021). https://analyticsindiamag.com/top-8-autonomous-driving-open-source-projects-one-must-try-hands-on. Accessed 30 January 2022
80. A. Luckow, M. Cook, N. Ashcraft, E. Weill, E. Djerekarov, B. Vorster, Deep learning in the automotive industry: Applications and tools, in 2016 IEEE International Conference on Big Data (Big Data) (IEEE, 2016), pp. 3759–3768

Computing Technology in Autonomous Vehicle

Fen Chen and Dong Zhao

Abstract The future of driving is quickly evolving toward AI-enabled, fully autonomous vehicles. The centralized Compute system will serve as a nerve center for all autonomous vehicles to meet stringent intelligence, performance, safety, security, and reliability requirements. We're seeing the complexity of autonomous driving systems growing at an unprecedented rate, and computational processing needs to keep pace with this growth. A high-performance, automotive-grade Compute system must be able to accommodate numerous sensor inputs from cameras, radars, light detection and ranging radars (LiDAR), ultrasonic sensors, inertial sensor module (ISM), acoustic sensors, and Vehicle-to-Vehicle (V2V)/Vehicle-to-Everything (V2X) communications concurrently to accurately and reliably perceive the environment around the vehicles. Also, it must be able to promptly enable better and safer driving decisions including prediction, planning, and control after analyzing all the perceived information. In this chapter, motivations, as well as various, Compute architectures and key components consisting of an advanced autonomous vehicle Compute system such as System on Chip (SoC), memory, storage, and network are reviewed. Furthermore, real-time operating system, onboard management, fault detection and diagnostics, security, and middleware will be illustrated. How to conduct rigid electrical tests and reliability validation to qualify autonomous vehicle Compute will be covered. Finally, challenges in Compute design, manufacturing, and validation including performance, power consumption, thermal management, size, cost, safety, security, quality, and reliability are explored for safe deployment of the autonomous vehicle at scale.

F. Chen (✉)
Cruise LLC, San Francisco, CA, USA
e-mail: fen.chen@getcruise.com

D. Zhao
Nio, San Jose, CA, USA
e-mail: Andy.zhao@nio.io

Y. Li and H. Shi (eds.), *Advanced Driver Assistance Systems and Autonomous Vehicles*,
https://doi.org/10.1007/978-981-19-5053-7_3

1 Introduction

The future of driving is quickly evolving toward AI-enabled, fully autonomous vehicles (AV). Autonomous driving (AD) is another great paradigm shift in the 100-year history of the automobile industry, which will redefine the rules of the automotive industry. The product definition of a vehicle will no longer be a "walking precision instrument" or a "computer on the wheels", but a "living space on the wheels". The role of the car OEMs will transform from a traditional car manufacturer to a Transportation as a service (TaaS) provider. Autonomous driving is an inevitable trend in the development of the industry. It is about time and life and is a key technology to reshape the future society. Since the second half of 2018, there has been a massive influx of capital into the global autonomous driving industry and the springing up of extensive new companies dedicated to the making of autonomous technologies. The prelude to the commercialization of AD has begun. The benefits of adopting AD are.

- Reduce transportation cost
- Reduce carbon emission
- Reduce riskily and distracted driving so to improve road safety
- Alleviate road congestion through higher throughput
- Offer accessibility, convenience, and independence for special needed people
- Improve human productivity and/or allow greater time for rest.

Maintaining consistent autonomous driving operations in all situations is challenging. The corner or unanticipated scenarios like the sudden onset of inclement weather or unsafe road conditions require vehicles to adapt in real-time. In general, such unanticipated cases are not the scenarios that you can code for. Only an onboard centralized Compute system that is capable of dynamically interpreting and quickly reacting can mitigate this kind of unusual scenario safely on time. Such a centralized Compute system requires the data and the ability to process that data in real-time using a combination of computing power and efficient deep learning neural networks. Therefore, the centralized Compute system will serve as a nerve center for all autonomous vehicles to meet stringent intelligence, performance, safety, security, and reliability requirements. A high- performance, automotive-grade Compute system houses the central Compute and connectivity to accommodate vision, radar, ultrasonic radar, acoustic sensors, ISM, and LiDAR signal transmissions. It must be able to accommodate massive data from numerous sensor inputs and V2X communications concurrently to accurately and reliably perceive the environment around the vehicles. Figure 1 illustrates how an AV's Compute sees and detects surrounding objects such as vehicles, pedestrians, and traffic lights by sensors on a rainy day. AD requires a much greater and more reliable awareness of everything around the vehicle as compared to traditional vehicles. The AV Compute is required to understand what they are "seeing" and the ability to control the vehicle to adapt to the situation evolving outside the car. It should be noted that this requirement is dramatically different from the Compute required by simpler Advanced driver assistance system (ADAS) functions like adaptive cruise control or emergency braking.

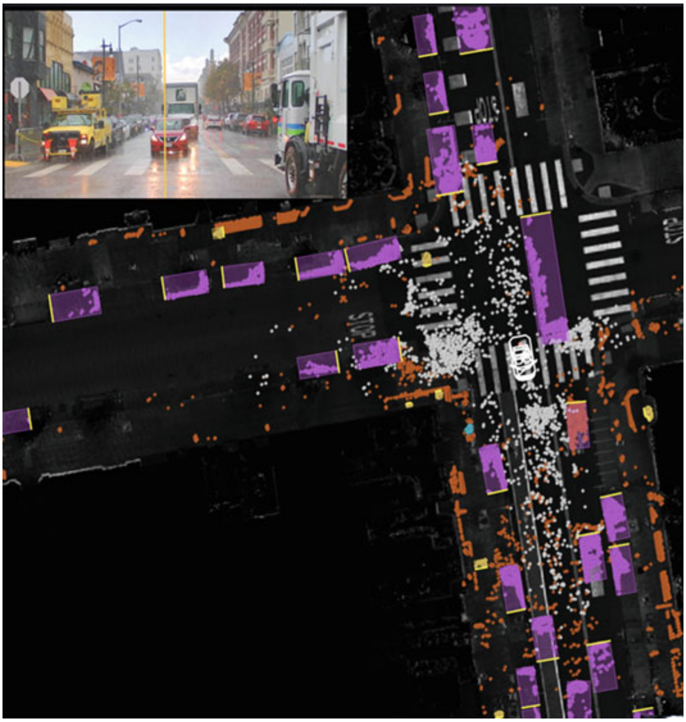


Fig. 1 Onboard Compute perception of surrounding environment around the AV

Maintaining consistent ADAS and autonomous driving (AD) operations in all situations requires machine learning, computer vision, and sensor fusion. Machine learning, computer vision, and sensor fusion will play critical roles in next-generation AVs. The high-performance AV Compute must be able to enable better and safer driving decisions including prediction and planning and must control promptly after analyzing all the perceived information as demonstrated in Fig. 2. To increase overall neural network capacity and boost the performance and responsiveness of automated driving perception systems, a Compute with high-speed parallel processing and massive processing acceleration becomes a key requirement.

One of the key missions of AV technology is to improve road safety to reduce the road fatality rate. World Health Organization (WHO) reported that about 1.27 million people die due to road traffic accidents each year [1]. Safety is a critical part of the AV Compute system. With AV, we are essentially to use a sophisticated Compute system to replace the human driver to make a safe decision. International standard ISO 26262 was developed for traditional automotive electric/electronic systems. ISO26262 defines functional safety features and requirements for all automotive electronic and electrical safety-related systems [2]. However, AV is not within its scope. For AV, there is a need to implement more rigorous safety standards and certifications to assure the highest levels of passenger and environmental safety. A

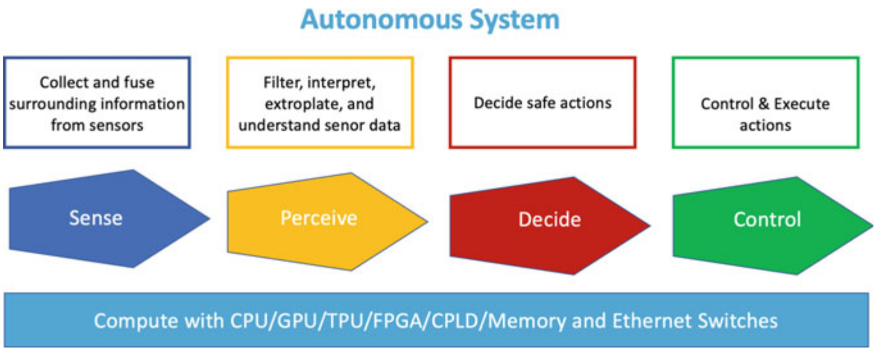
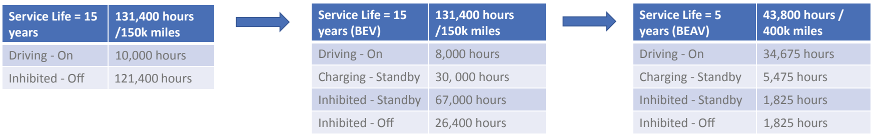

Autonomous System
Collect and fuse surrounding information from sensors
Filter, interpret, extroplate, and understand senor data
Decide safe actions
Control & Execute actions
Sense
Perceive
Decide
Control
Compute with CPU/GPU/TPU/FPGA/CPLD/Memory and Ethernet Switches

Fig. 2 An illustration of an autonomous system for AV

high-performance AV Compute must be built from the ground up to meet desired safety requirements, from development to validation to deployment. It is essential to address safety needs during the early stage of design cycles. A design for safety must be embedded in the design from day one to guarantee true automotive-grade safety conformance.

Reliability is another critical part of the AV Compute system, and it is closely tied to vehicle safety. Without a human driver, any hardware performance degradation and failures including soft failures due to performance regression and an intermittent malfunction could trigger a fatal accident. A Compute system must be able to be functional with top performance and must respond faster than the human driver at all times to guarantee AD safety. It is well known that hardware failure rates could be high during the early vehicle usage life and late wear-out period. As there have been no specific safety rules and inadequate field safety lessons learned for AV, improving an overall reliability target to reduce Compute failure rates of both hard and soft failures is essential to guard band the safety of AV. On the other hand, AV hardware, in general, will have its unique mission profiles. As shown in Fig. 3, the trend of vehicle operating time increases from traditional vehicles to Robo-taxi AVs. AV has 2–3 times of life mileages and vehicle operating hours as compared to traditional vehicles. In general, if AVs are used for road-sharing taxi business, vehicles will be required to be operative for over 20 h per day to maximize their business profit goal. With such longer daily continuous operation hours or mileages, the reliability specifications for AV hardware especially for Compute are high and challenging. A superior reliability resilience, which ensures the continuity of reliability and safety throughout the entire AV life cycle, is required for AV Compute.

This chapter presents state-of-the-art Compute systems for AD, covering five key performance metrics and nine key hardware enablement technologies, followed by validation and challenges to realize AV Compute operational performance. The remainder of this chapter is organized as follows. Section 3.2 discusses the general architectures of computing systems for AD. In Sect. 3.3, we show six key hardware and three key software constituents of an advanced Compute system. In



Service Life = 15 years	131,400 hours /150k miles
Driving - On	10,000 hours
Inhibited - Off	121,400 hours

Service Life = 15 years (BEV)	131,400 hours /150k miles
Driving - On	8,000 hours
Charging - Standby	30, 000 hours
Inhibited - Standby	67,000 hours
Inhibited - Off	26,400 hours

Service Life = 5 years (BEAV)	43,800 hours / 400k miles
Driving - On	34,675 hours
Charging - Standby	5,475 hours
Inhibited - Standby	1,825 hours
Inhibited - Off	1,825 hours

Fig. 3 Mission profile comparison among traditional vehicles, battery electric vehicles, and battery electric autonomous vehicles

Sect. 3.4, functional tests and validation of the AV computing system are introduced. Section 3.5 presents possible challenges for large-scale deployment. Finally, this chapter concludes in Sect. 3.6.

2 Compute and ADAS Technology

AV utilizes high-performance computing platforms together with a complex software system to enable a real-time AI-based perception and decision-making for vehicle maneuvers. A sophisticated AD algorithm in general requires a high volume of sensor data and a complex computational pipeline. Therefore, a Compute needs to process an enormous amount of data in real-time with extremely small latency. As the level of autonomy increases, the data generated by the AV will become larger and larger. According to Intel's estimation [3], assuming that an AV is equipped with GPS, ultrasonic sensors, camera, radar, and LiDAR sensors, the data generated by the above-mentioned sensors per second is shown in Table 1. For a vehicle driving about 20 h per day, AV Compute needs to process about 8 TB of data every day.

How to enable AV Compute to process such a large amount of data in real-time, and then based on the extracted information, make logical decisions that control safe driving behaviors is a challenge.

Two key questions to solve the above problems are:

1. Where the data processing is done: distributed-based architecture, centralized-based architecture with one central processing unit, or a hybrid-based architecture with several decentralized computing units?
2. How to transmit data from the sensors to the central processing unit: when data fusion is performed on multiple sensors that are not located in one place but spread

Table 1 Typical data size generated by various sensors per second for AV

GPS	50 kB
Ultrasonic radar	10–100 kB
Camera	20–40 MB
Radar	10–100 kB
LiDAR	10–70 MB

over the different locations of AV, the connectors and wiring cables between the sensors and the central processing unit need to be specially designed.

As shown in Fig. 4, in a complex network, three typical network structures: Centralized, Decentralized, and Distributed, are referenced.

Centralized Computing Architecture The central Compute will receive and process the raw data transmitted by each sensor. The central Compute will make and execute the decision. With the fusion of sensor data, each sensor knows what each other sensor is doing.

Decentralized Computing Architecture is a mixture of Centralized and Distributed computing solutions. Several preprocessing Computes process various sensor data before sending them to the high-level central Compute.

Distributed Computing Architecture each sensor processes its data to a certain extent, and also makes decisions locally. Only object data is transmitted from the sensor to the central Compute. The central Compute integrates the object data from each sensor first and then makes decisions and executions.

There are pros and cons for each architecture [4]. For distributed computing architecture, the advantages are that each sensor terminal processor does not have to process a large amount of data at once, and there is less demand on how to transmit data from the sensor to the central Compute safely and efficiently. A lower bandwidth, simpler and cheaper interface can be used between the terminal sensor and the central Compute. In most cases, a bandwidth of less than 1 MB per second is sufficient. Since a lot of data processing is done at the terminal processor, the increase in the number of sensors will not greatly decrease the performance of the central Compute. Since the central Compute only needs to integrate the object data, it has lower requirements on computing power and lower power consumption. It can combine various sensors in a cost-efficient way. Its disadvantages are that this computing architecture must distribute information at the same time and synchronize the information among all sensor nodes. When the number of nodes exceeds 3–4, this approach has almost become very difficult. The central Compute obtains object data rather than actual sensor data, so it cannot real-time track a specific "areas of interest" event. Also, as the terminal sensor needs to be equipped with a processor, its volume will be larger,

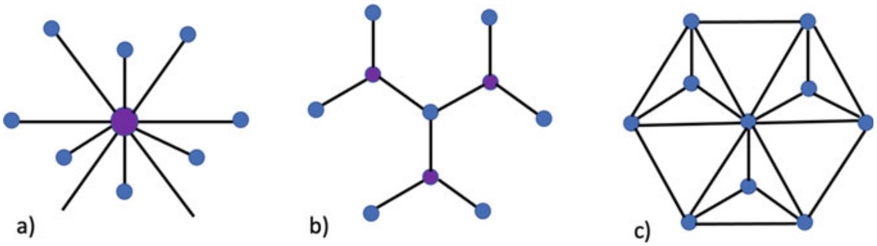


Fig. 4 Three typical network structures: **a** Centralized, **b** decentralized, **c** distributed

and the overall price and power consumption will also be higher. Since sensors need to process data and make decisions locally, their requirements for functional safety will also be higher.

The advantages of centralized computing architecture are that the cost and power consumption of terminal sensors are low since it only needs to complete the task of sensing and transmitting data. Therefore, the requirements for functional safety are low for terminal sensors. The sensor size can be small, so the installation space required is small. The installation position is also flexible and the replacement cost can be low. On the other hand, the central Compute will get the best quality information. The reason is that if the terminal sensor does not modify the data or filter the data, the central Compute can obtain the maximum possible information or original data to make the correct decision if needed. The disadvantages are that the central Compute will become a "big monster". GB-level data must be transmitted from various sensors with ultra-low latency and such massive data must be processed in time by the Compute without any delay. Broadband communication of up to several GB per second is required for data transmission and collection in real-time, which may result in high electromagnetic interferences. The central Compute needs powerful computing power and speed to process all the data transmitted from the terminal sensors, which consumes a lot of power and generates a lot of heat. In addition, the increase in the number of terminal sensors will greatly decrease the performance of the central Compute. If the central Compute is non-scalable, it will not be able to provide the required functional performance for AV needs with AV technology scaling up.

Terminal sensors are always needed to process data locally which can reduce bandwidth requirements and help to reduce AV costs. On the other hand, a centralized Compute is always needed to integrate the information of all terminal sensors to complete the overall perception of the vehicle's surrounding environment and make decisions for AV pathfinding, maneuvering, and motion trajectory. Therefore, the hybrid decentralized computing architecture that finds the optimal combination of distributed and centralized architectures is more likely to be the final technology path.

As currently, most AV Compute prototypes are centralized, we will use it as our AV computing system reference architecture. Generally, per functionality, the AV Compute can be divided into computation, network and communication, storage, and power supply management. The following sections will discuss the corresponding components in more detail.

2.1 Levels of Autonomous Driving

SAE International (Society of Automotive Engineers International) published the revised version of the autonomous vehicle classification standard in 2018. It defines six different levels of automation, ranging from Level 0 (no automation) to Level 5 (full automation), known as SAE J3016. Currently, as shown in Fig. 5, most vehicles

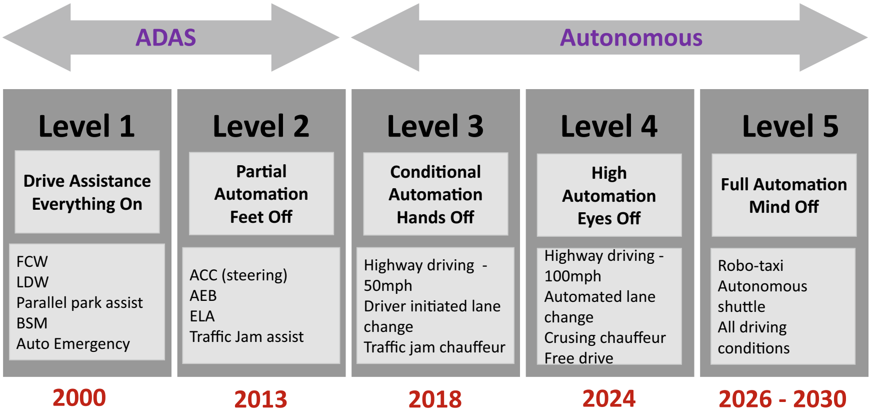


Fig. 5 Progress from Level 1 to Level 5 autonomous vehicle with timelines

on the road are only at SAE levels 0 to 2 with ADAS functionality. For such levels, human drivers remain the key controller of the vehicle to make driving decisions and are responsible for all potential hazards that occurred during driving. Every advancement in automation level requires substantial hardware and software technology advancements, and proper management of all safety–critical functions.

2.2 *Platform for Autonomous Driving System*

AV has two meanings: "intelligence" and "ability". The so-called "intelligence" refers to the ability of the vehicle to perceive, synthesize, judge, reason, decide and remember as intelligently as a human. The so-called "ability" means that the AV can ensure the effective execution of the "intelligence", implement active control, and be able to perform human–computer interaction and collaboration. Autonomous driving is an organic combination of "intelligence" and "ability". The two complement each other and are indispensable.

To realize "intelligence" and "ability", the core competencies of an autonomous vehicle system can be broadly categorized into four categories: environment perception, localization, decision-making and planning, and vehicle control. Similar to the human driver's perception of the driving environment and vehicle status through visual, auditory, and tactile sensory systems during driving, the AD system acquires its status and surrounding environment information by configuring internal and external sensors. Internal sensors mainly include vehicle speed sensors, acceleration sensors, wheel speed sensors, and yaw rate sensors. Mainstream external sensors include cameras, lidars, millimeter-wave radars, ultrasonic radars, and positioning systems. These sensors can provide massive amounts of information about the driving environment in all directions. To effectively use this kind of sensor information, it is necessary to use sensor fusion technology to combine the independent information, complementary information, and redundant information of a variety of sensors in space and time according to certain criteria, to provide an accurate understanding of the surrounding environment and own motion status. The decision-making planning subsystem represents the cognitive layer of autonomous driving technology, including two aspects of decision-making and planning, rule-based and AI-based. Figures 6 and 7 illustrate the AV process control flow and configuration schematic for AV software and hardware, respectively.

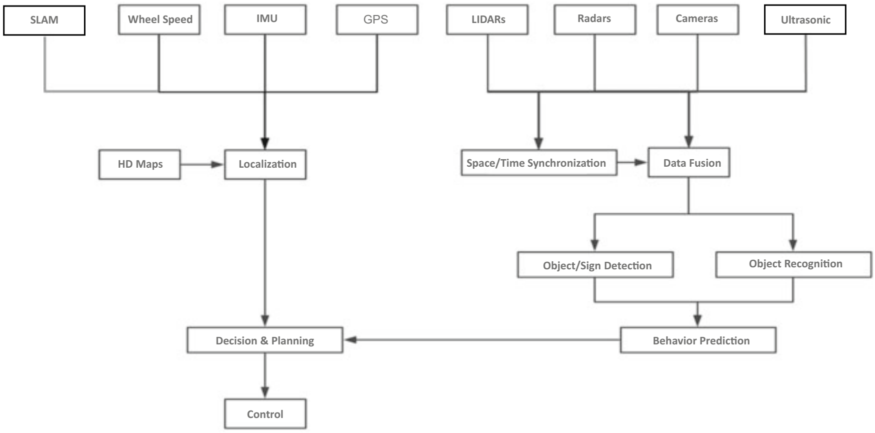


Fig. 6 The flow chart of an autonomous driving software system

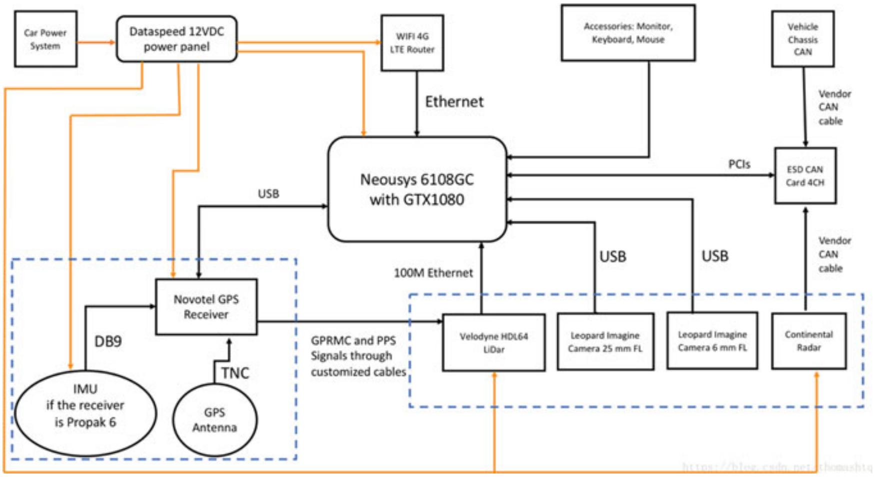

Car Power System
Dataspeed 12VDC power panel
WIFI 4G LTE Router
Accessories: Monitor, Keyboard, Mouse
Vehicle Chassis CAN
Ethernet
Vendor CAN cable
Neousys 6108GC with GTX1080
PCIs
ESD CAN Card 4CH
USB
100M Ethernet
USB
USB
Vendor CAN cable
Novotel GPS Receiver
DB9
TNC
GPRMC and PPS Signals through customized cables
Velodyne HDL64 LiDar
Leopard Imagine Camera 25 mm FL
Leopard Imagine Camera 6 mm FL
Continental Radar
IMU if the receiver is Propak 6
GPS Antenna

Fig. 7 The schematics of a hardware configuration for the Baidu Apollo AD system [5]

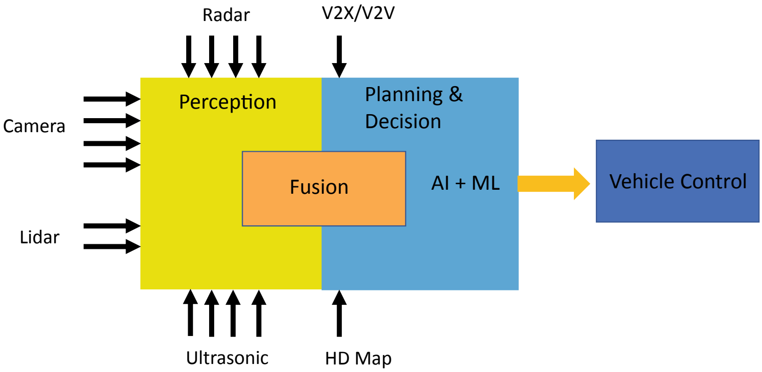


Fig. 8 AV with V2X and V2V for perception, planning, decision, and control

The decision-making system defines the interrelationship and function allocation between the various parts and determines the safe driving mode of the vehicle. The planning part is used to generate a safe, real-time collision-free trajectory. The vehicle control subsystem is used to realize the vehicle's longitudinal distance, vehicle speed control, lateral vehicle position control, etc. It is the final executive mechanism of vehicle intelligence. Environmental perception and decision planning correspond to the "intelligence" of the autonomous driving system, while vehicle control reflects its "ability".

To realize L4 or L5 autonomous driving, it may not be enough to rely just on the "smartness" of a single AV. As shown in Fig. 8, Vehicle-to-Vehicle (V2V) and Vehicle-to-everything (V2X) communications can be leveraged to achieve further improvements in areas of perception and/or planning through vehicle cooperation. Road conditions and traffic data through the V2X and V2V can provide more information than the internal and external sensors of a single AV. Such real-time data together with the help of a high-definition 3D dynamic map can enhance the perception of the environment for the no-line-of-sight situations. For example, under severe weather conditions such as rain, snow, and heavy fog, or in challenging scenes such as intersections and corners, radar and cameras cannot clearly distinguish the obstacles ahead. V2X and V2V can be used instead, which can realize an intelligent prediction of road conditions and avoid accidents.

2.3 Perception and Localization

Perception and localization are two of the most critical parts of AV Compute. Without the quantitative perception of the 3D environment around the vehicle, the decision-making initialized by Compute cannot work properly. Perception refers to the ability of an AV to collect information and extract relevant knowledge from the environment. AV needs to develop a capability to understand the surrounding environment such as road obstacles, road signs, and the movements of other road agents. Localization refers to the ability of the AV to determine its position concerning the environment. The main tasks of perception and localization include vehicle position, motion status, object detection, and object tracking.

Environment perception tasks can be fulfilled by using radars, ultrasonic sensors, LiDARs, cameras, and IR cameras, or a fusion among them to extract road traffic conditions and on-road object detection. Different sensors have different strengths and weaknesses. Ultrasonic radar is mainly used for vehicle reversing due to its limited reaction speed and resolution. Millimeter-wave radar and LiDAR are responsible for the medium and long-range environmental perception. LiDAR can produce 3D measurements and detect object traveling speed. But it offers little information on objects' appearance. The camera is mainly used for the identification of traffic lights and to provide rich appearance data with much more details on the objects. But its performance is not consistent especially under dark illumination conditions. It also does not implicitly provide 3D information. Therefore, sensor fusion is required to make full use of the advantages of each sensor.

Localization is the task to determine the pose of the ego vehicle (position and orientation) and measuring its motion. Knowledge of the ego vehicle's position is a critical piece of information that enables AV Compute to execute safety-related, AD maneuvers. One of the most popular ways of localizing a vehicle is the combination of satellite-based navigation systems, inertial measurement units (IMU), and a high-definition HD digital map. Satellite navigation systems, such as GPS and GLONASS, can provide a consistent outcome on the global position of the vehicle. However, the use of GPS and GLONASS requires reliable service signals from space satellites and the update rate is comparatively low. Inertial measurement units, which use an accelerometer, gyroscope, compass, and signal processing techniques to estimate the attitude of the vehicle at a very fast update rate (every 5 ms), do not require external infrastructure. However, IMU's accuracy is not great, and the error accumulates over time. In general, Kalman Filter techniques are used to combine the advantages of GPS/GLONASS and IMU to provide accurate and real-time position updates. Map aided localization algorithms use local features to achieve highly precise localization and have seen tremendous development in recent years. In particular, Simultaneous Localization and Mapping (SLAM) is a promising method, which refers to a process in which a moving object calculates its position based on the information from the sensors while constructing a real-time map of the environment. There is also a dynamic HD map-assisted localization method, which is a sensor system-based DeepMap with super dynamic perception capability. It can deliver road information (road geometry information, congestion information, construction information, etc.) and obstacle information (position, speed, type, etc.) to the AV in real-time through a high-performance AD cloud infrastructure. To implement this method, fast data collection and transmission capabilities are required.

2.4 Prediction, Planning, and Control

Environmental perception and localization mainly play a role in determining the state of the external environment and provide a basis for decision-making and planning. The task of prediction, planning, and control for Compute is to continuously provide collision-free decisions and motion trajectories from the current pose of the vehicle to the given destination, taking into account system dynamics, obstacles, and possibly desired criteria such as trip frequency, travel time, cost and conformable ride function. In the decision module, the main problem to be solved is how the vehicle should go. This is divided into two aspects, namely path planning and behavior planning.

1. Path planning

Path planning generally is a computational work to find a sequence of road paths to move the object from the source to the destination. It is a technology in the field of high-precision maps. In the traditional human-driving mode, if there is an error in the map navigation, it can be corrected by a human driver. In the field of autonomous driving, map accuracy and navigation accuracy will directly affect AV safety and user experience. Therefore, the high-precision map is very important. Path planning is the problem of finding the shortest path between two points. Commonly used algorithms for finding the shortest distance include Dijkstra, Floyd, A*, and RRT algorithms.

2. Behavior planning

The behavioral planner focuses on the AV on-road behaviors to assure it follows road rules and interacts with other agents safely. The prediction of traffic agents can be achieved through a variety of algorithms, and a set of motion models can be constructed. There is a lot of uncertainty in the behavior of other vehicles on the road such as accelerating and turning. The commonly used solution is to use Gaussian noise to represent the uncertainty of traffic participants. Because most of the participants' behavior must follow a normal distribution, the entire model construction can be regarded as a Gaussian process. The prediction of the behavior and intentions of traffic participants can be regarded as a dynamic time series process, and the corresponding problems can be solved by using convolutional neural networks (CNN).

Speaking of the vehicle itself, the local motion planning that requires decision-making includes: driving, following, turning, changing lanes, stopping, etc. How the vehicle makes decisions needs to be judged dynamically. The overall process of vehicle own motion planning should be divided into four steps as shown in Fig. 9. The first step is to perceive the changes in the environment. As an example, if a vehicle in front of the AV starts to merge into the lane that AV is currently using, per the perception of the local scene, a model should be used for prediction and decision. The final behavior output of AV maybe just slow down or change into another lane to assure safety with local goal setting. During the decision-making process, other vehicle behaviors and whether they comply with road rules and regulations must also be considered. The overall decision-making process of each behavior could be long, and each decision-making step affects the other. Therefore, the function of this

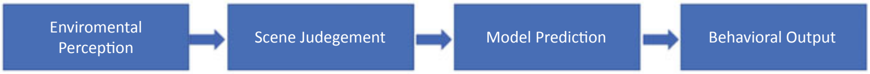


Fig. 9 Vehicle own motion behavior planning process

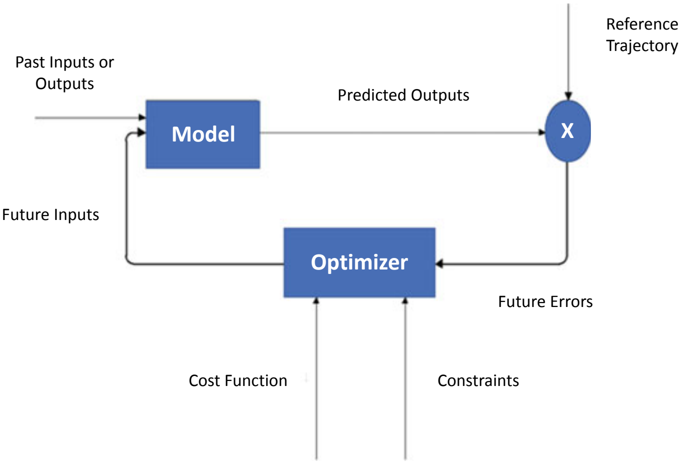


Fig. 10 Basic structure of MPC

kind of AV behavior decision-making can be regarded as a series of probabilistic additions, which can be modeled as a Markov decision process.

After environmental perception and decision-making planning, it comes the step of execution control. The execution control is a critical task for Compute. How to transmit the decision to the functional hardware components of the vehicle and implement the required accelerator, brake, steering, and shift commands are the keys to controlling the vehicle maneuvers. The most feasible solution for AV to control the behavior of each component is through the CAN power bus, and transmit instructions to each component through electronic signals. Autonomous systems need motion models for control execution. A control approach that uses system modeling is commonly referred to as Model Predictive Control (MPC). Figure 10 illustrates the basic structure of MPC. As shown in Fig. 10, it optimizes the control input by minimizing an index function while satisfying constraints and guaranteeing vehicle safe operation.

2.5 Functional Safety

As the complexity of the AV system continues to increase, new technologies will introduce new safety risks. Uber's self-driving car accidentally killed a pedestrian in March 2018. The US National Highway Transportation Safety Administration launched an investigation into Tesla's Autopilot feature, which has allegedly been involved in 11 collisions with stopped cars, resulting in one death and 17 injuries, over the past three-and-a-half years. Volvo issued a large-scale recall notice to the global market in March 2020. The number reached 700,000 vehicles, involving nine models on sale. The reason for the recall was that Volvo had previously conducted a safety test on the XC60 in Denmark. It was found that the Autonomous Emergency Braking (AEB) system did not stop the vehicle in time in the event of a collision as expected. A fatal crash involving NIO's autopilot function happened in August 2021, and the driver was killed while driving NIO's ES8 SUV. Therefore, the safety of AVs has received a lot of attention.

Function Safety (FUSA) refers to failure behaviors caused by hardware and software failures or unexpected behaviors in the design of automobiles. It is defined as safety due to the absence of unreasonable risk and is only concerned about malfunctioning systems. Currently, automotive safety frameworks include ISO 26262, the functional safety standard, and ISO/PAR 21448.1, the safety of the intended functionality. ISO 26262 standard defines the functional safety terms and activities of electrical and electronic systems in motor vehicles. Therefore, it can only solve the hardware and software hazards that may affect the safety of autonomous vehicles. The standard defines four automotive safety integrity levels (ASIL). ASIL D is the most stringent, and A is the smallest. Each level is associated with its specific development requirements, which must be complied with during certification. ISO 21448 SOTIF pays special attention to failure causes related to system performance limitations and predictable system misuse. Either hardware technical limitations (such as sensor performance limitations and noise) or software algorithm limitations (such as target detection failures and actuator technical limitations) could result in limited performances or insufficient functions for AV operation. User misuse such as overload and confusion could result in failures of AV operation as well. SOTIF is designed for Level 0–2 autonomy. SOTIF can be viewed as an extension of the functional safety process, specifically designed to solve the challenges of autonomous driving functions. SOTIF also uses hazard analysis and risk assessment (HARA) to identify hazards due to performance limitations and abuse. To demonstrate that the safety requirements are met for AV Compute, a process of design for safety, unit testing, and system verification needs to be thoroughly conducted.

Regarding safety risk mitigation, an intelligent driving safety system should be implemented to provide safety analysis and real-time monitoring services for potential problems in the perception, decision-making, and control modules of AV. Based on the concept of expected functional safety, the driving scene and system safety are analyzed and evaluated to improve the safety of AD. The driving behavior of AV highly depends on the stability, intelligence, and safety of the AV hardware and software systems. The main sources of safety risks for AV are as follows:

1. Hardware safety

Compared with traditional cars, AVs do not require the human driver to directly control the vehicle. But instead, it transfers part or all of the vehicle control to the automatic control system. AV vehicle motion perception and sensor data fusion functions play a decisive role in AD. Whether the hardware architecture setting is technically sound and sophisticated or not, whether the Compute and controller settings are comprehensive or not, and whether the sensors can quickly and accurately obtain road environment information or not, all of them would induce hardware safety risks.

2. Software reliability

Compared with traditional cars, the software development time for AVs is not long enough. Thus, it is lacking extensive supporting field data. The AV technology itself is still under development so it is not yet mature. The AV software system needs long-term reliability analysis. Therefore, its safety and stability still need long-term monitoring and validation.

3. Environmental security

When making driving decisions, AV still needs the correct driving of other on-road agents. Only when other agents have the correct driving behaviors, AV then will make its own correct decision and reasonable operation.

To ensure proper and safe functionality of AV, the development has to consider not only hardware but also software and user misuse at both component and vehicle levels. A holistic and traceable approach for risk analysis, risk mitigation, test specification, and validation is needed to orchestrate the behavior of single products in the function chain. For the AV Compute system, all embedded integrated circuits need to meet ASIL-C or even ASIL-D levels, and they shall be qualified by AEC-Q100 standards. AEC-Q100 is a set of stress test standards designed by AEC mainly for integrated circuit products for automotive applications. This specification is very important for improving product reliability and quality assurance. To prevent various conditions or potential failure states that may occur, AEC-Q100 conducts strict quality and reliability standard-based validation for each chip.

3 Advanced Computer System

3.1 Architecture Solution and Comparisons

The key to the success of an AV is to make a reliable decision in real-time quickly. Reasonable selection of AV Compute platforms to complete real-time large-scale sensor data processing, real-time driving prediction, and real-time control are essential to the safety, reliability, and durability of AV operation. During the early stages of AV development, most AV Compute solutions started with an Industrial Personal Computer (IPC) using the architectures based on Intel CPU + Nvidia GPU platform. IPC is a ruggedized and enhanced personal computer that can be used as an industrial controller to operate reliably in an industrial environment. The use of a fully sealed industrial chassis that meets the Electronic Industries Alliance (EIA) standard enhances the ability to resist electromagnetic interference. The CPU and various functional modules all use a plug-in structure with a soft locking lever to improve shock and vibration responses. The overall architecture design for AV needs to consider the requirements of ISO26262. The CPU, GPU, FPGA, and bus are all designed with redundancy to prevent single points faulty failure. Even when the IPC system fails, the MCU still can serve as a final guard bander, and directly send instructions to the vehicle Can bus to execute the emergency pull over or stop of the vehicle. At present time, this centralized architecture puts all computing tasks into one industrial computer. Therefore, the Compute size is large and the power consumption is high. But this architecture is very convenient. With the traditional X86 architecture, a computing platform can be built very quickly, and the card slot design is also convenient for hardware updates. As an example, Fig. 11 shows the Baidu Apollo AV adopted Neousys Nuvo-6108GC IPC for AV application. Nuvo-6108GC is the world's first industrial-grade Edge AI Computer supporting high-end graphics cards. It's designed to fuel emerging GPU-accelerated applications, such as autonomous driving, by accommodating Intel Xeon E5-2658V3 12-core CPU and Nvidia RTX 3070 GPU [6]. The peak CPU operating speed is 400 frame/s and it requires 400 W. Each GPU is capable of 8 TOPS performance computing and it requires 300 W. The whole system consists of two independent Computes. Therefore, the whole system can provide 64.5 TOPS Compute performance, and requires 3000 W. If both Computes operate at their maximum loading, a total of 5000 W will be required and it would produce excessive heat.

While providing high-performance data processing support, the Compute platform on a vehicle also needs to take into account issues such as power consumption, heat dissipation, and switch interface, which are equally important for continued safe driving.

- Power budget

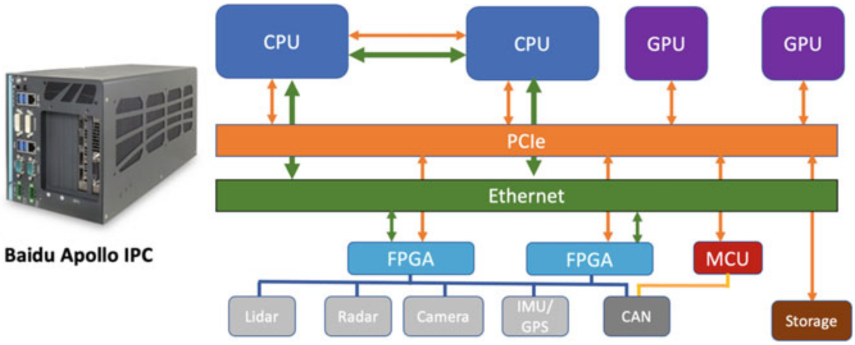


Fig. 11 Baidu Apollo IPC-based Compute platform block diagram

Vehicle platforms, especially electric vehicles, mostly provide a 12 or 48 V DC power supply. Compared with the traditional 220 V AC power supply for data center Computers, Compute used for vehicles needs to adapt to a 12 or 48 V DV power supply. For EV, the power consumption of Compute could reduce the vehicle mileage significantly. Therefore, the maximum power that an EV power supply can support is also an important issue for Compute design.

- Cooling solution

The heat dissipation solutions that can be used in vehicles are mainly air-cooled and liquid-cooled. Although the cost of air cooling is relatively low and the structure is simple, the disadvantage is that the heat dissipation efficiency is low and the noise is also relatively large. For most battery electric vehicles or hybrid vehicles, a liquid cooling loop for batteries already exists. Therefore, for battery-powered AV, the liquid cooling solution is a natural choice for a Compute mounted on a vehicle.

- Connector interface

In the existing AV Compute platform, different computing units are connected through an Ethernet Switch or PCIe Switch to transmit large amounts of data and complete coherent calculations. However, traditional ethernet mostly uses the RJ45 interface form, which will have a certain impact on the stability of the network during the long-term AV operation. In addition to the AI calculations, Compute also needs to coordinate and control various electronic control units (ECUs) and mechanical components in the vehicles to complete the driving control operations. This is achieved by the interconnection through the communication bus. Communication buses such as CAN, USB3.0, LIN, serial port, etc. are commonly used as interfaces to fulfill vehicle data sharing and effective transmission of control instructions from the Compute to vehicle ECUs and mechanical components. In general, with an increasing number of sensors on the vehicle, more interfaces are needed to connect all the required sensors.

- Real-time

AVs have very high requirements for system response in real-time. For example, in dangerous situations, the vehicle braking response time is directly related to the safety of vehicles, passengers, pedestrians, and roads. The braking reaction time includes not only the vehicle control time but the response time of the entire AD system including the time for perception, prediction, planning, and control. If the braking distance of the vehicle at a speed of 65 miles/h is to be less than 30 m, the overall response time of the system cannot exceed 100 ms, which is close to the response time of the best F1 players. Divide the response of AD into the requirements of each functional module of its Compute platform in real-time, including:

- Time for detection and precise positioning of surrounding targets: 15–20 ms.
- Time for data fusion and analysis of various sensors: 10–15 ms.
- Behavior and path planning time: 25–40 ms.

With the development of AD technology, the AD algorithm is constantly improving. After the algorithm is solidified, a dedicated ASIC chip or FPGA chip can be used to integrate the sensor and the algorithm to realize the edge calculation inside the sensor. This approach can further reduce the number of computing demands of the Compute to reduce power consumption and Compute size. Currently, there are two common schemes to construct an AV Compute platform:

1. Adopting off-the-shelf mature solutions

Some Compute platform solutions with different architecture designs for ADAS and AV applications are available. Designs are based on graphic processor unit (GPU), field-programmable gate arrays (FPTA), application-specific integrated circuits (ASIC), and digital signal processors (DSP). Among them, Nvidia's Xavier-based Drive PX2 and Drive AGX Pegasus platforms are two popular Compute solutions incorporating Nvidia's extensive experience in the field of Deep Learning for AV applications. Nvidia has been deeply involved in the field of autonomous driving for many years and is expected to become a new Tier 1 for the AV market. The main advantages of PX2 and AGX are:

- It is a complete system offering a turnkey solution. It is designed by the standards of vehicle regulations.
- Many sensors have been adapted already, especially image signal processor provided to adapt to many cameras. It can execute processes like demosaicing, noise reduction, auto exposure, autofocus, and the auto white balance at high speed and high quality.
- It has a relatively clear roadmap to facilitate subsequent iterative upgrades.

Nvidia Drive AGX is a powerful autonomous machine SoC [7]. Each Drive AGX consists of two Xavier SoCs and two Turing Tensor Core GPUs. Each Xavier has a custom 8-core Arm-based CPU. Drive AGX is capable of 320 trillion operations per second (TOPS) for Al computing and safe AV operation. The platform is designed and built for L4 and L5 autonomous systems.

GPU can provide tens to hundreds of times the CPU performance in terms of floating-point calculations and parallel calculations. Using GPUs to run machine learning models and perform localization and detection has greatly reduced the time consumed by CPUs. Relying on its powerful computing capabilities and driven by the rapid development of machine learning, GPUs are currently very popular in the deep learning chip market. Many car OEMs are also adopting GPUs as sensor data processing chips to develop AV. Therefore, GPUs have become the mainstream trend. However, the weakness of Nvidia's solution is that the performance of its CPU as a part of the SoC is still not powerful enough. Its CPU based on the Arm architecture has a main frequency of only 1.8 GHz and eight cores, which likely is difficult to meet the computing performance requirements for some AV applications.

Other commercially available solutions are Xilinx's Zynq UltraScale + ™ MPSoC ZCU104 product, TI's TDA3x, and Mobileye's EyeQ5 products as examples. Zynq is an FPGA-based SoC including 64-bit quad-core ARM Corte-A53 and dual-core ARM Cortex-R5. It is built by a 16 nm FinFET semiconductor technology node.

It is claimed to achieve 14 images/s/W for running convolutional neural network (CNN) tasks [8]. As a strong competitor of GPU in algorithm acceleration, FPGA has flexible hardware configuration, low power consumption, high-performance, and programmable advantages, which is very suitable for perceptual computing. More importantly, FPGAs are much cheaper than GPUs. In the case of energy consumption as a major concern, FPGAs have obvious performance versus energy consumption advantages over CPUs and GPUs. The low power consumption of FPGA makes it very suitable for sensor data preprocessing. In addition, the continuous development of perception algorithms means that the perception data processor needs to be constantly updated, and FPGAs have the advantage of hardware upgradability. One of the disadvantages of using FPGA is that it requires knowledge of hardware-level programming, which is difficult for many software developers. Therefore, FPGA is often considered an exclusive architecture for experts. However, some software platforms have emerged specifically for FPGA programming, which makes it possible for more software developers to use FPGAs. With the rapid popularization of the combination of FPGA and sensors, and the further optimization of vision, voice, and deep learning algorithms on FPGA, FPGA is very likely to gradually replace GPU and CPU as the mainstream AV chip, especially for perception.

TI TDA3x is a DSP-based solution for AV applications. It has two floating-point DSP cores with vision AccleerationPac to accelerate the image processing performance. Each TDA3x also has a dual Arm, Cortex-M4 image processor. The TDA3x SoC processor enables ADAS algorithms such as autonomous emergency braking (AEB), lane keeps assist, advanced cruise control (ACC), traffic sign recognition, pedestrian and object detection, forward collision warning, and back over prevention. It is for entry-to-mid-segment automobiles with L2 and L3 levels [9]. DSP can process a large amount of data with digital signals. It uses a Harvard architecture, that is, the processor is connected to two independent memory banks via two independent sets of buses, allowing the fetching and executing instructions in parallel. One memory bank holds program instructions and the other holds data. The next instruction can be fetched and decoded while the previous instruction is executed. This architecture greatly increases the speed of the microprocessor. In addition, it also allows transmission between processing space and data storage space, thus increasing the flexibility of the device. It not only has programmability, but its real-time running speed can reach tens of millions of complex instruction programs per second, far exceeding that of general-purpose microprocessors. Powerful data processing capabilities and a high operating speed are the two most commendable features of DSP. Because of its strong computing power, fast speed, small size, and high flexibility in software programming, it provides an effective way to engage in various complex applications.

Mobileye EyeQ5 is an ASIC-based SoC solution for AV applications. Its basic architecture is a combination of MIPS CPU core and vector acceleration unit. The overall computing performance is 24 TOPS with only a 10 W power budget. The power consumption is the brightest spot and obvious advantage of using EyeQ5. EyeQ5 is designed based on a start-of-art 7 nm FinFET IC technology, and this chip is aimed at L4 and L5 autonomous driving [10]. EyeQ5 is equipped with four heterogeneous fully programmed accelerators, which are optimized for proprietary algorithms, including computer vision, signal processing, and machine learning. EyeQ5 implements two PCIE ports at the same time to support multi-processor communication. This Compute architecture attempts to adapt the most suitable computing unit for each computing task. The diversity of hardware resources enables the fast operation of various applications and improves overall computing performance. However, the overall computing performance of EyeQ5 seems to be still far inferior to NVIDIA's solution.

2. Adopting self-designed, customized solutions

One customized solution is to take the x86 platform as the prototype, and directly integrate the Intel Xeon CPU and Nvidia's latest GPU architecture to achieve the highest computing performance. Another customized solution is to integrate GPU and FPGA to form a hybrid system by utilizing the benefits of short-latency, low power consumption, and high reliability of FPGA [11]. However, the disadvantage of those two solutions is that in addition to the need to customize the interfaces between CPU to GPU, and GPU to FPGA, heat dissipation, power distribution supply, sensor integration, functional safety, and connector interfaces are all required to be customized. Several companies like Tesla and Google/Waymo seek an edge in AV by making their own AI training silicon chips. Such a vertically integrated strategy for an AV company could be the ultimate solution for autonomous driving, which relies on deep neural networks (DNN) with huge computational demand. Google's Tensor Processing Unit (TPU) v3 is the latest ASIC-based AI accelerator mainly for DNN and machine learning [12]. It provides 420 TOPS computation performance for a single board. TPU is specially built for machine learning applications such as Google TensorFlow, which is designed to process more complex and powerful machine learning models in parallel at the price of reducing the accuracy of calculations. Compared to GPUs which are more suitable for machine learning and AI training, TPU is more suitable for analysis and decision-making after training. Tesla's D1 Dojo customized ASIC-based supercomputer chip can deliver 362 TOPS processing power [13]. Tesla places 25 of these chips on a single "training tile," and 120 tiles come together across several server cabinets (a total of more than an exaflop). The chip is built by a 7 nm semiconductor technology process and leaves the processor with an immense die size of 645mm^2, packing over 50 billion transistors.

How to choose the right Compute platform solution could depend on how to balance several metrics to achieve the best performance vs total life cost. According to Liangkai Liu et al. [14], 7 metrics shall be used to evaluate the computing system's effectiveness. They are accuracy, timeliness, power, cost, reliability, privacy, and security. The AV Compute platform integrates a variety of computing tasks with different attributes, such as precise geographic positioning and path planning, object recognition and detection based on deep learning, image preprocessing and feature extraction, sensor fusion and target tracking, etc. The performance and energy consumption ratios of these different computing tasks running on different hardware platforms are different. Generally speaking, for the convolution operation of object recognition and tracking, GPU has better performance and lower energy consumption than DSP and CPU. For feature extraction algorithms that generate positioning information, DSP is a better choice. Therefore, to improve the performance and energy consumption ratio of the AV Compute platform and reduce the calculation latency, it is very valuable to adopt heterogeneous computing architecture. Heterogeneous Compute selects appropriate hardware implementations for different computing tasks, makes full use of the advantages of different hardware platforms, and shields hardware diversity through a unified upper-layer software interface. Table 2 shows the comparison of GPU, DSP, FPGA, and ASIC-based Computes for architecture, performance, power consumption, and cost. Adopting a self-designed SoC with customized AI training silicon chips could be the game-changer for the future AV industry. Figure 12 shows the comparison of different SoC platforms suitable for different load tasks.

Table 2 Comparison of GPU, DSP, FPGA, and ASIC-based computes

Boards	Architecture	Performance	Power Consumption	Cost
Nvidia Drive AGX	GPU	320 TOPS	300 W	$30,000
Xilinx Zynq UltraScale + MPSoC	FPGA	14 images/s/W	–	$1295
TI TDA3x	DSP	–	30mW in 30fps	$549
Mobileye EyeQ5	ASIC	24 TOPS	10 W	$750
Google TPU v3	ASIC	420 TOPS	40 W	–
Tesla D1 Dojo	ASIC	362 TOPS	400 W	–
Qualcomm Snapdragon Ride L4/L5	ASIC	700 TOPS	130 W	–

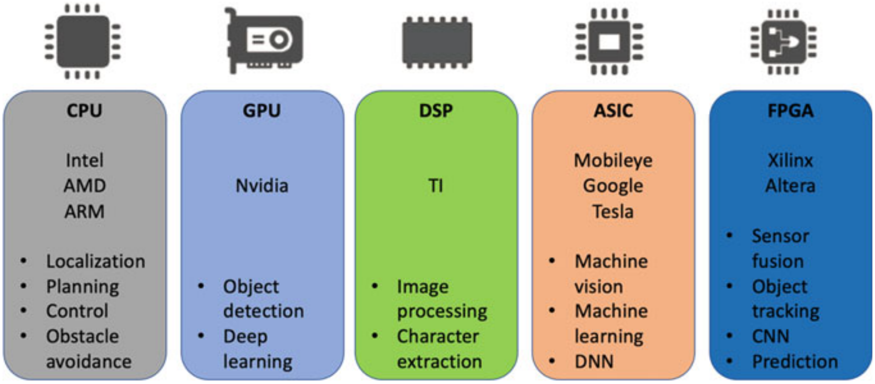
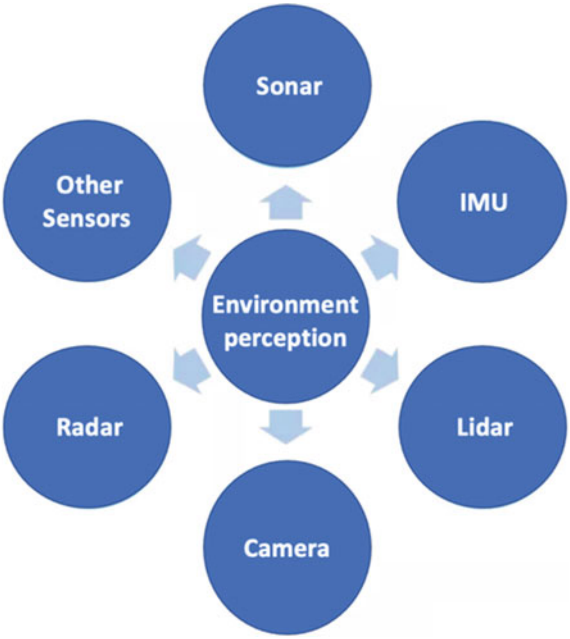


Fig. 12 Comparison of different SoC platforms suitable for different load tasks

3.2 Environment Perception Sensors

Sensors are to perceive the environment around the autonomous vehicle so AV can understand the environment correctly and make safe driving. To achieve the environmental perception, AV needs to obtain a large amount of surrounding information, specifically including the position and speed of surrounding vehicles, pedestrians, cyclists, and other moving agents, and possible behaviors of them at the next moment. AVs usually are equipped with cameras including IR camera, millimeter-wave radar, LiDAR, sonar, IMU, and GPS/GNSS to safely, accurately, and robustly collect such information as illustrated in Fig. 13. Moreover, typically there is more than one sensor of the same type. For example, to solve the blind spot and long-distance detection of LiDAR, both high-line-count radar, and low-line-count radar is generally used. As the horizontal viewing angle of a single camera is limited, multiple ($\geq$6) cameras are used to construct a 360° surround view. For millimeter-wave radar, due to the limitations of its horizontal viewing angle and distance factors, multiple radars are also used ($\geq$4). Furthermore, from a functional point of view, the redundancy between different types of sensors and from multiple same type sensors can improve the safety factor of the entire environment perception system. Perception generally is implemented by a chain of modules, comprising a sensor module, a microcontroller module, communication and networking infrastructure, and a Compute system. Perception needs to be robust and consistent across all use conditions. The requirements of perception are increasing with the increase in vehicle automation levels.

Fig. 13 AV environment perception hardware configuration



Technically speaking, first of all, the right sensing hardware needs to be chosen according to the needs of autonomous driving. This requires us to understand the advantages and disadvantages of different sensors. When sensors are available, we need to optimize the installation of these sensors to meet the needs of autonomous driving tasks. The environmental perception module is the most upstream of the automatic driving system. Through the analysis of data from different sensors, the analysis results are passed to the Compute module to realize the automatic driving of the vehicle. The sensing results of the environmental perception module include road dynamic and static target trajectories (such as vehicles, pedestrians, guardrails, etc.), traffic signal status (red, yellow, and green signal lights), traffic sign recognition related to traffic regulations, road lane line marking detection, and road surface detection.

According to the analysis of the output results of the environment perception module, we can get the key information related to the perception module task: 2D/3D target detection, scene semantic segmentation, instance segmentation, multi-sensor fusion, multi-target tracking, and trajectory prediction, as shown in Fig. 14. Although each technology can be designed independently, for the entire environment perception module, all different sensing technologies need to be fused to reduce the delay and memory storage consumption to achieve high efficiency, high precision, and low-cost objectives.

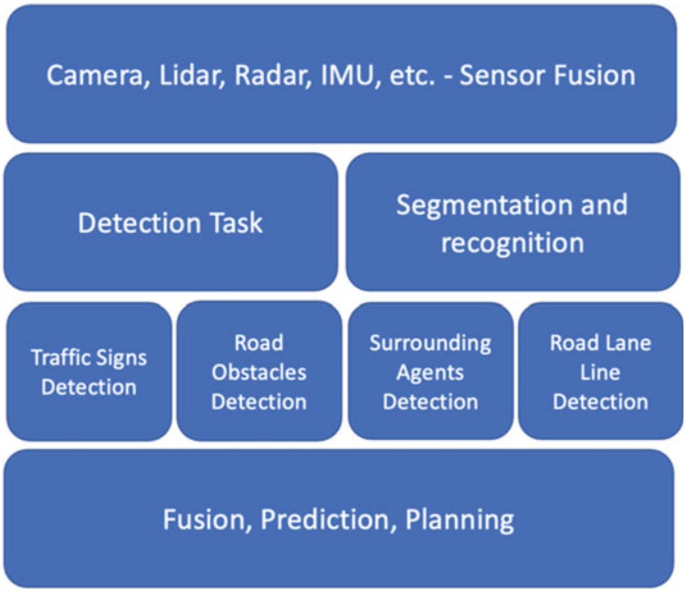

Camera, Lidar, Radar, IMU, etc. - Sensor Fusion
Detection Task
Segmentation and recognition
Traffic Signs Detection
Road Obstacles Detection
Surrounding Agents Detection
Road Lane Line Detection
Fusion, Prediction, Planning

Fig. 14 Environment perception module task flow chart

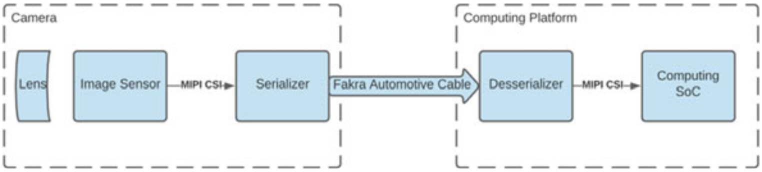


Fig. 15 The general camera block diagram and the data flow from the camera to the computing platform

3.2.1 Camera

Cameras are the most commonly used sensors to perceive the environment around the autonomous vehicle considering their relatively low cost and powerful usability. It is almost undisputedly adopted by all AV developers. The camera is the closest sensor type to the human eyes. The viewing range of the camera can vary from several centimeters to about 100 m. Also, they are often small, lightweight, and have low power consumption. The camera image provides a large amount of information at high frame rates, making it useful in tasks such as traffic light and pedestrian detection, lane tracking, object classification, traffic sign understanding, etc. Existing designs usually mount eight or more cameras around the vehicle to 360° to detect, recognize, and track objects. These cameras usually run at 60 Hz, making the total generated data a big challenge for Compute to real-time process such big data to get usable information about the environment. In addition, the quality of the camera's image is strongly affected by low lighting or bad weather conditions. The usability of the camera decreases significantly under heavy fog, rain, and snow. It is not good at long-distance vision as well.

In an AV, a camera in general consists of a Lens, image sensor, serializer, and power regulators. Different cameras may have different Lenses with different fields of view (FOV) and ranges. For example, 120-degree Len has a wide FOV but a short range. 30-degree Len has a narrow FOV but a long range. The image sensor is to detect and conveys information used to make an image by converting the variable attenuation of light waves into electronic signals. It has an active-pixel array-like 2MP or 8MP, and the Multiple Color filter array (CFA) like RGB Bayer, RCCB, etc. The Serializer is to convert the Mobile Industry Processor Interface (MIPI) Camera Serial Interface (CSI) to a single link. It can send the video frame data as well as receive the control data over the Fakra automotive cable as shown in Fig. 15.

3.2.2 Millimeter-Wave Radar

The radar is standard for Radio Detection and Ranging. It is a detection system to determine and calculate the distance and velocity using radio waves.

The Radar is mostly used to detect the distance to the objects around the vehicle. Once an object is detected too close to the vehicle, there may be a danger of collision so the autonomous vehicle should take action as soon as possible. Examples of actions are braking or turning to avoid a potential collision. The data generated by the Radar is not needed to process too much. It will feed into the Compute directly. The Compute could implement the emergency action, such as an autonomous emergency brake. In the autonomous vehicle, the Radars are deployed in different areas such as the front, rear, front-right, and front-left of the vehicle. The front radars are typically mid and long-range radars responsible for autonomous emergency braking (AEB) and adaptive cruise control (ACC). The side radars are typically short-range radars to handle the requirements of blind-spot detection (BSD), front/rear cross-traffic alert (F/RCTA), and lane-change assist (LCA) [15].

Generally, 24 and 77 GHz frequencies are used in the radar system. 24 GHz includes industrial, scientific, and medical banks from 24 to 24.25 GHz. For 77 GHz, it has a 76–77 GHz band available for radar application. Compared to the 24 GHz frequency, the 77 GHz frequency has a wider bandwidth available, which improves the range resolution and accuracy significantly [15]. High range resolution results in better separations of objects. It also results in a better minimum distance detection. A shorter minimum distance is very important for some AV functions such as AEB. The higher frequency also can provide a better velocity resolution and accuracy. Another benefit of higher frequency is that the radar size can be made smaller. Radar can work under any weather conditions, which makes it indispensable. It has its unique capability to penetrate dust, fog, rain, and snow, therefore has a firm foothold on the AV sensor module.

3.2.3 LiDAR

LiDAR is the heart of object detection for most of the existing AVs. The full name of LiDAR is light detection and ranging or laser imaging, detection, and ranging. It can be used to calculate the distance. The difference between radar and LiDAR is that LiDAR has the laser generator and receiver inside. It sends millions of light pulses per second in a well-designed pattern to the surface of an object and measures the reflection time return to the receiver. With its rotating axis, it can create a dynamic, three-dimensional map of the environment. In an AV, the LiDAR is commonly used to detect objects and pedestrians, determine the distance, make high-definition maps, and localize a vehicle aligned with the high-definition map [16].

Compared to the Camera, LiDAR generates a 3-dimensional cloud image of objects instead of a 2-dimensional image. It has a larger sensing range, and the performance is less impacted by bad weather and a low lighting environment. Point cloud output from the LiDAR provides the data for autonomous computing to determine where objects exist in the environment and where the vehicle is in relation to those objects. It can generate a lot of data for vehicle Compute to process in real-time.

3.2.4 Ultrasonic Sensor

The ultrasonic sensor is a kind of radar that is widely used in vehicles already. It is often installed on the bumper at the rear, front, and sides of the car for the reversing assist and parking assist functions as shown in Fig. 16. Its working principle is to transmit high-frequency sound waves to gauge the distance between objects within close range. The ultrasonic sensor shows good performance in bad weather and a low lighting environment. But ultrasonic radar's maximum range is only about 20 m so it is not suitable for long-distance ranging. Ultrasonic radars can be used to complement other vehicle sensors, including radars, cameras, and LiDARs, to get a full picture of the immediate surroundings of a vehicle.

Ultrasonic sensors are generally composed of an ultrasonic transmitter, an ultrasonic receiver, a timer, a temperature sensor, etc. The distance measurement principle is to use the propagation speed of ultrasonic waves in the air to be known (344 m/s at 20 °C) and measure the sound waves in the air. After the launch, the time when the obstacle is reflected is calculated, and the actual distance from the launch point to the obstacle is calculated according to the time difference between the launch and the reception. It can be seen that the principle of ultrasonic ranging is the same as that of radar.

The formula of ranging is expressed as:

$$L = C \times T \tag{1}$$

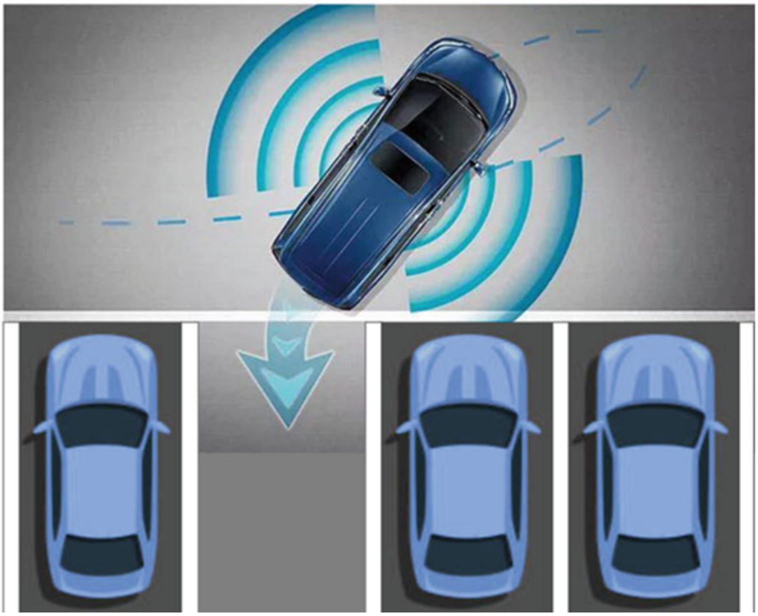


Fig. 16 Sonar-assisted parking illustration

where L is the measured distance length, C is the propagation speed of ultrasonic waves in the air, and T is the time difference of the measured distance propagation (T is half of the value of the time from emission to reception).

Table 3 illustrates a comparison of sensors, including camera, IR camera, radar, LiDAR, IMU, and ultrasonic sensors. The range of sensing distance for human eyes is 0–200 m. Human vision is poor during bad weather and low lighting condition. From the comparison, it shall be concluded that although humans have strength in the sensing range and show more advantaged functionality scenarios than any sensor, the combination of all the sensors can do a better job than human beings, especially in bad weather and low lighting conditions.

3.3 System on Chip (SoC)

A system on chip is a chip that integrates most components of a computer. The components consist of multiple cores of a central processing unit (CPU), graphics processing unit (GPU), artificial intelligence (AI) unit, multiple levels of cache, input/output ports of memory, high-speed I/O, internal connection between CPU, GPU, AI unit, memory, high-speed I/O, and the power management unit [17]. To support real-time data processing from various sensors, a powerful Compute is essential to AVs' success.

3.3.1 ASICS

In autonomous driving, Application-Specific Integrated Circuit (ASIC) consists of multiple units like the common CPU, the GPU, the unit for the deep learning, and the memory controller that connects the external memory through Low Power Double Data Rate 4 (LPDDR4). In general, different storages connect to the CPU. Flash is used to store the firmware. eMMC is for an application that needs more space. The UFS has a large size capability to store big data like a high-definition map. The camera data is transferred to the CPU through the Deserializer. The Deserializer converts the interface from Gigabit Multimedia Serial Link (GMSL) or Flat Panel Display Link (FPDlink) to the CSI. The LiDAR connects to the CPU through Ethernet Switch. The automotive 1GBase-T1 interface is from the Ethernet Switch. Sometimes, an ethernet physical layer (PHY) is needed to convert the 1GBase-T1 to other buses like Reduced Gigabit Media-independent Interface (RGMII) or Serial Gigabit Media-independent Interface (SGMII) if the Ethernet Switch can't support the 1GBase-T1.

The Micro Controller Unit (MCU) is used to manage the board. For example, monitor the health of the board like the voltage, current, and temperature, control the power on/off and reset. The radar data is going to the MCU through the Controller Area Network (CAN) bus. Ethernet Switch is used to communicate between CPU and MCU. The radar data is transferred to the CPU from MCU through the Ethernet Switch.

Table 3 A comparison of sensors, including camera, IR camera, radar, LiDAR, IMU, and ultrasonic sensors

	Advantages	Disadvantages	Detection distance (m)	Functionality
Lidar	High accuracy, wide detection range, 3D model of the surrounding environment, speed and distance estimates	Can be affected by bad weather such as rain, snow and fog. Less matured technology with high cost	200	Obstacle detection and recognition, road agents' speed and distance measurements, 3D model of surrounding environment
Camera	Identify the geometry and color of the objects. Recognize texts and symbols. Mature technology with low cost	Affected by changes in light, vulnerable to bad weather such as rain, snow and fog. Can't measure distance accurately	100	Obstacle detection and recognition, lane tracking, auxiliary positioning, road information understanding, map construction
Radar	Strong penetrating ability to smoke and dust, strong anti-interference, high accuracy of speed and distance estimates	Unable to apply visual recognition, such as size and shape of objects, Detection range is narrower than lidar	200	Obstacle detection—Medium and long distance
Ultrasonic sensor	Matured technology, low cost, strong antiinterference, less affected by weather	Poor measurement accuracy, small measurement range, short distance	3	Obstacle detection—Short distance, useful for BSM, parking assistance, and reversing assistance
IR/Thermal camera	Good vision at night or blind sun glare, reliable detect persons/animals	High cost	200	IR camera sees heat, reducing the impact of occlusion on classification of pedestrians
GPS/IMU	Localize vehicle position by combining satellite triangulation and inertial navigation	vulnerable to building and tunnel interferences, high cost	10	Localization

Figure 17 illustrates an ASIC-based Compute block diagram.

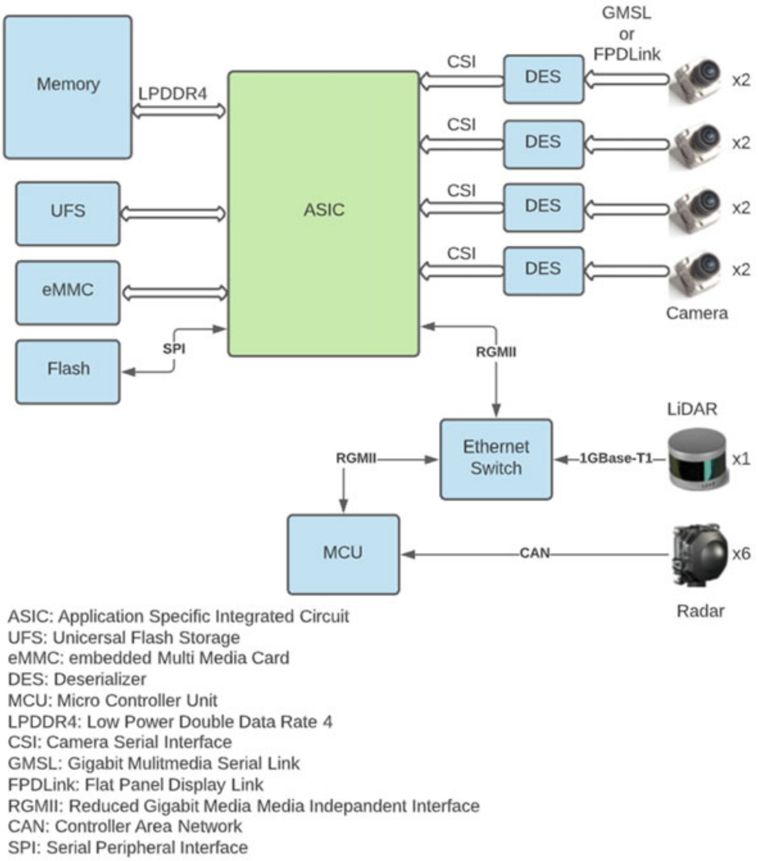

Memory
LPDDR4
UFS
eMMC
Flash
SPI
ASIC
CSI
CSI
CSI
CSI
DES
DES
DES
DES
GMSL
or
FPDLink
x2
x2
x2
x2
Camera
RGMII
LiDAR
Ethernet Switch
1GBase-T1
x1
RGMII
MCU
CAN
x6
Radar
ASIC: Application Specific Integrated Circuit
UFS: Unicersal Flash Storage
eMMC: embedded Multi Media Card
DES: Deserializer
MCU: Micro Controller Unit
LPDDR4: Low Power Double Data Rate 4
CSI: Camera Serial Interface
GMSL: Gigabit Mulitmedia Serial Link
FPDLink: Flat Panel Display Link
RGMII: Reduced Gigabit Media Media Indepandent Interface
CAN: Controller Area Network
SPI: Serial Peripheral Interface

Fig. 17 ASIC-based block diagram

3.3.2 x86

x86 is a family of computer processing instruction set architectures (ISA) developed by Intel. ISA is computer architecture. It is an abstract model of a computer that defines the data types, registers supported, how to manage the memory and memory consistency, and how to access the input/output model. It also specifies the behavior of the machine code consisting of instructions. It is a low-level programming language used to control a computer processor [18].

The 8086 was developed in 1978 for 16-bit processors. Many additions and extensions have been added to the x86 instruction set over the years. In 1985, it grew to a 32-bit instruction set of the 80,386. The bit in both 32-bit and 16-bit is 32 or 16 binary digits. Today, an x86 microprocessor is used in almost any type of computer. It is also used as the computing platform in AV. In the computing platform used for AV, there are CPU and GPU. CPU performs basic arithmetic, logic, and control. The GPU is more focusing on the artificial intelligence algorithm. Depending on the performance requirement, the computing platform can have a single CPU or dual-CPU solution.

In general, a single CPU solution has 1 CPU and GPU. CPU consists of the multiple cores, cache, memory controller which connects the memory devices, input/output controllers like the Peripheral Component Interconnect Express (PCIe) root complex which connects to the GPU and ethernet controller, and 10Gbps or 1Gbps ethernet outputs from ethernet controller. The camera, LiDAR, or radar data can be transferred to the CPU and GPU through the ethernet interface.

Platform Controller Hub (PCH) is Intel's signal chipset. It is the successor to the Intel Hub architecture that used two chips—northbridge and southbridge instead. It includes a clocking generator, PCIe interface, and storage interfaces like SATA and USB hub. The different storage devices can be connected by different interfaces such as the PCIe based hard disk or M.2 to PCH through PCIe and the SATA based hard disk or M.2 to PCH through SATA. The Direct Media Interface (DMI) is an interface that connects the CPU and PCH.

The firmware to perform the hardware initialization during the booting process is called the Basic Input/Output System (BIOS). It stores in the flash connected to PCH. It provides the runtime service for operating systems and programs.

The platform needs to be managed by monitoring the health of the system, controlling the power on/off, and resetting. It is done by the Baseboard Management Controller (BMC). The BMC has its memory, and flash with firmware. The 1Gbps ethernet to BMC can be used for remote access.

Figure 18 illustrates an x86-based Compute with a single CPU block diagram.

Besides the single CPU solution, to have more performance for the AI algorithm, there is a dual CPUs solution with dual GPUs. It can provide more CPU and GPU cores to increase workloads and performances. The communication between CPUs uses the high-speed interface Ultra Path Interconnect (UPI) to provide the high bandwidth between CPUs.

Figure 19 illustrates an × 86-based Compute with a dual-CPU block diagram.

3.4 Memory

The Synchronous Dynamic Random-Access Memory (SDRAM) is mostly used in the autonomous computing platform. The read and write operation is through an interface synchronous with the system bus. The data and control signals are aligned with the clock signal.

There are different standards of the SDRAM such as Single Data Rate (SDR) SDRAM and Double Data Rate (DDR) SDRAM. SDR reads/writes one time in one clock cycle. DDR SDRAM is the next-generation of SDR SDRAM. The data can be transferred two times in one clock cycle, at the rising and falling edges of the clock signal. Thus, it achieves higher bandwidth as compared with the SDR. It doubles the data rate without increasing the frequency of the clock.

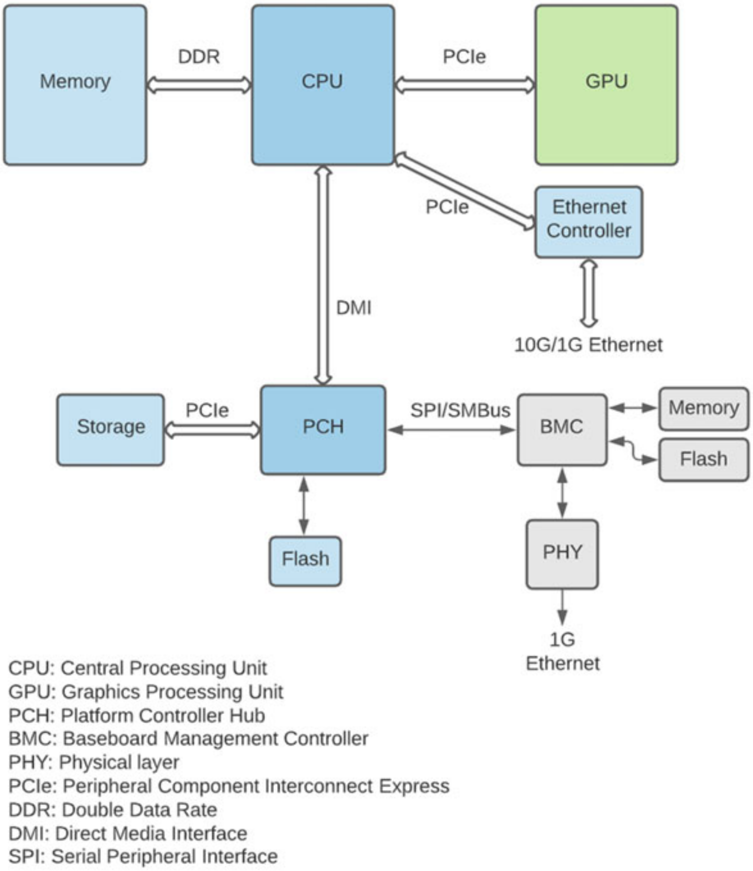

Memory
DDR
CPU
PCIe
GPU
PCIe
Ethernet Controller
DMI
10G/1G Ethernet
Storage
PCIe
PCH
SPI/SMBus
BMC
Memory
Flash
Flash
PHY
1G Ethernet
CPU: Central Processing Unit
GPU: Graphics Processing Unit
PCH: Platform Controller Hub
BMC: Baseboard Management Controller
PHY: Physical layer
PCIe: Peripheral Component Interconnect Express
DDR: Double Data Rate
DMI: Direct Media Interface
SPI: Serial Peripheral Interface

Fig. 18 x86 block diagram with single CPU

DDR2 SDRAM has a data rate twice as fast as DDR SDRAM. It is achieved by doubling the prefetch buffer to 4 bits. The DDR3 SDRAM has an 8-bit prefetch buffer. As a result, the data rate doubles based on the DDR2. The DDR3 SDRAM reduces power consumption by lowering operation voltage. DDR4 SDRAM lowers its operating voltage. It adds four new back groups to achieve a higher data rate. Table 4 compares different SDRAM.

DDR SDRAM is mostly used in the x86 platform. In the ASIC-based platform, the Low Power Double Data Rate (LPDDR) like LPDDR4 is used to save power.

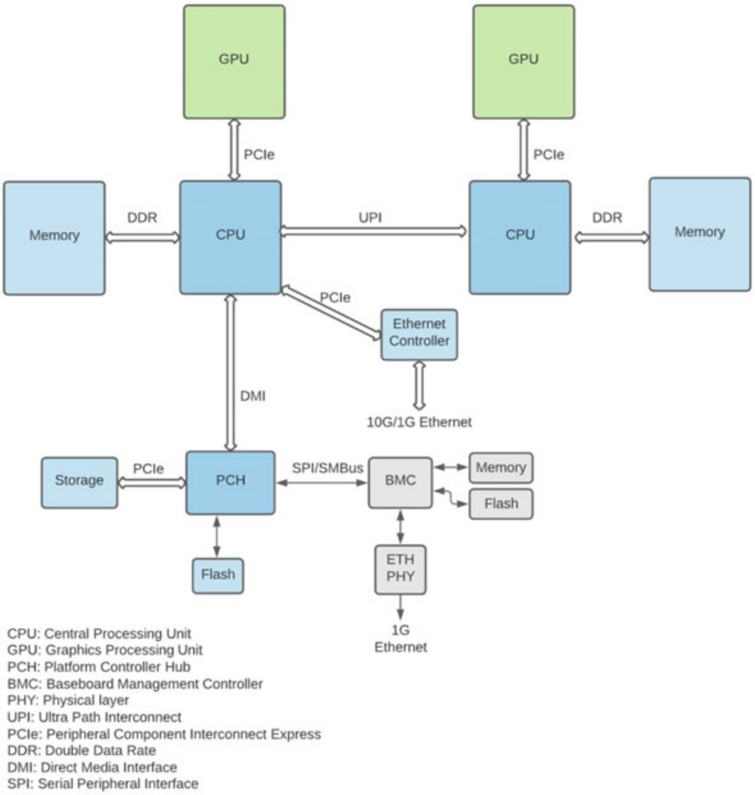

GPU
GPU
PCIe
PCIe
Memory
DDR
CPU
UPI
CPU
DDR
Memory
PCIe
Ethernet Controller
DMI
10G/1G Ethernet
Storage
PCIe
PCH
SPI/SMBus
BMC
Memory
Flash
Flash
ETH PHY
1G Ethernet
CPU: Central Processing Unit
GPU: Graphics Processing Unit
PCH: Platform Controller Hub
BMC: Baseboard Management Controller
PHY: Physical layer
UPI: Ultra Path Interconnect
PCIe: Peripheral Component Interconnect Express
DDR: Double Data Rate
DMI: Direct Media Interface
SPI: Serial Peripheral Interface

Fig. 19 x86 block diagram with dual CPUs

Table 4 Comparison of different SDRAM

SDRAM standard	Internal data rate (MHz)	Interface clock data rate (MHz)	Prefetch	Interface data rate (MT/s)	Operation voltage (V)
SDR	100–166	100–166	1*n*	100–166	3.3
DDR	133–200	133–200	2*n*	266–400	2.5
DDR2	133–200	266–400	4*n*	533–800	1.8
DDR3	133–200	533–800	8*n*	1066–1600	1.5
DDR4	133–200	1066–1600	8*n*	2133–3200	1.2

3.5 Storage

In autonomous driving Compute, there are different kinds of storage devices for different purposes. For example, Serial Peripheral Interface (SPI)/Quad Serial

Peripheral Interface (QSPI) Flash, Embedded Multimedia Card (eMMC), Universal Flash Storage (UFS), and Solid State Drive (SSD) are commonly used.

SPI/QSPI Flash is to store the SoC firmware. During the system boot-up, to initialize components inside the SoC as well as provide the runtime service for the operating system and programs, the SoC loads and executes the firmware from this Flash through SPI/QSPI bus.

eMMC is similar to the SPI/QSPI Flash. The SoC firmware can be stored in it. Generally, eMMC has a larger capability. Except for the firmware, it can store the operating system and applications running in the operating system.

UFS has a larger capacity and higher bandwidth with a differential signal interface. Except for storing the SoC firmware, it can store the high-definition map as well as the training data for autonomous driving Artificial Intelligence Algorithm.

In autonomous driving, the data captured from the cameras, radar, and LiDAR is huge. Collecting all the sensors data as much data as possible will help to train the AI model. The SSD has a big capacity. It normally is used to collect the data captured from camera, radar, and LiDAR.

For AV applications, memory must have enough read and write endurance to match the excessive data logging requirements of a vehicle over its lifetime. Consider an SSD with an endurance of 10^6 accesses before a cell typically degrades. At a record rate of 0.2 s, an SSD block would wear out in less than three days. To extend the effective endurance of SSD, wear-leveling has been used. Wear-leveling involves tracking the reliability of each memory block and moving data to a new block when the current block begins to experience errors beyond a certain threshold.

3.6 Network

In the autonomous vehicle, mainly two networks are used. One is the ethernet network. The communication between the controlling process and the computing process is through this network. LiDAR data is captured and transferred through an ethernet network. In the x86 platform, the Camera data also uses the ethernet to transfer to CPUs. It offers high bandwidth interfaces like 1 or 10 Gbps. Normally, the automotive-grade Ethernet Switch chip and the media convertor physical layer (PHY) are used in the Ethernet network.

The other network is the controller area network (CAN). It is mainly used in the communication between electronic control units (ECU). Radar data is captured and transferred to the Compute through the CAN network.

With the development of AD technology, more and more data need to be transferred between computing processer, control processor, and different ECUs. More Camera, LiDAR, and Radar data are captured and transferred. The bandwidth of the network inside the vehicle becomes more and more important. The ethernet network will be more popular in the AD platform.

3.7 Real-Time Operating System

To make AD safe, perception, prediction, and deciding in real-time are important. So, a real-time operating system (RTOS) is used in an AV. RTOS is fast and responsive. It is intended to run a real-time application. The RTOS is mainly used in many embedded systems. It requires real-time processing. Due to the hardware resource, performance and efficiency are high priorities. The scheduler in an RTOS is designed to provide a predictable (normally described as deterministic) execution pattern. This is useful for embedded systems.

RTLinux and QNX are two popular RTOS systems. RTLinux runs all the operation system (OS) components in the kernel space, including memory management, file management, networking, and drivers. It can improve performance. It also can respond faster and more reliably. The downside is that since all the components are in the kernel space, a single failure can cause the OS to crash [19].

Compared with RTLinux, QNX has a core RTOS kernel to access the whole system. It allocates the memory for other processes. All the other components run in their own isolated space. It improves reliability and security. Also, it isolates the error in one component from other components.

3.8 Management, Failure Detection, and Diagnostics

Safety is very critical in an autonomous vehicle. To make the vehicle safe, failure detection, diagnostics, and platform management become impotent. There are different kinds of failure detection to cover the autonomous computing platform. The voltage, current, and temperature monitoring is the hardware-level fault detection mechanism. Run time diagnostics like CPU internal self-test, memory bit error detection, and storage bit error detection is important to detect and report any failure that happened.

The MCU in the ASIC-based platform, as well as the BMC in the x86-based platform, are mainly used to detect the failure, manage the power on/off, and reset other domains like the computing domain, network domain, camera domain, etc. They run diagnostics applications to monitor critical functions.

3.9 Security and Middleware

For AV, security is very important. It is extremely dangerous for any AVs to get on the road without meeting the rigid security requirements. At present, there are a variety of methods for AV to be attacked and the attacks can happen at every level of the AD system, including sensors, Compute system, control system, and vehicle networking communication system. First of all, the attack on the sensors does not need to enter the AD system. Therefore, the technical threshold of this external attack method is quite low, that is, it is simple and direct. Second, if hackers enter the AD system remotely, they can crash the system to cease the vehicle operation. They also can directly steal sensitive vehicle information. Third, if hackers enter the AV control system, they can directly manipulate and control the mechanical components so that they can hijack the vehicle to make terroristic attacks, which is extremely dangerous. Fourth, the Internet of Vehicles links different AVs and the central cloud platform system. Hijacking the Internet of Vehicles communication system can also cause communication chaos in the AVs. Therefore, car interconnection through V2V and V2X can bring great convenience to users, but it also exposes vehicle systems to the risks brought by the internet. The security requirements of AV become more and more challenging due to:

- More and more networked and intelligent vehicle controllers used: BCM, IMMO, PKE/RKE, TBOX, IVI, ADAS, etc.
- More and more networked and intelligent vehicle sensors are used: TPMS, Camera, LIDAR, RADAR, etc.
- More and more input ports, interface layers, and codes used: OBD, CAN, wireless, mobile phones, cloud, etc.
- More and more cloud control, AV remote control used: remote management, frequent OTA, remote driving, remote mobile phone control, etc.
- More and more vehicle communication protocols are used: 4G/5G, Wi-Fi, Bluetooth, NFC, RFID, etc.

The automotive security categories can be classified as component/sensor security, network security, and control security as shown in Table 5. The sensor security includes jamming or spoofing the sensors like Cameras, Radars, LiDARs, and GPSs. Network security includes attacking the network and sending the wrong message to the network. Multiple services are running in the autonomous driving system. To facilitate the dependencies between the services. The middleware is impotent to simplify the communication between different autonomous driving services. It is on top of the RTOS.

SAE's J3061 procedure "Cyber-physical Convergence System Cyber Security Guidelines" released in January 2016 is the first guidance document formulated for automotive cyber security. The supporting document J3101 "Hardware Protection Safety Requirements for Road Vehicle Applications" allows designers to take some measures to provide multiple protections for vehicles, such as storing the verification key in the protected area of the microcontroller. For AV, safety, and security, in general, are considered the top items in the development of AD technology. To reduce and avoid the risks in actual road operation, adequate simulation, bench, and closed field testing and verification must be done before actual road deployment.

Table 5 Classification of component and system-level security

Security category		Security content
Component security		Authentication protocols such as verification key Secure start and communication Security certification and upgrade Security monitoring Embedded with TEE and HSM Intrusion detection system Hardware root of trust
Vehicle information system security	In-vehicle network security	Sub-networks Gateways Visit control Protocol encryption and authorization Abnormal vehicle control detection
	OS and software control security	Anti-flash FW Prevent the denial of service and attacks Anti-sniffing Protocol authorization and management Data encryption

4 Electrical Functional and Reliability Validation

The automotive industry is a highly regulated industry across the globe. To survive in the market for a long period, automotive OEMs and component manufacturers need to be constantly innovative in terms of quality, durability, reliability, and safety. They also need to ensure that the system and components of the automobiles must function properly throughout their working life. With the fast-growing innovations in the industry such as EV and AV, new testing solutions and methodologies are constantly needed accordingly. From a testing and validation perspective, AV Compute brings together two previously separate validation standards: the automotive industry standards such as GMW3172 and ISO16750 standards, and the electrical industry standards such as IEC and JEDEC standards. The benefits of conducting automotive level electrical functional and reliability validation are (1) ensure the electrical safety of users during the product operation, (2) verify that the products comply with the state-of-art industry standards, (3) evaluate the conformance, interoperability, and electromagnetic compatibility, (4) validate product durability and reliability along with the cost of the warranty. Functional and reliability tests are the ways to identify manufacturing faults and design weaknesses that could compromise the electrical safety and durability of a Compute out in the field. Thorough functional and reliability tests protect against the risk of safety and reliability issues so that Compute can be used for its intended purpose with minimal chance of accidents and failures occurring.

In general, electrical AV electronics could present significant challenges for automotive testing. High currents and voltages are present in the form of complex signals both as stimuli and as measurements. Much of the circuitry involves asynchronous timing and events. In this section, we will introduce Compute electrical functional test and reliability validation based on GMW3172 and ISO16750. GMW3172 and ISO16750 are automotive industry well-established and accepted electronic components testing standards. Those standards have been used to systematically qualify electronic components for the life cycle of all GM and other vehicle OEM manufactured vehicles with a set of testing environmental conditions and pass/fail requirements. In the process of forming the standards, various environmental factors, world climate conditions, vehicle types, vehicle operating conditions and working modes, product life cycle, vehicle power supply voltage, and component installation locations in the vehicle were taken into consideration. We are going to describe the overall Compute validation testing in three categories: EE functional testing, reliability validation testing, and EMC/ESD compliance testing.

4.1 Automotive Level EE Functional Tests

4.1.1 Five-Point Functional/parameter Check

For fully functional/parameter testing, a 5-point check is required. This test is to let Compute be exposed to three temperatures and three voltages. The operating types are 2.1 defined by GMW3172, Compute functions are not activated to confirm functionality in sleep mode/off mode, and 3.2 defined by GMW3172, Compute with electric operation and control in typical operating mode. The five points are defined as:

1. T_{min}, V_{min}
2. T_{min}, V_{max}
3. T_{room}, V_{nom}, where U_{nom} is V_{B} for Operating Type 2.1, and U_{A} for Operating Type 3.2
4. T_{max}, V_{min}
5. T_{max}, V_{max}

The test condition for this test is:

(a) Step 1:

Test temperature for Chamber: T_{min}
Testing time: 75 min
Operating type: 3.2
Test voltage: 9 VDC and 18 VDC

(b) Step 2:

Test temperature for Chamber: 23 °C (Room Temperature)
Testing time: 75 min
Operating type: 2.1 for 12 VDC test voltage -> 3.2 for 14 V test voltage
Test voltage: 14 VDC

(c) Step 3:

Test temperature for Chamber: $T_{\max}$
Testing time: 75 min
Operating type: 3.2
Test voltage: 9 VDC and 18VDC.

4.1.2 One-Point Functional/Parameter Check

One-point functional/parametric check is to verify Compute full functionality under one single temperature and one single voltage condition. It is a special case of the 5-point check. The 1-point check shall be performed at room temperature under a nominal voltage unless otherwise specified. The temperature shall be stabilized before the 1-point Functional/Parametric Check.

4.1.3 Continuous Monitoring

Continuous monitoring shall monitor the functional status of the Compute during the test environment continuously. Continuous monitoring shall record all input and output signals, serial data messages, all transmitted packets, voltages, frequencies, powers, and temperatures from all critical components, and erroneous Input/Output (I/O) commands or states.

4.1.4 Electrical Load Testing

In Table 6, selected electrical load testing items are listed specifically for Compute used for battery-powered AV.

4.2 Reliability Validation Tests Based on AV Mission Profiles

The reliability of AV hardware, especially the Compute, is one of the critical enablers for AV business. Reliability is defined as the probability that a product will perform its required function for a given time at the desired confidence level under the specified use conditions. The failure of a Compute is defined as the termination of the ability of the Compute to perform a required function. The function usually is specified

Table 6 List of electrical load tests for compute

Test item	Standards	Purpose	Requirement
Direct current supply voltage	ISO16750-2	Validate Compute functionality at minimum and maximum input voltages	All functions of the device/system perform as designed during and after the test
Overvoltage	GMW3172	Verify Compute immunity to overvoltage conditions	One or more functions of a device/system do not perform as designed during the test but return automatically to normal operation after the test
State change waveform characterization	GMW3172	Verify that the Compute behaves adequately during state changes (e.g., Compute cold start, shutdown, etc.)	All functions of the device/system perform as designed during and after the test
Reverse polarity	GMW3172	Check the ability of the Compute to withstand against the connection of a reversed battery in case of using an auxiliary starting device	One or more functions of a device/system do not perform as designed during the test but return automatically to normal operation after the test
Jump start	GMW3172	Verify the Compute's immunity to positive overvoltage. This condition can be caused by a double-battery start assist	One or more functions of a device/system do not perform as designed during the test but return automatically to normal operation after the test
Slow decrease and increase of supply voltage	ISO16750-2	Simulate a gradual discharge and recharge of the battery	One or more functions of a device/system do not perform as designed during the test but return automatically to normal operation after the test
Ground reference and supply offset	ISO 16750-2	Verify reliable operation of the Compute if two or more power supply paths exist. For instance, a component may have a power ground and a signal ground that are output on different circuits	All functions of the device/system perform as designed during and after the test

(continued)

Table 6 (continued)

Test item	Standards	Purpose	Requirement
Parasitic current	GMW3172	Verify that the compute's power consumption complies with the specification for Ignition OFF state. This is to support power management and engine start ability following long-term storage and parking conditions	The maximum allowable average parasitic current shall be 0.125 mA. Analyze the stored current waveforms for any random fluctuations. Unintentional wakeups are not allowed
Power supply interruptions	GMW3172	Verify the proper reset behavior of the compute. This test shall also be used for all microprocessor-based components to quantify the robustness of the design to sustain short-duration low voltage dwells	All functions of the device/system perform as designed during and after the test
Battery voltage dropout	GMW3172	Verify the compute's immunity to voltage decrease and increase that occur during discharge and charging of the vehicle battery	There shall be no inadvertent behavior during the transitions. Different functional statuses are required pending on the zone
Pulse superimposed voltage	GMW3172	Verify the compute's immunity to supply voltage pulses that occur on battery supply in the normal operating voltage range	All functions of the device/system perform as designed during and after the test
Intermittent short circuit to battery and to Ground for I/O	GMW3172	Verify the Compute's immunity to intermittent short circuit events on Input/Output (I/O) lines as well as the component's ability to recover automatically from these events	One or more functions of a device/system do not perform as designed during the test but return automatically to normal operation after the test
Continuous Short circuit to battery and to ground for I/O	GMW3172	Verify the Compute's immunity to continuous short circuit events on Input/Output (I/O) lines	One or more functions of a device/system do not perform as designed during the test but return automatically to normal operation after the test

(continued)

Table 6 (continued)

Test item	Standards	Purpose	Requirement
Open circuit—Single-line interruption	ISO16750-2	Verify that the Compute is immune to single-line open circuit conditions	One or more functions of a device/system do not perform as designed during the test but return automatically to normal operation after the test
Open circuit—Multiple line interruption	ISO16750-2	Ensure functional status as defined in the specification of the Compute when the Compute is subjected to a rapid multiple line interruption	One or more functions of a device/system do not perform as designed during the test but return automatically to normal operation after the test
Ground offset	GMW3172	Verify the Compute's ability to function properly when subjected to ground offsets	All functions of the device/system perform as designed during and after the test
Discrete digital input threshold voltage	GMW3172	Verify the capability of discrete digital input circuits (including switch interfaces) to withstand minor voltage fluctuations without causing a change of active/inactive state	All discrete digital input interfaces shall be able to correctly detect the logic levels
Over load—All circuits	GMW3172	Verify the component's ability to withstand overload situations or open circuits in a safe manner	If an output is over-current protected: one or more functions of the component do not perform as designed during the test and do not return to normal operation after the test until the component is reset by any "operator/use" action If an output is not over-current protected: one or more functions of the component do not perform as designed during and after the test and cannot be returned to proper operation without repairing or replacing the component

(continued)

Table 6 (continued)

Test item	Standards	Purpose	Requirement
Insulation resistance	GMW3172	Verify the component's immunity to loss of insulation	One or more functions of a device/system do not perform as designed during the test but return automatically to normal operation after the test The insulation resistance shall be greater than 10 MΩ
Power offset	GMW3172	Verify the component's ability to function properly when subjected to power offsets	All functions of the device/system perform as designed during and after the test

in Compute technical specification or operation manual. Failures could be the loss of whole or partial functions either permanently or intermittently, the deterioration of the whole or partial function over time, and a field surprise. The purposes of performing reliability validation tests on the Compute are to quantify the factors limiting the life of a Compute significantly less than the total expected life and to provide guidelines for design for reliability (DfR) and field replacement. There are two important concepts generally involved with reliability validation tests, the acceleration concept and the statistics concept. Usually, the products last so long that their lifetime can't be verified by direct measurement, therefore, accelerated tests, as well as the extrapolation procedures, are mandatory for reliability engineering. On the other hand, the reliability test is a sampling test, which is not testing the entire population. Thus, the true probability of failure can't be obtained. The probability of failure of a population can only be inferred. As a result, the concepts of uncertainty and confidence arise from the fact that it can test only a limited sample from a large population. The statistical theory for reliability such as the reliability function, the probability density function, the cumulative density function, the hazard rate, the conditional reliability function, and mean time to or between failure (MTTF or MTBF) are needed.

In general, two different reliability validation approaches are used in the automotive industry, the knowledge-based approach and the standard-based approach. For the knowledge-based approach, the failure rates are quantitatively determined under various use conditions based on DFMEA, physics of failure models, continuous probabilistic model, and prior knowledge with customized stress conditions. Usually, failures are needed so are good. In contrast, the standard-based approach proves that a defined failure rate is met based on specifications, experience, and shipped product field return knowledge. It is a zero-failure test or test of Bogey. *N* parts are tested to one life and no failures are allowed so failures are bad. The mathematical interpretation is that the product has unknown inherent reliability, *R*. The reliability test verifies that *R* exceeds a critical value with a specified probability or confidence, C, as shown in Eq. 2.

$$\Pr(R > R_{\text{critical}}) = C \tag{2}$$

For the automotive industry, standard-based validation has been popularly used. Many standards such as AECQ, GM3172, ISO16750, JEDEC, IEC, etc. form a framework throughout the industry for easy implementation. It is a simple digital "pass or fail" method. It requires a sample of a predetermined size to be tested for a specific length of time under a specific test condition. The required reliability then is demonstrated if no failures occur at the end of the test. In this section, we will introduce a set of standard-based reliability validation tests based on GMW3172 and ISO16750 standards.

4.2.1 AV Mission Profiles

Automotive electronic component reliability validation tests start and end with the mission profile. When specifying a component, it is common for OEMs and their suppliers to develop a specific mission profile, which is essentially a summary of all the expected environmental and functional conditions that the component will face during its service life. As AV largely will be used for Robo-taxi rideshare service, the fleet can be deployed at a specific location following its unique operational design domain (ODD). Furthermore, it can be controlled 100% by service providers. Therefore, it will have its customized mission profiles to mimic a particular type of field stress, as well as its related severity. In addition to the operation life mission profile that was discussed in the introduction section, the most commonly referenced stresses are related to temperature, humidity, dynamic loads, thermomechanical stress, chemical load, UV radiation, dust ingression, water ingression, and EMC load.

The customized mechanical random vibration mission profile is usually obtained by installing a series of accelerometers at various vehicle locations to record the transfer function through the vehicle operation. As shown in Fig. 20, an example of dynamic responses from a vehicle driving at Gomentum Station proving ground located in Concord, California to mimic smooth suburb and city driving roads was recorded [20]. The solid colorful wavy lines are acceleration power spectral density (PSD) curves instrumented on the roof rail of a vehicle from different runs, the solid black line is the envelope profile, the dash black line is the margin profile, and the solid red line is the accelerated profile.

The customized temperature mission profile is usually obtained by installing a series of temperature loggers at various vehicle locations to record the temperatures through the vehicle operation. As an example of the temperature profiles of San Francisco shown in Fig. 21, roof temperature and trunk temperature are dependent on ambient air temperature, vehicle driving speed (airflow), solar loading, and heating sources from the surrounding components. The highest peaks are associated with vehicle idles. The trunk temperature has less temperature swing as compared to the

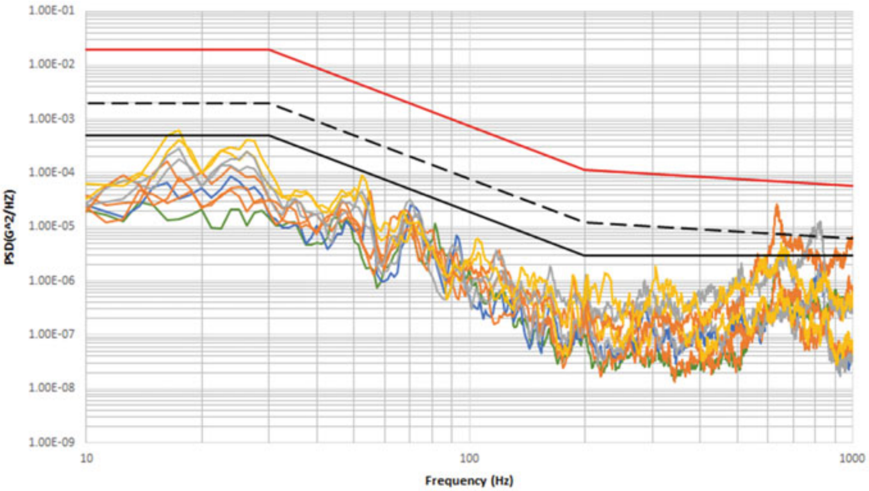


Fig. 20 Example of customized random vibration profile

roof temperature. By comparing the CDFs of roof and trunk temperatures as shown in Fig. 22, both of them exhibit a multimodal distribution.

Customized temperature cycling (TC) mission profiles can be inferred from the time-dependent temperature loggings. Endo and Matsuishi [21] developed the Rainflow Counting (RFC) method by relating stress reversal cycles to streams of rainwater flowing down a Pagoda. The rainflow counting algorithm is one of the popular counting methods used in fatigue and failure analysis from a time history for cycle counting and was adopted as a standard by ASTM E 1049-85. The rainflow counting method allows the application of Miner's rule to assess the fatigue life of a structure

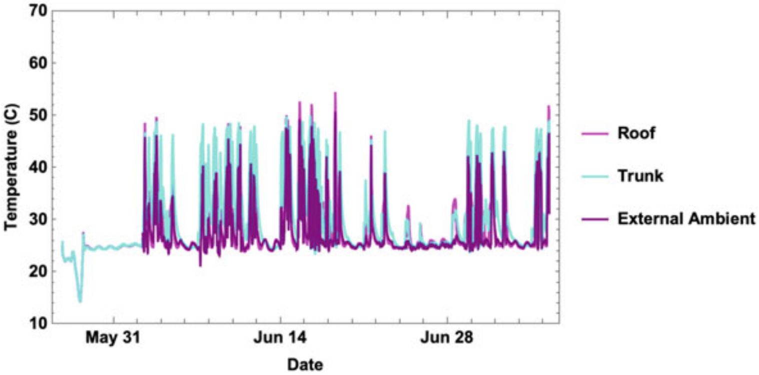


Fig. 21 Temperature time-dependent profile for roof and trunk locations together with the external air ambient temperature

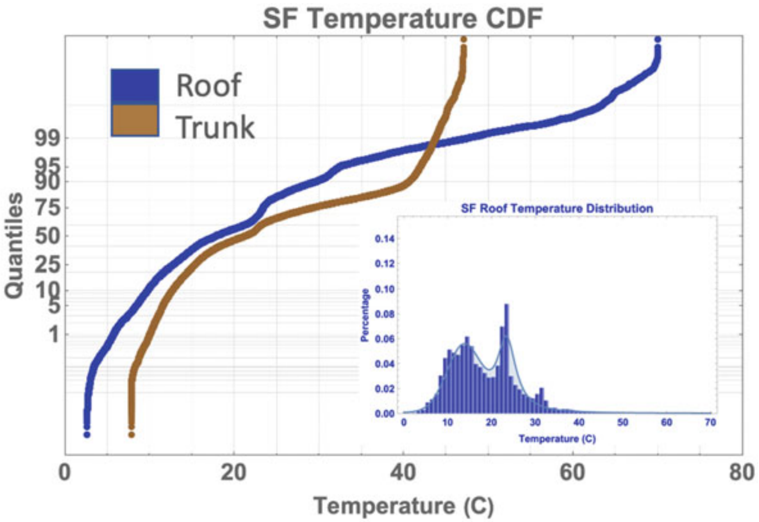


Fig. 22 CDF of roof and trunk temperatures. Inset: PDF of root temperature

subject to complex loading. For TC mission profile establishment, RFC is recommended to avoid potential over-stress and under-stress TC damages. Figure 23 shows an example of computing operating temperature in the trunk recorded up to 1600 h and calculated dT and average temperature distribution based on RFC. Accordingly, based on Miner's rule of linear accumulation of the damage, when the damage fraction (LC) reaches 1, failure occurs per Eq. 3, the effective *dT*, therefore, can be determined for stress to field condition transformation.

$$\text{Total Damage} = \sum_{i}^{m}\left(\frac{\Delta T_{\text{stress}}}{\Delta T_{\text{ref}-i}}\right)^{n} \times P_i = \left(\frac{\Delta T_{\text{stress}}}{\Delta T_{\text{ref}-\text{eff}}}\right)^{n} \tag{3}$$

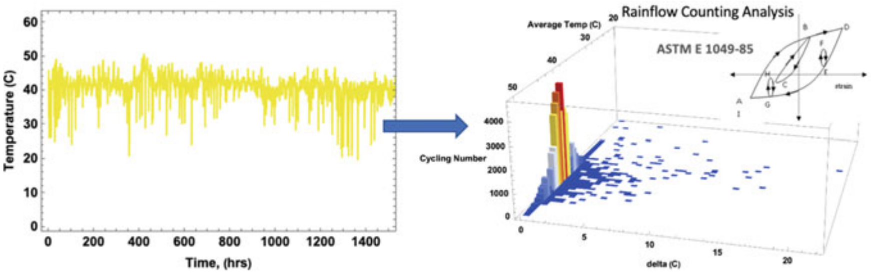


Fig. 23 An example of Compute operating temperature in the trunk, and calculated *dT* and average temperature distribution based on RFC

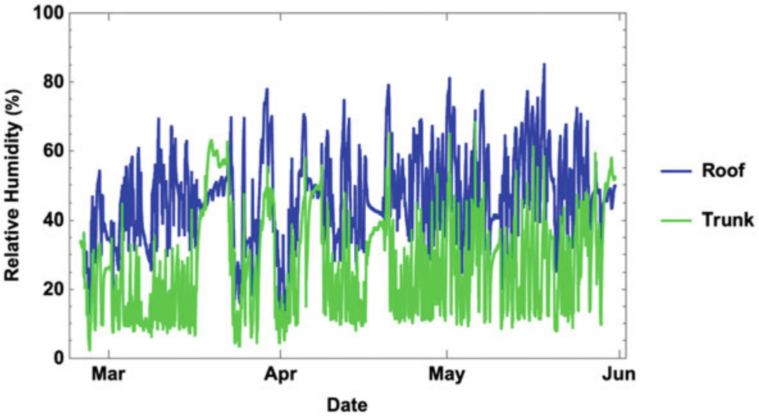


Fig. 24 Time-dependent RH levels from roof and trunk recorded during vehicle operation in San Francisco

where ΔT_{ref} is the reference temperature change from the stress condition, and ΔT_{eff} is the effective temperature change for field operation derived from TC mission profile data and Miner's rule.

The customized humidity mission profile is usually obtained by installing a series of humidity loggers at various vehicle locations to record the relative humidity (RH) through the vehicle operation. Figures 24 and 25 show the time-dependent relative humidity level from roof and trunk locations during vehicle operation in San Francisco, and the CDF of those RH values. Interestingly, by plotting RH vs temperature as shown in Fig. 26, it was found that high-temperature high humidity conditions likely can't co-exist even in a coastal city like San Francisco. Also, it was found that the location is important. Vehicle trunk location is much drier with less RH change than roof location.

Lastly, a customized solar loading mission profile is usually obtained by installing a set of pyranometers at various vehicle locations to record the solar intensity through the vehicle operation. Figure 27 shows an example of solar intensity in downtown San Francisco on a day in September 2018. It indicates that solar loading has a strong dependence on vehicle speed, and location (indoor vs. outdoor). Furthermore, it was found that component surface finish color has a profound impact on component temperature under solar loading as shown in Fig. 28. Overall, we found solar load effect has geometric location, seasonal, daytime, surface finish, indoor vs. outdoor, and the state of the atmosphere dependences.

4.2.2 Reliability Validation Tests

Reliability is a method to determine how long a product will last. Therefore, reliability engineering is the prediction of the life of the product. It is different from quality

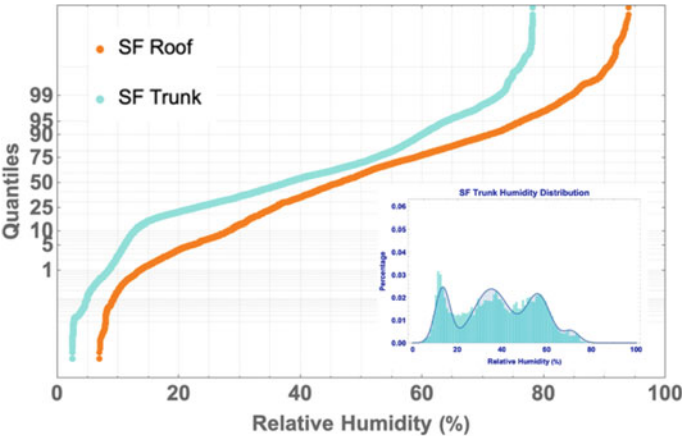


Fig. 25 CDF of RH values shown in Fig. 28

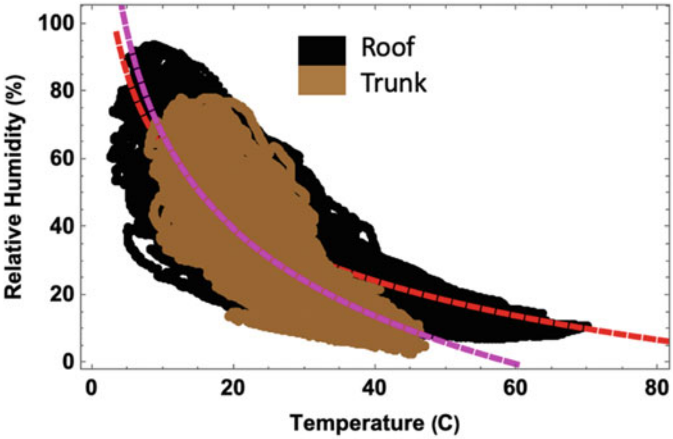


Fig. 26 An inverse relationship between RH and temperature for both roof and trunk locations

which means how closely the product meets user needs. Reliability testing is the testing of the life of the product by repeatedly making the product go through the stresses for an estimated duration or number of cycles and checking for failures.

As we mentioned earlier, reliability testing is sampling testing. The probability of failure of a population can only be inferred. Thus, the concepts of uncertainty and confidence arise. Practically, we are dealing with two kinds of confidence during our work. The first one is engineering confidence, which is mostly a matter of judgment and experience based on people. The second one is statistical confidence, which is used to make inferences about a population based on the sampling data. Statistical confidence is the one that directly impacts the reliability testing plan. A good reliability testing plan should be able to build a statistical sample size, meet a particular

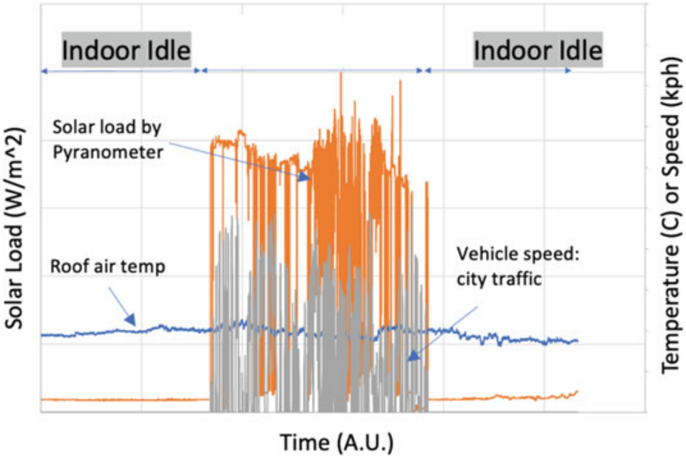


Fig. 27 Solar intensity recorded by a pyranometer in downtown San Francisco

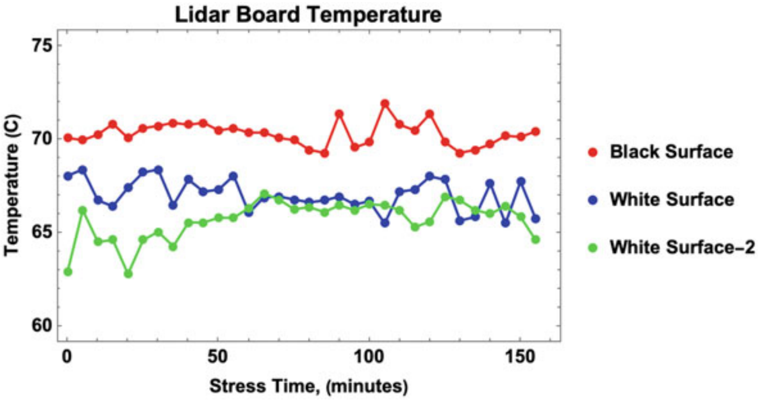


Fig. 28 LiDAR board temperatures with different surface finishes

reliability objective or goal, and achieve a specific confidence level. In general, two statistical approaches can be used for developing a reliability testing plan for sample size N, the chi-squared testing approach and the Weibull Bayesian estimate of zero-failure approach. The chi-squared testing approach is for the flat part of the failure rate bathtub curve or random failures with a constant failure rate as shown in Eq. 4:

$$N = \frac{\chi^2(\alpha, 2n+2)}{2\overline{\lambda} \times \mathrm{AF} \times t} \tag{4}$$

where $\overline{\lambda} = 1/\mathrm{MTBF}$ and is an upper bound failure rate objective, AF is acceleration factor, n is a number of failures, and α is confidence level.

Weibull Bayesian estimate can model the entire bathtub curve with different Weibull slope values as shown in Eq. 5:

$$N = \frac{Ln[1-\alpha]}{Ln[R] \times \left(\frac{t_{\text{test}} \times \text{AF}}{t_{\text{spec}}}\right)^{\beta}} \tag{5}$$

where R is lower bound reliability objective, β is the Weibull slope with $\beta < 1$ for early life failures, $\beta = 1$ for random failures, and $\beta > 1$ for wear-out failures, t_{test} is the total test time, and t_{spec} is the specification life. Weibull Bayesian estimate can't allow any failures.

Different acceleration factors are used for different failure mechanisms to determine the sample size or testing duration per different reliability tests. For thermal shock, temperature cycling, and power temperature cycling, a modified Norris-Landzberg model or simple Coffin-Manson model could be used. Modified Norris Landzberg consists of four parts as shown in Eq. 6.

$$\text{AF} = \left(\frac{\Delta T_{\text{test}}}{\Delta T_{\text{field}}}\right)^{a} \times \left(\frac{\text{Dwell}_{\text{test}}}{\text{Dwell}_{\text{field}}}\right)^{b} \times \left(c \times \text{RampRate}^{d}\right) \times \mathrm{e}^{f \times \left(\frac{1}{T_{\text{fieldmax}+273}} - \frac{1}{T_{\text{testmax}+273}}\right)} \tag{6}$$

The parameters of Norris Landzberg for lead-free solder and lead solder are listed in Table 7 for reference. The first part in Eq. 6 is the Coffin-Manson acceleration.

For high temperature and/or high humidity, Arrhenius and/or Peck equation usually could be used as shown in Eq. 7.

$$\text{AF} = \left(\frac{RH_{\text{low}}}{RH_{\text{high}}}\right)^{n} \times \mathrm{e}^{\left(\frac{E_{\text{a}}}{k_{\text{B}}}\right)\left(\frac{1}{T_{\text{low}}} - \frac{1}{T_{\text{high}}}\right)} \tag{7}$$

The common parameters for Arrhenius and Peck models are listed in Table 8 for reference.

For random vibration, the Basquin model usually could be used to scale the vibration testing time versus G level as shown in Eq. 8.

Table 7 Referenced parameters for Norris-Landzberg model

	AF	Parameter	Lead-free solder	Leaded solder
1	Coffin-Manson	a	2.65	2.5
2	Dwell time	b	0.136	0.0667
3	Ramp rate	c	1.22	0.80094
		d	−0.0757	0.0964
4	Highest temperature	f	2185	1414

Table 8 Referenced parameters for Arrhenius-Peck model

	Parameter	Value
1	n (humidity exponent)	−2.66
2	E_a (activation energy) (eV)	0.8 (average conservative value)
3	k (Boltzmann's constant) (eV/K)	8.6173×10^{-5}

$$G_{\text{RMS - accelerated}} = G_{\text{RMS - normal}} \times \left(\frac{T_{\text{normal}}}{T_{\text{accelerated}}}\right)^{m/2} \tag{8}$$

where m is the Basquin's exponent or fatigue parameter. Some reference numbers for different materials are listed in Table 9.

Table 10 illustrates an example of environmental and mechanical reliability design validation (DV) testing plan for a Compute located in the trunk or vehicle's rear compartment. Total five water-fall legs covering 22 testing items are required for Compute engineering and design validations. Leg 0 includes vibration transmissibility demonstration and thermal cycle profile development, temperature measurement, visual inspection, and design review based on test results (DRBTR), and cross-section (x-section). Leg 1 includes low-temperature wakeup, high-temperature degradation, pothole shock, and random vibration with temperature cycling. Leg 2 includes low-temperature wakeup, non-operational thermal shock or temperature cycling, power temperature cycling (PTC), humidity heat cyclic (HHC), humidity heat constant (HHCO), and salt mist. Leg 3 includes low-temperature wakeup, minimum temperature non-operation temperature storage, dust ingression (IP5k), and water ingression (IP2). Leg 4 includes low-temperature wakeup, collision shock, elbow load, and fretting corrosion for connectors. Among all the tests listed in Leg 1–4 in Table 10, high-temperature degradation, mechanical shock, random vibration, TS, and PTC are considered stress tests or quantitative accelerated life tests. Their acceleration factors can be calculated based on industry-accepted models as described above. The rest tests are considered as performance indicator tests or qualitative accelerated tests. For stress tests, the test duration and sample sizes can be calculated based on Eqs. 4–8 and customized mission profiles per each stress test. For performance indicator tests, the testing durations and sample sizes are recommended to follow the GMW3172 standard.

Table 9 Referenced m values for Basquin vibration model

	Materials	m – Material fatigue constant
1	Aluminum leads in electronic assemblies	6.4
2	Overall usage value	5
3	Connector fatigue or fretting corrosion	4
4	Highly accelerated vibration for metal fatigue (>3.3x original stress)	3.3

Table 10 Compute environmental and mechanical stress tests

	Test A	Test B	Test C	Test D
Test Leg				
DV Leg 0	VTD and TCPD	Temperature measurement	Visual Inspection and DRBTR	X-section
DV Leg 1 water-fall	Low temperature wakeup	High temperature degradation	Shock—Pothole	Vibe w/TC
DV Leg 2 water-fall	Low temperature wakeup	Thermal shock	PTC	Humidity heat cyclic humidity heat constant salt mist
DV Leg 3 water-fall	Low temperature wakeup	Min non-op temperature	Dust (IP5k)	Water (IP2)
DV Leg 4 water-fall	Low temperature wakeup	Shock collision	Elbow Load	Fretting corrosion

It should be noted that a 5-point check before and after each leg is required for all the legs, and a 1-point check at the end of Tests A, B, and C is required per each leg listed in Table 10 except Leg 0.

Compute may be exposed to a variety of different fluids. Exposure to these fluids may affect the functionality of the Compute. The chemical load tests or fluid compatibility tests are intended to assure that vehicle operating liquids, chemicals and oils will not degrade the materials, identification, or function of the Compute. Although other fluids beyond those in the list in Table 11 could come into contact with the Compute, these fluids were considered more aggressive. The following list of fluids in Table 11 was selected based on the likelihood of exposure and the severity of exposure for Compute using liquid coolant located in the trunk or vehicle's rear compartment.

Table 11 Compute chemical load list

Fluid/Chemical/Substance	Specification/Part number	Method
Commercial vehicle cleaning agent-interior	Genuine GM Fluid 88,861,405, Leather, Vinyl and plastic cleaner, Formula 409, Fantastik multi-purpose cleaner, Sonax car interior cleaner	Normal cleaning
Engine Coolant	Ethylene glycol (EG) base fluids, 50:50%	Pour test
Grease, Electrical connector, Dielectric lubricant	9,985,821	Brush test
Ammonia based cleaner	Windex, Sonax glass clear, Glass cleaner	Normal cleaning
Coca-Cola classic		Pour test
Coffee (10 oz., 0.5 oz. Cream, 2 tsp. Sugar)		Pour test

4.3 EMC/ESD Validation

AV Compute EMC testing is the process of measuring the electromagnetic compatibility of a Compute and its components. Automotive EMC testing is now more important than ever as AV component RF design grows in complexity. The main purpose of Compute EMC testing is to test the mutual influence between the Compute and the surrounding electromagnetic environment, which includes the ability of the Compute to resist a given electromagnetic disturbance and the indicators of the electromagnetic disturbance generated by the Compute. That is, the Compute is not affected by the electromagnetic disturbance emitted by other equipment in the electromagnetic environment, and the Compute cannot generate electromagnetic disturbance that exceeds the prescribed limit. Electro Static Discharge (ESD) testing of Compute refers to the transfer of unbalanced charges on the surface of the Compute. When the charge voltage difference is higher than a certain level, the insulating medium will undergo an electrical breakdown process, which will cause a localized conductive path to form inside the insulating medium. Such localized conductive path can induce high current passing through. The main destructive force of electrostatic discharge is the thermal effect from the instantaneous peak current, which can easily cause the Compute electronic components to be broken down or burned, and then cause malfunction of the entire Compute. For safety concerns in automotive electronics, automotive ESD compliance standards have higher voltage test limits than commercial electronics.

The internationally accepted automotive EMC regulations include ECE R10 regulated by the United Nations Economic Commission for Europe (ECE), 97/24/EEC and 95/54/EEC regulated by the European Union, and CISPR (French: Comité International Spécial des Perturbations Radioélectriques), Society of Automotive Engineers (SAE), Japanese Automobile Standards Organization (JASO) and ISO. Generally speaking, EMC/EMI is tested according to customer requirements and specifications in the state of the whole system. Different vehicle OEMs have their own EMC testing specifications such as GMW3091 and 3097 from GM, ES-XW7T-1A278-AC from Ford, TSC3351 from Toyota, DC-10614 and DC10615 from Daimler Chrysler, etc. In this section, we will use GMW3097 as our EMC validation baseline. For EMC testing laboratories, the major U.S. automakers have requested that EMC testing of all components must be performed in a laboratory accredited by Automotive EMC Laboratory Recognition Program (AEMCLRP).

A Compute EMC Test plan should be developed outlining the following per IEC 11451-2:2015:

1. test setup
2. frequency range
3. the reference point(s) (or line if a four-probe method is used)
4. vehicle mode of operation
5. vehicle acceptance criteria
6. definition of test severity levels
7. vehicle monitoring conditions

Table 12 Compute test mode descriptions

DUT test mode number	Operation and description	Function	Used for GMW 3097:2019 procedures
0	Unpowered	Off	ESD handling
1	Full operation	On—Continuous high utilization processing cycle running with traffic on ethernet ports	Emissions tests, Radiated immunity tests, ESD power-up mode
2	Ethernet traffic only	On—Low utilization of processors, full ethernet traffic	Conducted immunity tests

8. modulation
9. polarization
10. Compute orientation
11. antenna location
12. test report content.

For Compute on EV, three modes of operation such as unpowered, ethernet traffic only, and full operation full load should be tested as shown in Table 12.

The full Compute shall be tested in the below set of tests covering both EMC and ESD as shown in Table 13. A generic Compute test setup and a test configuration to be used for all RE tests are shown in Fig. 29. The test should be conducted twice, once with a grounded enclosure and once with an un-grounded enclosure. For isolation, the Compute shall be placed on a non-conductive, low relative permittivity material ($\varepsilon_r \leq 1.4$), at (50 ± 5) mm above the reference ground plane. During the test, all Compute shall not exceed the limits defined by Radiated Emissions Absorber-Lined Chamber (ALSE) Non-Spark Requirements in GMW3097:2019, by Conducted Emissions Artificial Network (AN) Non-Spark Requirements in GMW3097:2019, the "Level 2" requirement for all frequencies and modulations. If Compute passes the component level EMC tests but does not pass the vehicle level EMC tests, the vehicle level test results will be the determining factor for validation test pass/fail status.

ESD test shall verify the immunity of lines, pins, or Compute enclosure locations, which are to be subjected to ESD discharge events. ESD test shall identify the potential ESD discharge points and list all individual pins, case discharge locations, discharge type, simulator voltages, discharge network type, and a description of the pin signal.

Table 14 defines Compute ESD testing for Power-On Mode Setup per GMW 3097:2019 3.6.1 as an example. The test should be conducted twice, once with a grounded enclosure and once with an un-grounded enclosure. For isolation, the Compute shall be placed on a non-conductive, low relative permittivity material ($\varepsilon_r \leq 1.4$), at (50 ± 5) mm above the reference ground plane. During the test, for the test mode(s) given in Table 13, all Compute locations must comply with the performance standards defined in GMW3097:2019. After the test, no permanent

Table 13 Summary of EMC/ESD tests

GMW3097 section	Description
3.3.1	Radiated Emissions—Absorber-Lined Shielded Enclosure
3.3.2	Radio Frequency Conducted Emissions (via Artificial Network)
3.4.1	RF Immunity—Bulk Current Injection
3.4.2	RF Immunity—Anechoic Chamber
3.5.2	Transients Conducted Immunity, Nominal 12 V Lines
3.5.3	CI, Fast Transient Coupling
3.5.4	CI, 30 V DCC Transient Coupling
3.6.1	Electrostatic Discharge, Power on Mode
3.6.2	Electrostatic Discharge, Remote I/O
3.6.3	Electrostatic Discharge, Handling of Devices

Compute damage or performance deviations shall be observed. The Compute ESD power-on test configuration is shown in Fig. 30.

- The DUT is inaccessible from the outside of the vehicle
- Capacitance = 150 pF
- Resistance = 2 kΩ.

Table 15 defines Compute ESD testing for Remote I/O—Inputs/Outputs Setup per GMW 3097:2019 3.6.2 as an example. Remote I/O testing is to be completed on pins 1–8 of each of the two RJ45 Ethernet Service port connectors as well as both PDB LIN lines as shown in Fig. 31. During the test, for the test mode(s) given in Table 15, all Compute pins must comply with the performance standards defined in GMW3097:2019. After the test, no permanent Compute damage or performance deviations shall be observed. This includes changes in rising edge shape in pre/post serial bus plots.

For 4–15 kV.

- Capacitance = 150 pF
- Resistance = 2 kΩ
- Human Body Model (HBM) = 330 pF/2 kΩ for $\leq$ 15 kV; 150 pF/2 kΩ for $>$ 15 kV, unless otherwise specified by GM EMC Engineer.

Table 16 defines Compute ESD testing for Handling of Devices Setup per GMW 3097:2019 3.6.3 as an example. Remote handling testing will be performed on all contactable ports. Both Contact and Air Discharge methods should be attempted. During the test, for the test mode(s) given in Table 16, all Compute ports must comply with the performance standards defined in GMW3097:2019. After the test, no permanent Compute damage or performance deviations shall be observed after exposure.

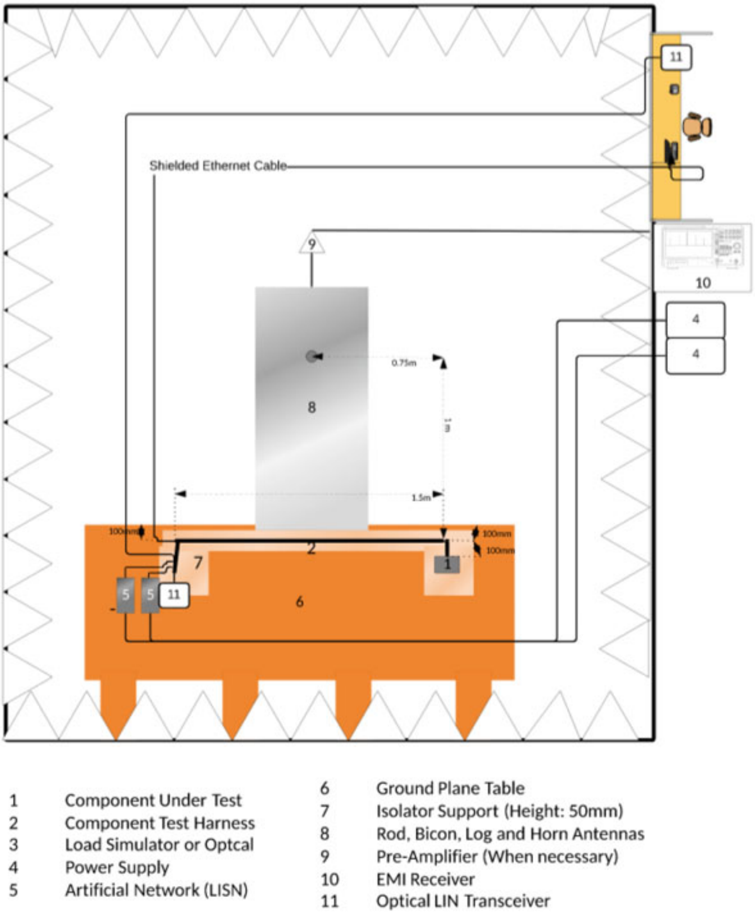

11
Shielded Ethernet Cable
9
10
4
4
0.75m
8
1m
1.5m
100mm
100mm
100mm
2
7
1
5
5
11
6
1 Component Under Test
2 Component Test Harness
3 Load Simulator or Optcal
4 Power Supply
5 Artificial Network (LISN)
6 Ground Plane Table
7 Isolator Support (Height: 50mm)
8 Rod, Bicon, Log and Horn Antennas
9 Pre-Amplifier (When necessary)
10 EMI Receiver
11 Optical LIN Transceiver

Fig. 29 RE configuration show monopole antenna as a reference, follow CISPR25 for other antennas location

- Capacitance = 150 pF
- Resistance = 2 kΩ.

5 Challenges to Safe Deployment at Scale

5.1 Artificial Intelligence: Perception and Prediction

The perception and prediction rely on the sensors on the autonomous vehicle such as cameras, LiDARs, and radars. Their performance is different in different scenarios. For example, the resolution of 2D image from the camera becomes low in the dark.

Table 14 ESD, test during operation of the device (power-on mode) test

Mode	Location (Pin/Case)	Discharge type (Air/Contact)	ESD simulator voltage (kV)	Signal/Pin description
1	Enclosure surface points	Air	±4	Screw holes/Enclosure edge
1	Enclosure surface points	Contact	±4	Screw holes/Enclosure edge
1	Enclosure surface points	Air	±6	Screw holes/Enclosure edge
1	Enclosure surface points	Contact	±6	Screw holes/Enclosure edge
1	Enclosure surface points	Air	±8	Screw holes/Enclosure edge
1	Enclosure surface points	Contact	±8	Screw holes/Enclosure edge
1	Enclosure surface points	Air	±15	Screw holes/Enclosure edge

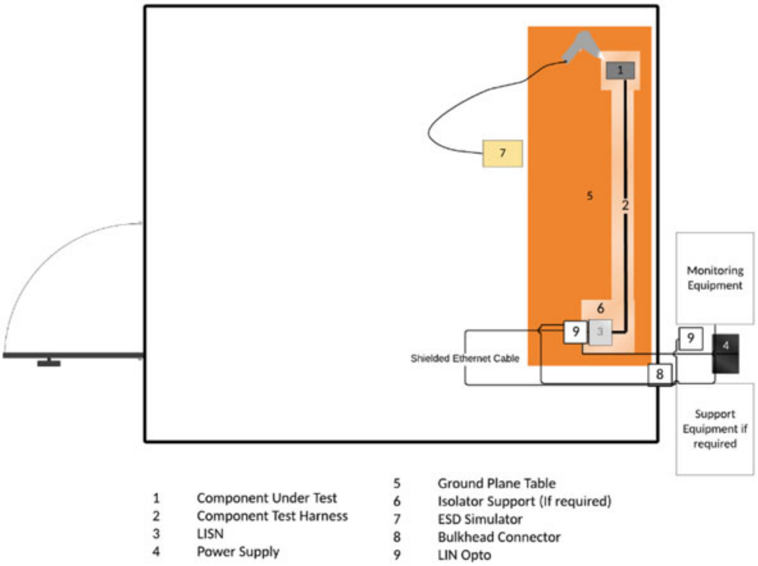


Fig. 30 ESD power-on configuration

Table 15 ESD, remote inputs/outputs test

Mode	Location (Pin/Case)	Discharge type (Air/Contact)	ESD simulator voltage (kV)	Signal/Pin Description/Name
1	Ethernet pins 1–8 on both service ports and both LIN lines	Contact	±4	Ethernet cable for service RJ45 connector and LIN lines
1	Ethernet pins 1–8 on both service ports and both LIN lines	Contact	±6	Ethernet cable for service RJ45 connector and LIN lines
1	Ethernet pins 1–8 on both service ports and both LIN lines	Contact	±8	Ethernet cable for service RJ45 connector and LIN lines
1	Ethernet pins 1–8 on both service ports and both LIN lines	Air	±15	Ethernet cable for service RJ45 connector and LIN lines

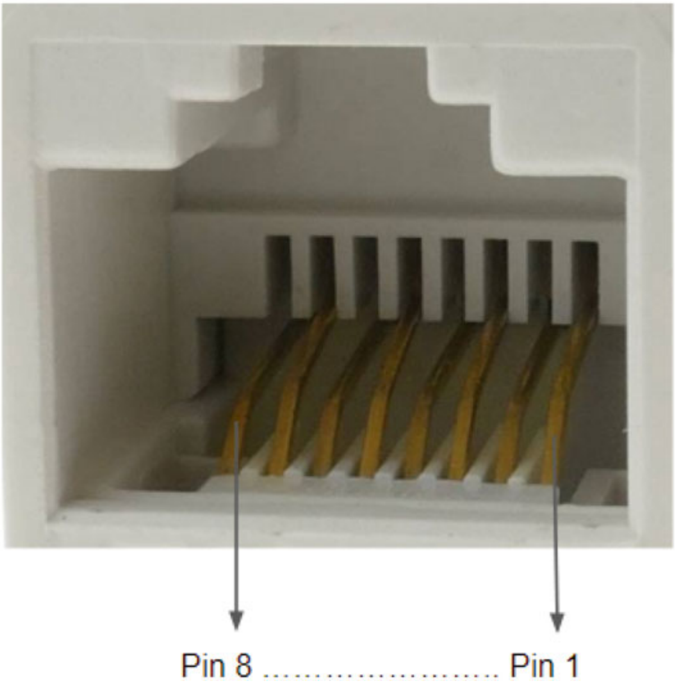


Fig. 31 RJ45 Ethernet port pins

Table 16 ESD, handling of devices test

Mode	Location (Pin/Case)	Discharge type (Air/Contact)	ESD simulator voltage (kV)	Signal/Pin Description/Name
1	All ports	Contact	±4	All ports
1	All ports	Contact	±6	All ports
1	All ports	Air	±8	All ports

On a rainy day or foggy day, the sensor performance will be low. It will impact the perception of an AV. It is challenging for an AV to operate in complex urban streets such as busy intersections in the urban street. Many pedestrians and vehicles appear to be moving in different directions. It is difficult for an autonomous vehicle to do perception, prediction, and make decisions.

5.2 *Power Consumption*

With government policies for carbon emissions and environmental protection, more and more autonomous vehicles are BEVs. The electric vehicle battery range becomes very impotent. How to increase EV maximum range with autonomous L4 and L5 driving is a big challenge. One option is to increase the battery range. Another option is to reduce autonomous vehicle power consumption.To achieve fully autonomous driving, the autonomous computing platform needs more performance. More performance means more power consumption. For example, the Nvidia Drive AGX is 300 W with 320 TOPS performance. The Tesla D1 Dojo is 400 W with 362TOPS performance. As every watt matters, it is required to design Compute with EVs in mind. One way to do this is to improve efficiencies in the system themselves by designing Compute from the ground up with the EV power platform in mind. It is imperative to have a custom-designed, high density, functionally safe chip, but with lower power consumption to give AV maximum miles on the road. As an example, a new application-specific integrated circuit (ASIC) can achieve more performance but less power consumption as shown in Table 17.

Table 17 Power, performance, and TOPS per watt comparisons of different ASIC chips

ASIC	Power consumption (W)	Performance (TOPS)	TOPS/W
Mobileye Eye Q5	10	24	2.4
Google TPU v3	40	420	10.5
Qualcomm Snapdragon Ride L4/L5	130	700	5.38

5.3 Thermal Management

The autonomous computing platform could generate tremendous heat that increases the component operating temperatures above their temperature limit. Such overheating prevents the components from functioning efficiently, safely, accurately, and reliably. It is critical to control the temperature below the maximum operating temperature limits to prevent them from degrading and malfunctioning.

As long as the autonomous computing platform has large power consumption, it becomes a challenge to cool the temperature below its operating temperature limit. If passive cooling is not able to reduce the temperature, active cooling should be required. Cooling the temperature below the maximum operating temperature is needed to ensure the performance of a Compute. Generally, liquid cooling is used in the autonomous computing platform.

5.4 Manufacturing, Assembly, and Quality Control

After the design phase of the autonomous driving system, it is going to the manufacture and assembly phase. Generally, the manufacturing includes PCB process, PCBA process, function test, enclosure assembly, burin in test, package, and shipping as shown in Fig. 32.

The PCBA process (Fig. 33), includes solder paste printing, place components, reflow soldering, automated optical inspection (AOI), in-circuit test (ICT), image programming, and function test. After that, it is going to the enclosure assembly.

The autonomous driving system is going through a lot of process steps during manufacturing and assembly. It is important how to do quality control and make sure the system has no issues during each process step. Especially, in mass production, how to improve the yield rate becomes challenging.

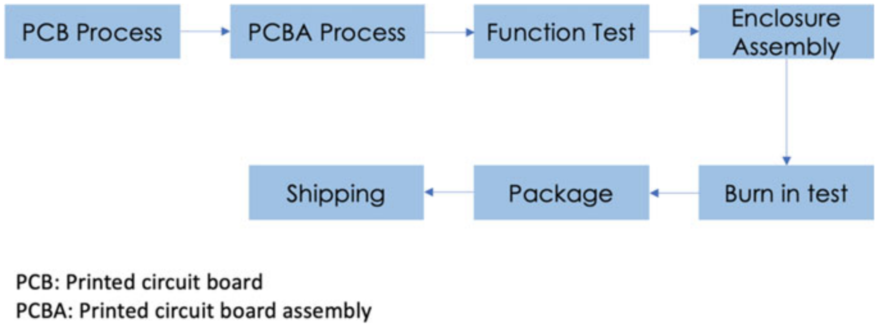


Fig. 32 Autonomous computing system manufacturing process

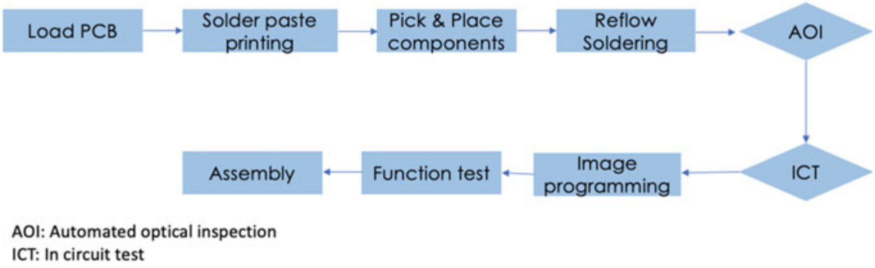


Fig. 33 Autonomous computing system PCBA process

5.5 *Size and Cost*

For fully autonomous driving, to achieve the performance for perception and prediction, the computing platform needs to use several CPUs, GPUs, and memory to meet the performance requirement. The board and enclosure sizes become larger. Considering the redundancy to make the computing platform safe, several boards are needed in the encloser unit. It will increase the total cost of an AD system.

To reduce the CPU/GPU temperature below the maximum operating temperature, the liquid cooling system is generally used. The enclosure is designed to have heat pipes as well as liquid pipelines or channels. All of them will add to the overall cost of the autonomous driving system.

5.6 *Quality and Reliability*

The strict reliability standards for AV Compute are critical for road safety and human safety. Since AV Compute would run software with more than 1 billion lines of code during life, hardware reliability is an absolute necessity because a blue screen of a system crush at 60 MPH could mean actual death. Currently, there is no industry well-established reliability target for AV electronic modules such as Compute. Most OEMs and tier-1s adopt automotive industry established traditional vehicle reliability specifications such as 99% or 95% reliability at end of vehicle life to qualify AV electronics. The risk classification scheme of Automotive Safety Integrity Level D (ASIL-D) defined by ISO 26262 places a more stringent reliability standard on self-driving vehicles. For these vehicles to be ASIL-D compliant, the maximum acceptable probabilistic metric for random hardware failure (PMHF) is 10FIT. In other words, these vehicles can only make ten errors in 1 billion hours of operation, while an average U.S. driver makes 10,000 mistakes in the same duration. As an example, for a Compute to achieve a 10FIT failure rate at the end of 5 years of life with an 80% duty factor, the reliability target will be 99.965% instead of 99 or 95%, which is a great challenge.

Failures in computer systems are broadly categorized into permanent hard failure and intermittent recoverable soft faults. Permanent hard failures are repeatable and occur the same way every time. On the contrary, intermittent soft faults are temporary and are a function of the operating environment and stress loading. While permanent hard failures sound scary, they are relatively easier to handle in general. A diligent reliability testing framework usually can expose permanent hard failures, therefore they can be mitigated by design and process optimization. In the worst case, they can be monitored, diagnosed, and quarantined by on-vehicle safety measures. But intermittent faults are often harder to be diagnosed so to be prevented since they are a function of the unique operating environment and stress loading. How to address intermitted recoverable faults is another great challenge for Compute validation and usage.

As mentioned in an early section, AV Compute's operation time and mileage mission profiles could be 2–3 times of the traditional human-driving vehicles. With such longer daily continuous operation hours or mileages, the reliability specifications for AV hardware especially Compute shall be higher, therefore it will be challenging. Furthermore, with such long continuous operation, the probability of a vehicle hitting extreme road conditions or corner cases increases drastically. For environmental loads such as thermal, mechanical, radiation, dust, water, humidity, chemical, etc., we can't just use traditional values such as 95th or 98th, or even 99th percentiles to model such loads for the use conditions. We may have to adopt the absolute worst case from a 5- or 10-year period to truly guard band Compute's durability. In addition to bad environmental conditions, poor infrastructure and chaotic road conditions are also proving to be tremendously challenging for Compute operation.

In the worst case, a Compute with redundant GPUs and CPUs could consume more than 2000 W of power. Therefore, an enormous amount of heat would be generated. As autonomous driving functions rely heavily on the Compute power of the data processing units, the clock speeds must stay in the optimal range all the time. Therefore, air cooling usually is not adequate to meet the thermal management requirement. Instead, liquid cooling offers a much higher cooling capacity and is therefore generally chosen as the Compute thermal solution. Using a liquid cooling coldplate to enclose a Compute will impose several new reliability challenges. First, the coldplate usually serves as a good "moat" to insulate the temperatures of boards/components inside from external temperature changes as shown in Fig. 34. Therefore, the traditional reliability testing methods described by GMW3172 do not apply to liquid cooling Compute. The boards and components inside the coldplate will see liquid-to-liquid induced temperature changes instead of air-to-air induced temperature changes. Since a liquid is used as the thermal medium, very high thermal ramp rates can be achieved with liquid-to-liquid temperature change, when compared to the air-to-air temperature change. As a result, the liquid-to-liquid temperature change is considered more stringent stress than air-to-air in terms of acceleration. As a consequence, a more severe board and coldplate interaction would be generated. Second, for a liquid cooling Compute, during its validation testing and field operation/maintenance, condensation effect, hydrolocked state effect, and water hammer effect all need to be diligently investigated and assessed. As illustrated in Fig. 35,

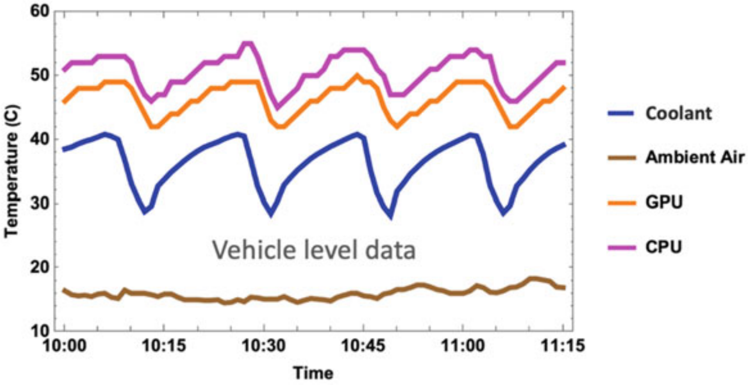


Fig. 34 Compute GPU and CPU temperatures follow coolant closely, while they are independent of ambient air temperature

a hydrolocked state-induced hydrostatic pressure under a high-temperature stress condition can deform the coldplate severely to cause Compute failure. Third, special unit handling, chiller pump operating, and transporting procedures also need to be carefully developed. As an example, at the end of Compute testing, wait until coolant temperature reaches room temperature before unplugging the Compute connector. To prevent a water hammer, first, unplug the connector of the water inlet, then unplug the connector of the water outlet.

For a Compute to be automotive qualified, all the ICs and electronic components used on the board need to meet AECQ standards throughout the manufacturing and testing process first. AECQ is a set of failure mechanism-based stress test qualifications. Among them, AEC-Q100 is for packaged integrated circuits, AEC-Q101 is for active components, AEC-Q102 is for optoelectronic devices, AEC-Q200 is for passive components, and AEC-Q104 is for multi-chip module (MCM) used in automotive applications. This specification has been established by the Automotive Electronics Council (AEC) to define qualification requirements and procedures for ICs and electronic components used in the automotive industry. An AEC-qualified device means that the device has passed the specified stress tests and guarantees a certain level of quality and reliability. Unfortunately, currently on the market, there has been none AEC-qualified CPUs ever available. Furthermore, due to supply chain shortage issues and some other reasons, it is not common that non-automotive-grade ICs and components have to be used for Compute. Such usages bring a great challenge for Compute for its long-term reliability and defect-free quality requirement for harsh environment operation. In general, a thorough gap analysis with component and board level validations is needed to assess the real risks if non-automotive-grade ICs and components are going to be used for AV Compute.

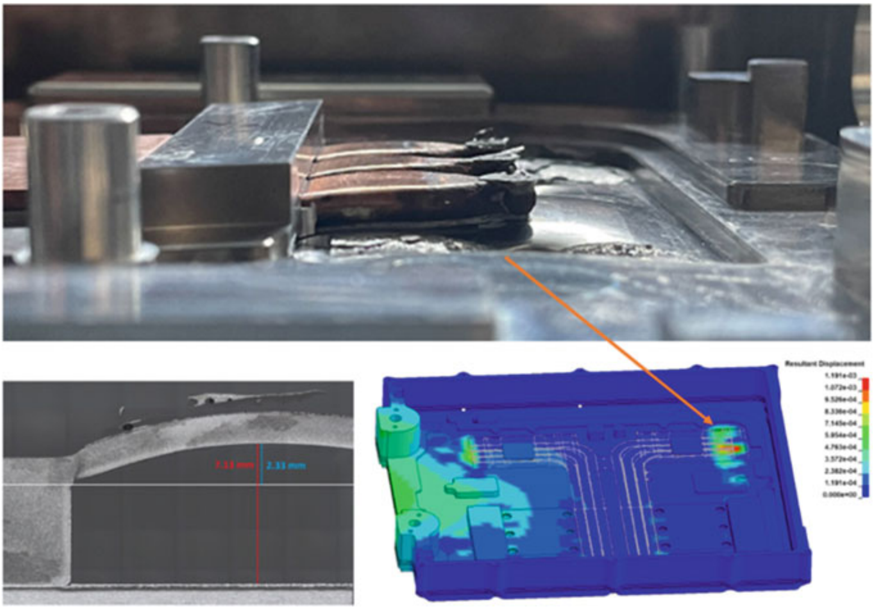


Fig. 35 Compute failure caused by coldplate buckling due to liquid thermal expansion under a hydrolocked state. Maximum in-plane stress of 134,121 psi was generated that caused buckling of 2.57 mm based on simulation, in good agreement with the actual cross-section measurement

Reliability has sometimes been classified as "how quality changes over time". Building Compute to achieve high reliability requires setting and achieving standards for precise process and assembly. Keeping Compute stresses within the design envelope during operation requires setting and meeting precise operating domain that delivers the least-stress operating performance. These standards are called quality standards. For AV Compute, if we want it to be highly reliable, we must first set the appropriate process and operating quality standards. Then we must achieve those standards. Hence, a high level of quality assurance is required to deliver its matching reliability. Setting and achieving world-class quality standards would bring world-class reliability. Higher reliability then brings higher safety. For a component to be automotive qualified, manufacturers have to meet specific industry standards throughout the manufacturing and testing process. IATF 16949 is a global automotive industry standard for such quality management and control. The automotive industry generally expects parts to be manufactured, assembled, and tested in IATF 16949 qualified facilities. However, currently, not all AV component suppliers and contract manufacturers are IAFT 16949 certified. The AV companies who are not traditional vehicle OEMs are also likely not IAFT 16949 certified. Furthermore, there is still a question on if IAFT 16949 is adequate for building AV hardware. AV's high reliability and safety standards require a matured supply chain with a higher level of the quality management system. The current IATF 16949 quality management system focuses mainly on the quality part, not being adapted to effectively include the security activities and safety aspects. This could be the main weakness of the current IATF 16949. Therefore, there still is a great challenge to integrate quality, security, and safety standards to synthesize a coherent quality management system for the development of AV Compute.

5.7 Security and Safety

Each AV is equipped with or supported by Compute to process the sensor data, monitor the vehicle's status, and control the mechanical components. Hence, the security threats against the Compute are of serious concern. Specifically, the attacks targeting AVs could cause fatal traffic accidents, and threaten both personal and public safety. There are many methods for AV attacks. How to defend against these attacks to ensure the safety and security of AV is of paramount challenge.

The safety of AV is at risk if security is compromised at any level. As each AV is equipped with numerous sensors and a Compute, an attacker targets one of the sensors, the Compute, or the communication networks to confuse, mislead, or even take over control of the vehicle under attack, leading to fatal accidents. It is extremely dangerous for any AVs to go on the road if it fails to meet the safety and security requirements. Generally speaking, it is extremely difficult to enter the Compute system. However, the vehicle infotainment system and the OBD-II port of the overhaul system are all connected to the CAN bus, and the CAN bus is connected to Compute, which allows hackers to enter Compute. The methods of attack include the following:

- Onboard diagnostics (OBD)-II intrusion: the OBD-II port is mainly used to diagnose the status of the vehicle, firmware update, and vehicle control. Usually, when the vehicle is in service, the technician will use the detection software developed by each car OEM to access the OBD-II port and exam the vehicle. Since OBD-II is connected to the CAN bus, as long as hackers obtain such detection software, they can easily hack into the vehicle system.
- Attack the AV remote control management platform: car schedule and resource allocation are all controlled by this cloud platform Therefore, once the platform is attacked by hackers, the entire AV dispatch and control system of a city may be disrupted, and traffic paralysis and accidents are prone to occur.
- Invasion of electric vehicle chargers: with electric vehicles becoming more and more popular, charging equipment has become an indispensable core component of the electric vehicle ecosystem. Since the EV charging unit will communicate with an external charging station during charging, and the charging unit will be connected to the CAN bus, this allows hackers to invade the CAN system through the external charging station.
- Car media player intrusion: there has been an attack case where the attack code is encoded into the burned music CD [22]. When the user plays the CD, the

malicious attack code will invade the CAN bus through the CD player, to obtain bus control and steal the core information of the vehicle.

- USB invasion: USB and other input and output interfaces. Plug a special USB into the car's USB port to complete certain car functions. If a USB is compromised with built-in chips, ROM, RAM, and wireless network functions, as well as written malicious control programs. If the line connection and signal transmission are large enough, and the Compute and other important ECU modules are involved, the safety of the AV and the safety of information can be damaged.
- Bluetooth intrusion: another entry point for attacks is Bluetooth. Nowadays, Bluetooth connection of mobile phones and car communication and entertainment systems has become standard. Since users can send and read information to and from CAN via Bluetooth, this also gives hackers a window to attack. In addition to gaining control of the owner's mobile phone, because the effective range of Bluetooth is 10 m, hackers can also use Bluetooth to carry out remote attacks.
- TPMS invasion: TPMS is a wheel pressure management system. Hackers can also launch attacks on TPMS. In this attack method, the hacker first places the attack code in the vehicle TPMS ECU, and then when the TPMS detects a certain tire pressure value, the malicious code will be activated to attack the vehicle.

A general solution is to encrypt and verify the information received by the Compute to ensure that the information is sent by a trusted MCU or component, not by a hacker. Using encrypted authentication, symmetric or asymmetric ciphers can be chosen. The symmetric cipher has a small amount of calculation, but the two parties in the communication need to know the cipher in advance. The asymmetric key does not require the password to be known in advance, but it is computationally intensive. Such additional safety authentication and encryption may cause Compute processing latency and communication timeout to impact AV operation. Therefore, it is necessary to consider increasing the delay caused by the safety mechanism while verifying the safety. Finally, the distribution and management of ciphers are also crucial but challenging. Although in recent years there have been some interesting proposals regarding protecting the security of AVs, more research is required before we deploy AVs on a large scale.

6 Summary

More and more fatalities associated with early developed AVs arise recently, which reveals the big gap between the current AV Compute system and the expected robust system for L4 and L5 full AD. In this chapter, we gave a high-level review of computing systems for autonomous driving, including an ADAS overview, centralized Compute system architecture, functional test and validation, and challenges. Safety and reliability are the most important requirements for autonomous vehicles.

Hence, the challenge of designing a Compute ecosystem for AVs is to deliver enough computing power, redundancy, and security to guarantee the safety and reliability of AVs while consuming less power.

Acknowledgements The authors want to thank Dr. Daniel Braun from BMW Group for his critical reviews of this chapter.

References

1. *World Health Organization*, [online] Available: https://www.who.int/news/item/11-12-2010-pedestrians-cyclists-among-main-road-traffic-crash-victims
2. *International Organization for Standardization (ISO)*, [online] Available: https://www.iso.org/standard/43464.html
3. Brian Krzanich. Data is the New Oil in the Future of Automated Driving. Intel Newsroom. 2016.11.15
4. Charles Murray. What's the Best Computing Architecture for the Autonomous Car? Design-News. 2017.08.17
5. Apollo Auto, https://github.com/ApolloAuto/apollo
6. https://apollo.auto/platform/hardware.html
7. (2018). *Meet NVIDIA Xavier: A New Brain for Self-Driving, AI, and AR Cars.* [Online]. Available: https://www.slashgear.com/meet-nvidia-xavier-a-new-brain-for-self-driving-ai-and-ar-cars-07513987/
8. (2020). *Enabling Next Generation ADAS and AD Systems.* [Online]. Available: https://www.xilinx.com/products/silicon-devices/soc/xa-zynq-ultrascale-mpsoc.html
9. *Texas Instruments TDA*. https://www.ti.com/lit/wp/spry272a/spry272a.pdf
10. (2020). *The Evolution of EyeQ.* [Online]. Available: https://www.mobileye.com/our-technology/evolution-eyeq-chip/
11. Cong Hao, et. al., 2019 IEEE International Workshop on Signal Processing Systems, pg 121–126, 2019
12. [Online]. Available: https://cloud.google.com/blog/products/gcp/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu
13. https://www.youtube.com/watch?v=j0z4FweCy4M&t=6750s
14. L. Liu, S. Lu, R. Zhong, B. Wu, Y. Yao, Q. Zhang, and W. Shi, IEEE INTERNET OF THINGS JOURNAL, VOL. 8, NO. 8, APRIL 15, 2021
15. (2017) Moving from 24 GHz to 77 GHz radar. [Online]. Available: https://www.edn.com/moving-from-24-ghz-to-77-ghz-radar/
16. S. Liu , L. Liu, J. Tang, B. Yu, Y. Wang, and W. Shi. Edge Computing for Autonomous Driving: Opportunities and Challenges
17. System on a chip. [Online]. Available: https://en.wikipedia.org/wiki/System_on_a_chip
18. x86. [Online]. Available: https://en.wikipedia.org/wiki/X86
19. Ultimate Guide to Real-time Operating Systems (RTOS). [Online]. Available: https://blackberry.qnx.com/en/rtos/what-is-real-time-operating-system/
20. Hualiang, et. al., ECTC 2021–1559
21. Endo, Tatsuo; Mitsunaga, Koichi; Takahashi, Kiyohum; Kobayashi, Kakuichi; Matsuishi, Masanori (1974). "Damage evaluation of metals for random or varying loading—three aspects of rain flow method". *Mechanical Behavior of Materials.* **1**: 371–380.
22. S. Checkoway et al., "Comprehensive experimental analyses of automotive attack surfaces," in Proc. USENIX Secur. Symp., San Francisco, CA, SA, 2011, pp. 77–92.

Overview of Packaging Technologies and Cooling Solutions in ADAS Market

Sandeep Sane, Shalabh Tandon, Erich Ewy, and Luisa Cabrera Maynez

Abstract Automotive electronics is among the fastest growing segments of semi-conductor industry (https://www.pwc.com/gx/en/technology/publications/assets/pwc-semiconductor-survey-interactive.pdf). It is primarily driven by increasing dependence on advanced electronics for a wide range of functions such as safety, control, and driver assist. The critical-to-function Advanced Driver-Assist System (ADAS) must meet stringent reliability and thermal requirements while providing high-performance deterministic compute. This chapter will cover the packaging technologies employed to meet the lifetime performance, quality, and reliability demands of automotive applications, as well as the system packaging and thermal management strategies for control units that dissipate between 10 and 500 W of power depending on the application. The high operating ambient temperatures for passenger vehicles can range from 65 to 85 °C, and this influences the choice of thermal solution ranging from low-power fan-less designs, mid-power forced-air designs to high-power liquid-cooled solutions. While significant innovations in technology for in-vehicle infotainment to autonomous driving are happening today, more are required in the near future to translate into reality the visions of automobile design that will revolutionize the auto industry and more broadly the way we live.

S. Sane (✉)
3563 S. Cox Court, Chandler, AZ 85248, USA
e-mail: sandeep.b.sane@intel.com

S. Tandon
4033 S. Windstream Place, Chandler, AZ 85249, USA

E. Ewy
3914 E. Cherokee St., Phoenix, AZ 85044, USA

L. C. Maynez
4564 S. Emerson St., Chandler, AZ 85248, USA

Y. Li and H. Shi (eds.), *Advanced Driver Assistance Systems and Autonomous Vehicles*,
https://doi.org/10.1007/978-981-19-5053-7_4

1 Introduction

Commoditization of chips and the consumer desire to have latest technology available in automobiles is leading to increased usage of semiconductors in automobiles. Transportation medium such as vehicles (even trucks) are gradually transforming from mechanical chassis on wheels to sophisticated devices that are akin to *server on wheels* given the total computational power that will be required to operate autonomous vehicles. The inclusion of semiconductors to manage entertainment functions, safety controls to autonomous features are increasing the cost of chips such that they may cost about a third of the total value of the entire vehicle [2]. This automotive evolution is due to three vectors: (1) desire for autonomous vehicles that are safe to operate, (2) increase in demand for high-speed communications, infotainment, and secure, and (3) the transformation to environment friendly electric vehicles. The authors will primarily focus on the first two vectors in this chapter.

1.1 Market Opportunity and Trends

Multiple surveys and analyses have been conducted to understand the overall market opportunities available for semi-conductor industry. One such survey is shown in Fig. 1. This was the analysis conducted by Morgan Stanley on the impact of new auto industry paradigms on social and economic benefits. The survey shows that the benefit of self-driving vehicles to the US economy is greater than 1 trillion dollars. This is due to a combination of accident reductions, productivity gains, congestion reduction, and improvements in fuel efficiency and savings, to list a few.

The actual social-economic benefits will likely depend upon the how the automotive industry evolves, both the rate of transition to all-electric vehicles, as well as the wider adoption of autonomous features as well as fully autonomous vehicles. If history is any guide, the transformation to both electric vehicles and autonomous driving will happen, albeit it gradually. Adoption of safety features such as seat belts,

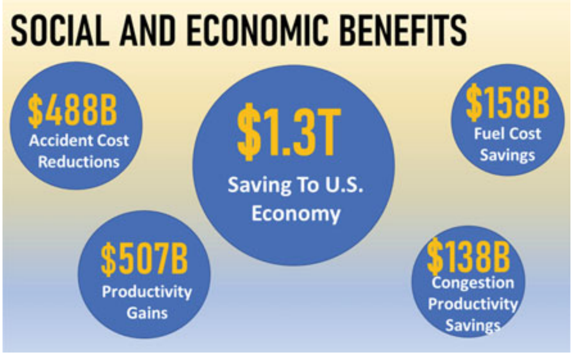


Fig. 1 Social and economic benefits autonomous cars. *Source* Morgan Stanley

Fig. 2 Evolving market for auto-industry

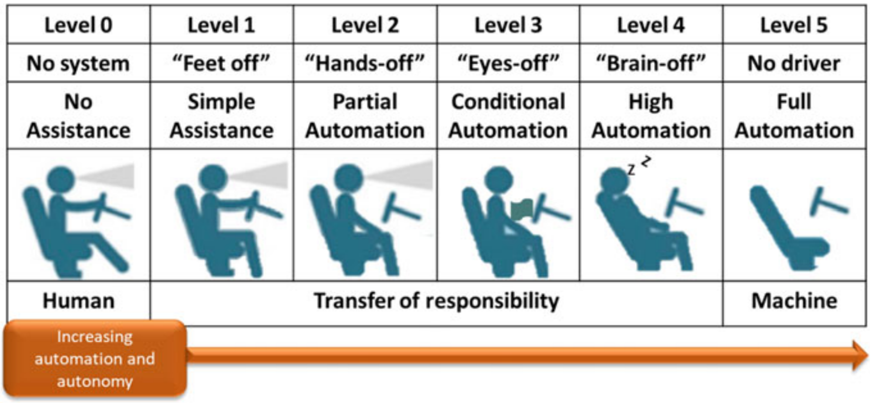
airbags took years, if not decades for general adoption. Similarly, market dynamics and business models of transformation will determine how fast self-driving vehicles become the norm and create the new business opportunities. Figure 2 shows the benefits of autonomy in the form of personal vehicles as well as fleet accommodations (taxis) that are likely to bring business opportunities. The traditional car market is likely to continue with a potential self-driving option with a concurrent evolution of a new service referred to as Transportation As a Service (TaaS or also termed as Mobility as a Service, MaaS) as the autonomous features and ingredients become more cost effective and reliable instead of owning a vehicle, many next generation consumers may opt for a reliant TaaS service or fully autonomous vehicles to meet their needs, as is evident with the growth of Uber/Lyft options in big cities [3]. Another factor enabling automation is ability to deliver content fast. It is expected that passengers would want high bandwidth, efficient and secure connectivity for entertainment, business and safety reasons. The demands and requirements for infotainment segment are also evolving fast.

These disruptive trends in auto-industry are only possible due to the availability of high computation power at reasonable cost, high bandwidth memory with low latency and fast and high bandwidth network connectivity that delivers a high quality and high reliability experience to the user. Overall, it indicates ~$70B TAM opportunity and is one of the quickest growing segments in the semi-conductor industry.

1.2 Road to Autonomy: ADAS Architecture

Society of automotive engineers defines levels of autonomy from L1 to L5. The requirements and definitions for these five levels are illustrated in Fig. 3. As the level of automation expands, compute demands increase dramatically from L3 to L5.

While there are numerous methods to reach L4-L5 autonomy, all of them require some common, basic ingredients. Figure 4 summarizes these ingredients into five



Level 0	Level 1	Level 2	Level 3	Level 4	Level 5
No system	"Feet off"	"Hands-off"	"Eyes-off"	"Brain-off"	No driver
No Assistance	Simple Assistance	Partial Automation	Conditional Automation	High Automation	Full Automation
Human	Transfer of responsibility				Machine

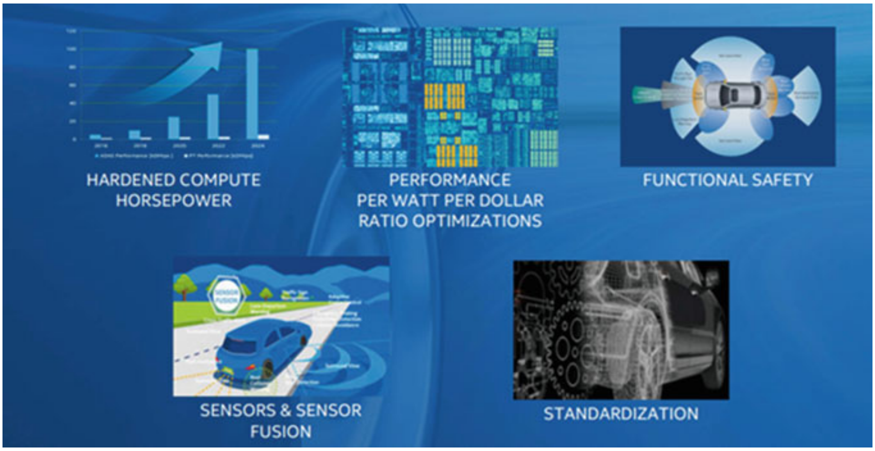
z z
Increasing automation and autonomy

Fig. 3 Vehicle automation levels [4]

categories, namely, standardization, functional safety, performance per watt per dollar, compute power, as well as sensors and sensor fusion.

The components highlighted require a focused development effort to meet the high automotive expectations of quality and reliability that are different from general personal computer requirements where these components are typically used. For uniform application and availability, it is crucial to exploit standards that allow multiple component suppliers to design-in to the requirements upfront. Without a concerted effort in the industry, enabling these technologies at reasonable costs for mass production may remain elusive.

The functional building blocks critical to enable autonomous driving are categorized into four buckets: specifically, acquisition, perception, cognition, and action.



HARDENED COMPUTE HORSEPOWER
PERFORMANCE PER WATT PER DOLLAR RATIO OPTIMIZATIONS
FUNCTIONAL SAFETY
SENSOR FUSION
SENSORS & SENSOR FUSION
STANDARDIZATION

Fig. 4 Prerequisite components to achieve L4-L5 autonomy in driving

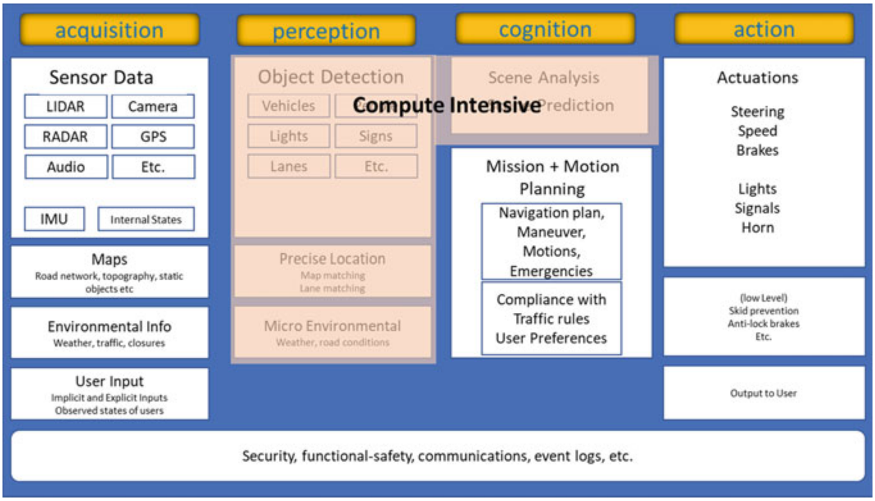


Fig. 5 Illustration of necessary building blocks for autonomous vehicles

The acquisition function represents a host of sensors that provide data on the immediate surroundings of the vehicle. Perception and cognition are essentially internalization of the data and subsequent processing to translate raw information to an actionable element by the vehicle (HW and SW) such as braking (if an obstacle), turning (if directional), and accelerating. While each functional block brings unique challenges, it is the cognition and perception states that are computationally intensive and requires low latency for safe and consistent vehicle operation. Figure 5 illustrates these functional blocks pictorially.

While it is always hard to define clear key performance indicators (KPIs) for any nascent market, it is particularly difficult to do so for ADAS type requirements because of the complicated interplay between customer needs, regulatory bodies, safety compliance, cost elements to be superseded by geo-politically drive regulatory compliance needs. Metrics of interest, as understood from studies conducted by Waymo, Mobileye, and Academia [5] can be categorized as:

1. Fast response times—low end-to-end latency determined as <100 ms relative to human response times of approximate 600 ms.
2. Performance Predictability: tail latency numbers (99.99%) should be less than or equal to 100 ms.
3. Navigation being critical to autonomy, high-definition maps drive high-storage requirements within the vehicle rather than relying on connectivity to access cloud-based maps. Best estimates: range from 40 to 50 TB of storage capability necessary in the vehicle.
4. Thermal Power: Platform and component thermal management to assure proper dissipation of heat to deliver seamless computational functionality and vehicle performance.

5. Extremely high availability driven by high quality and reliability of semiconductor components.
6. Power budget: compute efficiency required to minimize power consumption and significantly reduce impact to vehicle fuel efficiency.

As a typical scenario, the computational element in the vehicle (e.g., a server) must consume all the sensor (such as LiDAR, Radar, and camera,) data analyze it with high fidelity, provide the vehicle a frame of reference (its location) and define a safe course of action. This must be achieved continuously, with low latency (~100 ms) and extreme reliability for a safe vehicle operation. These requirements drive the overall vehicle computational demands, as well as impose strict requirements on network, sensors, and associated software stack up to optimize performance with good reliability. The amount of computational power required relies on the quantity and quality of software stacks (AI/ML algorithms), security requirements, input data, etc. The computational and bandwidth requirements will depend upon the level of autonomy the vehicle has and can increase many fold from L1 to L5, as highlighted in Fig. 6. One of the critical challenges will be power and thermal requirements for these computations platforms since both are essential to delivery consistent performance and reliability.

The new trend in automotive electronics requires high performance with high reliability with extreme out-going quality in terms allowable defects per million samples; this poses both a significant challenge and an opportunity for semi-conductor industry both for silicon and packaging technologies.

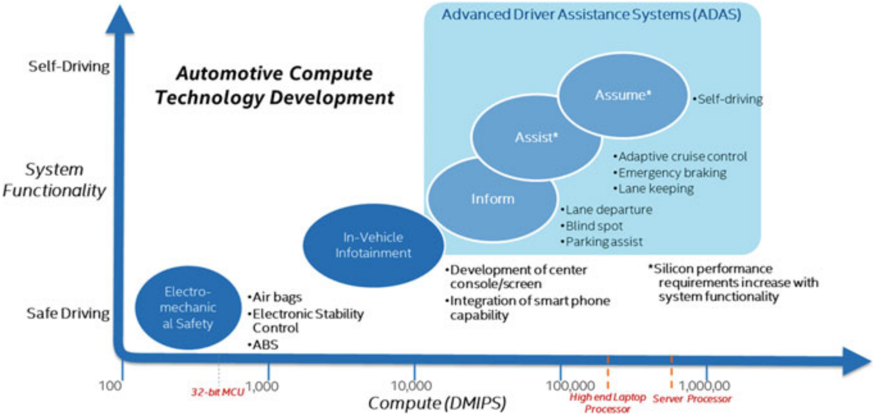

Automotive Compute Technology Development
Self-Driving
System Functionality
Safe Driving
Advanced Driver Assistance Systems (ADAS)
Assume*
•Self-driving
Assist*
•Adaptive cruise control
•Emergency braking
•Lane keeping
Inform
•Lane departure
•Blind spot
•Parking assist
In-Vehicle Infotainment
•Development of center console/screen
•Integration of smart phone capability
*Silicon performance requirements increase with system functionality
Electro-mechanical Safety
•Air bags
•Electronic Stability Control
•ABS
100
32-bit MCU
1,000
10,000
100,000
High end Laptop Processor
Server Processor
1,000,00
Compute (DMIPS)

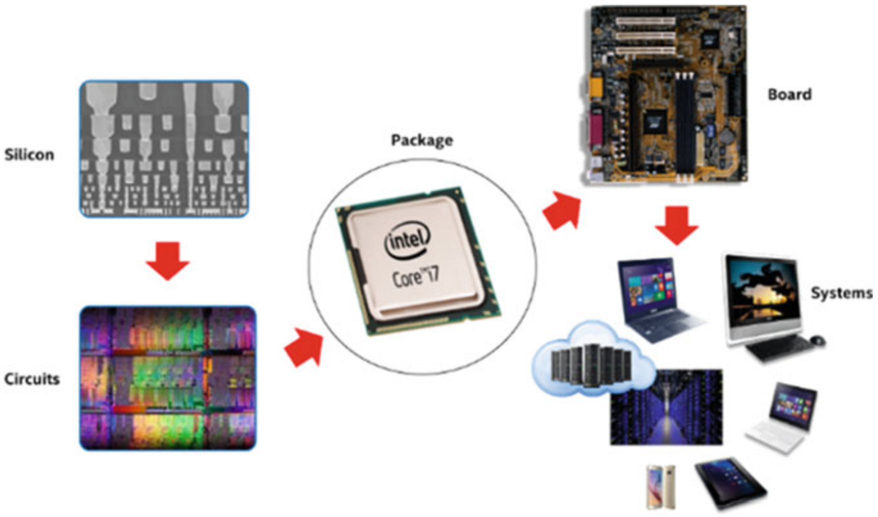
Fig. 6 Graphical representation of increased system functionality requiring higher computational (DMIPS—Dhrystone million instructions per second) power [6]

2 Package Technology and AD Requirements

Packaging technologies in semiconductor industry is going through massive changes, and it is imperative to first understand the board trends to determine packaging technology needs for AD market. The explosive growth in the semiconductor industry in the last couple of decades has been the result of consumer demand for products that deliver user experiences in entertainment and communication. This growth has been enabled primarily by consistent transistor node scaling. Traditional role of electronic packaging is to act as a bridge between the silicon and platform (or system) providing space transformation between the dimensionally fine interconnect features on the silicon and the significantly coarser interconnect features on the system board (Fig. 7). The package also serves to provide mechanical protection to the fragile silicon while facilitating power delivery to and power removal from the semiconductor device. It enables the transmission of high bandwidth signals to and from the silicon. To keep up with Moore's Lay, aka cadence with transistor density increase, the packaging technologies evolved from wire bond (low IO density) to flip chip technologies (higher IO density) and from ceramic substrates to organic to meet increasing demand for I/O count and chip performance demands. Detailed review of packaging technologies outside the scope of this chapter, readers are encouraged to review any textbook on packaging technology [7].

In recent years, demand for advantaged packaging technologies is increasing primarily driven by product need for package-level heterogeneous integration. There are primarily four reasons for these trends, namely: (1) the need to integrate dissimilar Si-technologies/IP blocks on a single package, (2) Need to deliver monolithic SOC like performance by reducing power and latency between chiplet by using smaller interconnect technologies, (3) Improve time to market using the chiplet approach and ability to re-use existing IPs, and (4) desire to build in yield resiliency vis-à-vis the introduction of advanced Si-Technology nodes. These vectors continue to increase demand for advanced technologies, thus driving significant advancements in packaging technologies such as Foveros, EMIB, CoWoS, Info_oS, and many others [8].

In general, advanced packaging technologies are delineated into two categories, two-dimensional (2D) (note: some packaging technologies are referred as 2.5D packaging and are lumped together with 2D technologies as they also offer lateral die-to-die connection) and three-dimensional (3D) technologies. In the former case, lateral connections between two dice are achieved by mounting the dice on an organic multi laminate substrate as shown in Fig. 8. In 3D technologies, vertical connections are achieved across dice by stacking the dice on top of each other, and the die complex is then mounted on an organic multi-laminate substrate as shown in Fig. 9. Packaging technologies such as multi-chip package (MCP), Fanout, or EMIB technology that delivers planar connections falls in 2D category while interposer-based technologies such as Intel's Foveros or TSMC CoWoS that offer out-of-plane connectivity falls in 3D category.



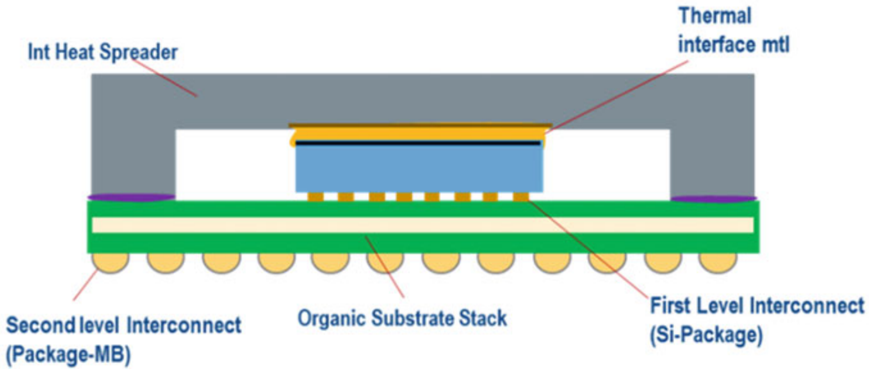

Int Heat Spreader
Thermal interface mtl
Second level Interconnect (Package-MB)
Organic Substrate Stack
First Level Interconnect (Si-Package)

Fig. 7 Illustration of a typical cross section of flip chip (land or ball) grid array, FCxGA package

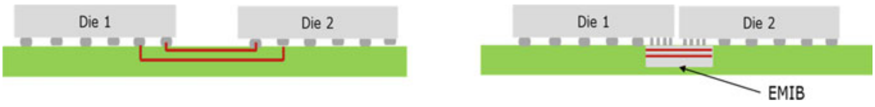


Fig. 8 Lateral or planar die-to-die connection between two dice are categorized as 2D package technologies

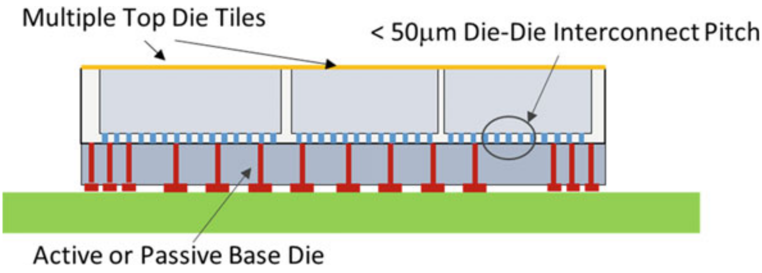


Fig. 9 Die-to-die connections in vertical as well as planar between two dice; such technologies are categorized as 3D packaging technologies

While the traditional fan-out packaging and MCP technologies are currently mainstream with significant volumes of products, the demand for advanced packaging technologies based on Si-Interposer, fan out bridge, and EMIB is increasing abruptly as the desire to use mixed Si-node technologies and product complexities with die-disaggregation grows. It is anticipated that over the next few decades, these advanced packaging technologies are likely to be adopted aggressively over the next decade because of the increased utility they offer. Figure 10 displays the evolution of advanced microelectronic packaging technology driven by machine learning and/or artificial intelligence needs. The various heterogeneous packaging technologies provide more freedom to the product architects for TTM improvement, cost reduction, and performance optimization [9].

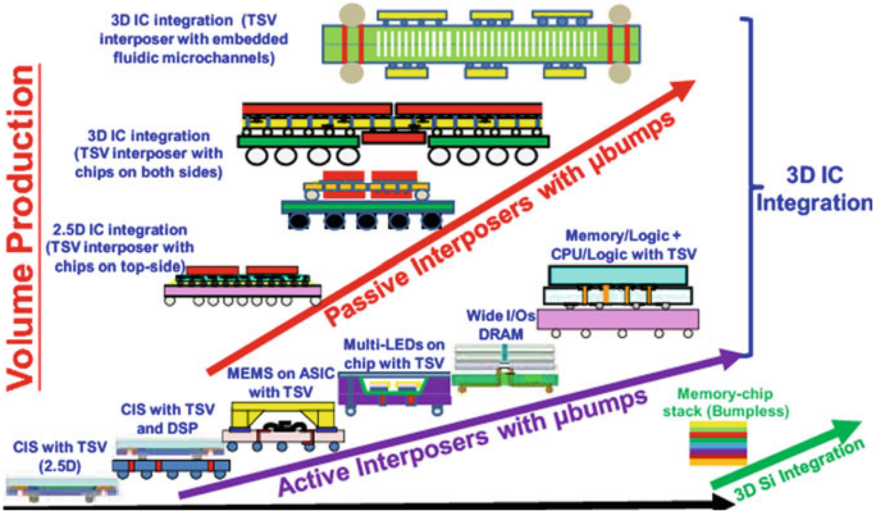

3D IC integration (TSV interposer with embedded fluidic microchannels)
3D IC integration (TSV interposer with chips on both sides)
2.5D IC integration (TSV interposer with chips on top-side)
Volume Production
Passive Interposers with µbumps
3D IC Integration
Memory/Logic + CPU/Logic with TSV
Wide I/Os DRAM
Multi-LEDs on chip with TSV
MEMS on ASIC with TSV
CIS with TSV and DSP
CIS with TSV (2.5D)
Active Interposers with µbumps
Memory-chip stack (Bumpless)
3D Si Integration

Fig. 10 Schematic of the packaging technology growth trends driven by AI/ML product demands and high performance computing

The mainstream products such as server and laptops are also benefiting from these packaging technology advancements since they allow integration of multiple chiplets at high interconnect densities to provide greater IO performance. Given the rate of increase in functionality, the expectations in the ADAS market segment are similar. The following sections will describe what impact these global trends in package technologies will have on the needs of the AD market segment.

2.1 Role of Advantage Packaging Technology in AD Market

As mentioned earlier, traditional mechanical vehicles are transforming into autonomous systems that are akin to server on wheels with high compute power, bandwidth, latency, and thermal requirements. Integration of silicon on various 2.5D/3D package architectures are already in development today, with mass production expected over the next 5 years, demanded by main-stream product segments [10]. For autonomous driving to proliferate and not be limited to luxury vehicles, delivering good performance at an affordable price is critical. This implies system on chip (SOC) approach for the AD market utilizing chiplet architecture or re-use and using advanced packaging technologies to integrate all together. Figure 11 shows current integration trends moving from board level to onto package level. Heterogeneous integration with advanced packaging technologies allows end products to deliver high bandwidth at low latency and low power and provides a compact platform to integrate IP from different vendors built on different silicon technology nodes. Additionally, heterogeneous integration improves product time to market.

Figures of merit (FOM) to judge microelectronic packaging technology utility are two-fold: the first includes performance based FOM for example die-to-die I/O power, interconnect densities, power delivery schemes, off-package I/O scalability, and cooling capabilities. The second includes reliability, yields, manufacturability, and cost. These FOMs need to continue to improve to meet future product demands. However, it is also critical to understand the trade-offs between performance and manufacturing based FOMs for the delivery of a viable product that balances the product performance and affordability goals. For example, re-using the IP or chiplet

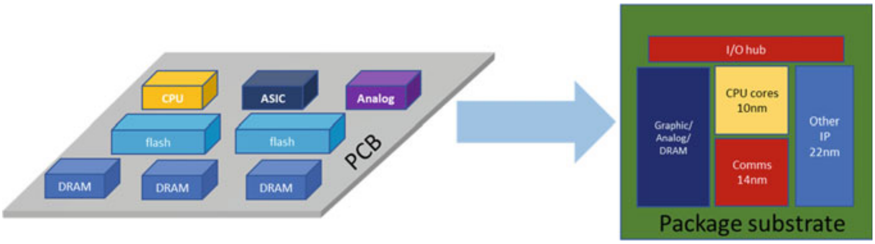


Fig. 11 Board level to package level integration with adv. Packaging technologies; delivers higher bandwidth with low power and latency, integrates different IP blocks to deliver functionality

from server market to AD market for development cost and time reduction is a good idea. However, it is imperative that the IP blocks in the chiplets need to meet to meet the AD specific reliability requirements, not just server requirements. At the same time, advanced microelectronic packaging technologies need to be developed including materials, design rules, process conditions, and metrologies to meet AD reliability requirement and DPM goals. This will be discussed further in Sect. 4.

2.2 Smaller System Level Footprints

AD customers will prefer compact form factors given the lighter weight and easier fitting capability into the automobile given the space and weight constraints of a vehicle. Balancing product functionality at diminutive package footprints does increase silicon to package area ratios bringing unique challenges in developing assembly processes at package and mother board levels, scaling board and substrate DRs such as BGA pitch, pad, line/space to allow breakout/fanout, innovative materials to continue to deliver high reliability, and high yielding processes [11] (Fig. 12).

3 Thermal Management

AD market thermal management requirements are unique and require significant innovation on package/system level. The high computation requirements drive high TDP on one hand, while harsh environmental conditions in the car, having ambient temperature significantly higher relative to the conventional server use conditions (e.g., in cloud server applications), further acerbate the heat dissipation challenge. Figure 13 presents a comparison of thermal boundary conditions between a conventional server environment vs an automotive environment. Typically, in a server

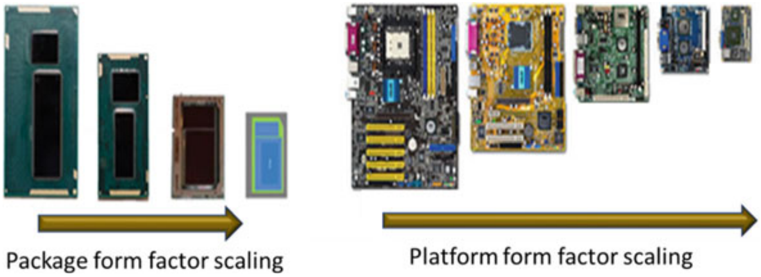


Fig. 12 Demonstrate the scaling of package and platform footprint in client markets. Similar trends and demands are expected in AD market segment

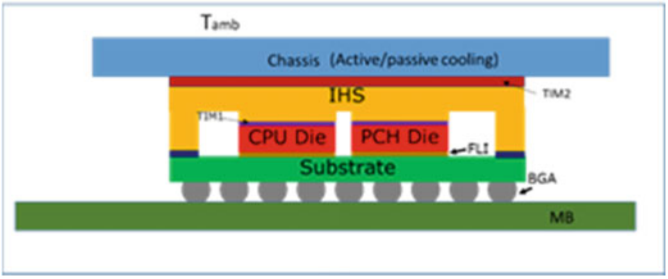


Fig. 13 Thermal head room comparison between a typical client and server market versus AD or Infotainment segment. Note that thermal head room reduced by 15–20 °C

ambient temperature (Tamb) is ~40–45 °C which gives a thermal headroom of 60–65 °C (with Tjmax ~105 °C). In comparison in AD segments, Tamb estimates ~75–80 °C, thus giving thermal headroom of less than 50 °C (with Tjmax ~125 °C). Moreover, any additional power consumed for system cooling purposes takes away power from overall car, likely impacting fuel efficiency.

- For Server/Client Market: Tamb ~40–45 °C
  - Tja budget (thermal headroom): 60–65 °C (assuming Tj: 105 °C)
- AD/IVI Market: Tamb ~75–80 °C
  - Tja budget (thermal headroom): 45–50 °C (assuming Tj: 125 °C)

Therefore, the AD market poses a unique challenge for package as well system level thermal management. Innovative package and system level solutions are necessary, such as high conductive thermal interface materials, heat pipes, heat sink design, and materials. There also must be a strong collaboration between silicon, package, and platform design teams to deliver a co-optimized SOC Silicon floor plan to system layout to assure thermal constraints are resolved successfully.

3.1 ECU Thermal Fundamentals

From a thermal design perspective, the typical controller employs a common heat sink approach. The mechanical assembly of the controller housing is sometimes referred to as a “clam shell” design, where the printed circuit board assembly is held within two halves or a housing assembly, a simple schematic is shown in Fig. 14. This type of construction provides a level of ingress protection from particulates/dust and moisture, and it also provides a thermal environment with a common heat sink design where components that need thermal management all contact the same housing wall. In this configuration, there will be conduction heat transfer from the component into a common heat sink (housing wall). At the housing external surface, heat transfer is

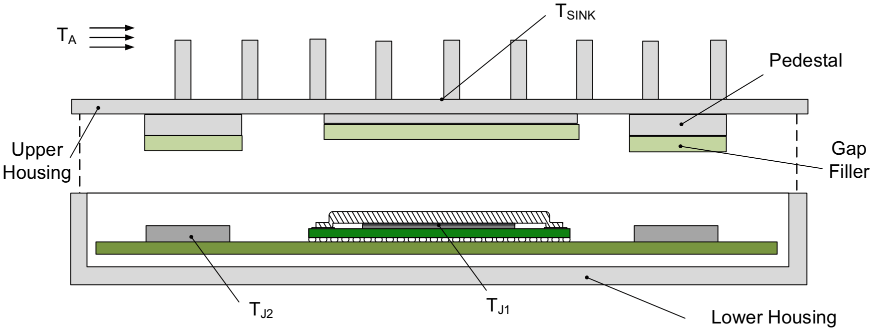


Fig. 14 ECU controller with components contacting a common sink (chassis wall) through gap fillers

accomplished using convection or convection/radiation to the local vehicle ambient is realized.

Internally, all heat is transferred by conduction from the silicon, through packaging materials (TIM1, integrated heat spreader or lid), through the gap filler (TIM2) and into the housing heat spreader/wall. The housing material is typically cast aluminum which has good heat spreading capability.

Since all components conduct heat into a common heat sink, mutual heating effects must be considered. For example, a high-performance processor with an operating junction temperature of 125 °C will raise the operating temperature of nearby components like memory and power supply controllers since they are all “connected” to the same heat sink.

Externally, depending on the dissipated power of the controller, operating ambient temperature, and available area for heat transfer, there are several options for the system (controller) thermal solution:

- Free convection (fanless): when system power is less than ~15 W, the controller can operate under free convection conditions.
- Forced convection: when system power is as high as ~150 W some means of forced airflow will be required, either an integrated fan/blower or airflow provided by the vehicle.
- Liquid cooling: very high power ECUs (>150 W) will have to rely on liquid cooling for overall thermal management.

From a thermal model perspective, the controller with a common heat sink can be represented as a thermal resistor network as shown below in Fig. 15. While not applied directly in terms of determining operating points, the model is useful to show relative resistances (thermal characterization parameter psi).

From component junction/case to sink, each component (1, 2, …, *n*) will have a unique psi_js which is a function of package type (bare die, lidded, overmold, etc.) and interface material used (TIM2) between the component and the heat sink. From sink to ambient (or fluid for liquid cooling), the magnitude of the heat sink characterization

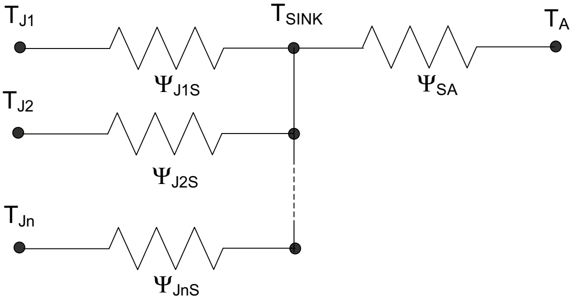


Fig. 15 Simplified, common heat sink ECU thermal model

parameter psi_sa indicates system thermal solution type, with approximate ranges given here:

- Free convection (fanless): Ysa > 3 °C/W
- Forced convection: Ysa 0.3 to 3 °C/W
- Liquid cooling: Ysf < 0.3 °C/W (references minimum fluid temperature '*f*' instead of ambient)

Early in the controller design phase, it is imperative to know how a controller will operate thermally. If requirements call for fanless operation, for example, steps must be taken early to ensure components operate with their respective specifications for junction/case temperature under free convection conditions at maximum operating ambient. An assessment of controller thermal performance can be realized with a detailed model using computational fluid dynamics (CFD) software. Running such a model, where component operating temperature is of interest, requires component package characteristics or detailed model, component power, system configuration/layout (mechanical assembly), interface material properties, housing material, and operating ambient temperature. Figure 16 shows an example of a CFD model for a fanless system. Note that under the assumed boundary conditions with include layout details, component placements and orientation, motherboard design, prescribed power maps, the peak temperature of the component is predicted to be 116 °C.

Developing a controller system thermal model requires close collaboration between disciplines: mechanical, thermal, hardware and system engineering. Early assessments of ECU thermal performance are critical when there is time to make tradeoffs in component/system power, material selection, PCB layout, housing size/features and eventually get to a point of optimization once the design is deemed thermally feasible.

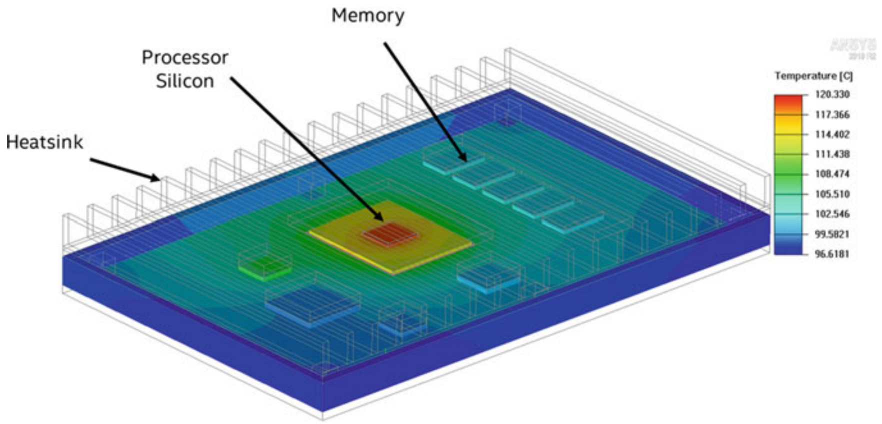


Fig. 16 CFD analysis of a fanless ECU showing a peak component temperature of 120 °C (limit is 125 °C)

3.2 Component Level Fundamentals

The increasing reliability demands of automotive segments require that package thermals be addressed in a separate manner from that of system cooling in the ECU. A focused approach to package thermals allows for a detailed characterization on the effects of interface performance and results in shorter turnaround times for optimization. This is due to a potential simplification that can be achieved in package-focused simulation models.

When addressing the specifics of package thermal analysis, details such as thermal interface performance and power dissipation should be included for in a comprehensive manner. Overall, packaging components are key to maintaining device performance as power densities in automotive applications are typically highly variable-ranging from <1 W/cm^2 to >400 W/cm^2 [12].

The analysis of a lidded package may then be independent from that of the cooling solution through a discretization of the model, where an assessment is performed with simplified approximations of system conditions. From a package-centric perspective, the resistance network in Fig. 17 is rearranged (as compared to Fig. 15), where total resistance Psi-ja is divided into junction to case (psi-jc) and case to ambient (psi-ca) resistance. The junction to case resistance is composed by the impedance of the die, TIM1 and integrated heat spreader (IHS).

Computational methods can be employed to simplify the model such that results are quickly integrated into the cooling solution approximations of thermal performance of the ECU. This allows for quick and easy estimates, to be able to evaluate multiple power scenarios and package layouts during the design stage. An example of this is defined in Fig. 18, showing a schematic of a potential approach for package thermal assessment. The characterization can be approximated through a conduction-based model spanning from the bottom of the second layer interface to the top of

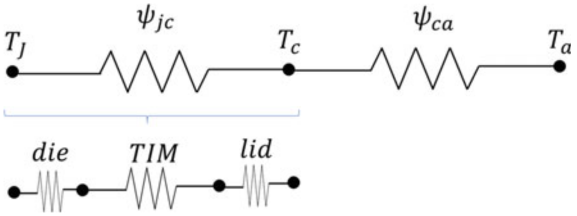


Fig. 17 Simplified resistance network of a package thermal model

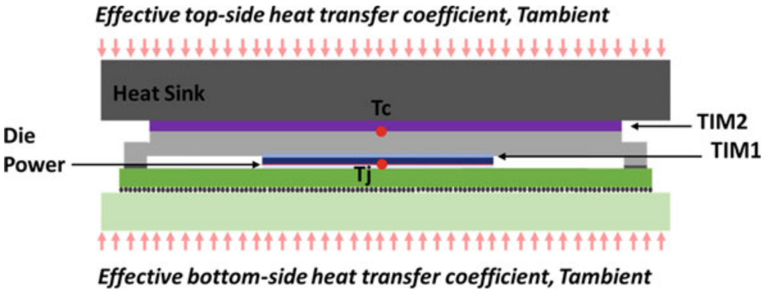

Effective top-side heat transfer coefficient, Tambient
Heat Sink
Tc
TIM2
Die
Power
TIM1
Tj
Effective bottom-side heat transfer coefficient, Tambient

Fig. 18 Basic approach schematic for a conduction-based package thermal model

the integrated heat spreader as the region of interest. Specifications of local package features and conductivities are modeled in detail while the system simplified into a set of local boundary conditions.

The psi-jc metric is calculated based on the simulation results using (1), where Tj is the highest temperature at the bottom of the die, and Tc is defined to be a point atop the integrated heat spreader. TDP corresponds to the thermal design power of the package.

$$\psi_{\text{JC}} = \frac{T_J - T_C}{\text{TDP}} \tag{1}$$

The first step toward a thermal-focused package design is to take into consideration the resistance budget. Overall, psi-jc allocation is closely related to the ambient temperature, TDP, and the type of cooling solution employed. The specific application and its corresponding psi-ca will dictate the total allowed package thermal resistance.

Table 1 presents a comparison of sample products in the automotive space for ADAS, IVI, and AD applications. This table demonstrates a set of representative thermal requirements assuming typical values of a product in the subsegment. For a high-powered AD package (product C) the total thermal budget is highly constrained. A case like this warrants the use of a liquid cooling approach with a psi-jc budget under 0.1 C/W. With such a tight margin the package design variables are highly impactful toward the total resistance. In comparison, a low-powered package as seen in products A and B may allow for a less expensive approach to system cooling

Table 1 Thermal requirements for sample products in the automotive space

	Product A	Product B	Product C
Application	ADAS	IVI	Autonomous (AD)
Cooling solution type	Fanless	Fan(s)	Liquid cooling
Tjmax (C)	125	110	105
Tambient (C)	85	65	65
TDP (W)	10	12.5	200
Psi-ja thermal budget (C/W)	4	3.6	0.2
Psi-ca (C/W)	3.3	2.5	0.1
Psi-jc budget (C/W)	0.7	1.1	0.1

with more headroom in package optimization. Overall, the relative contribution and impact of decisions in packaging design must be carefully addressed.

Designing a lidded package in the automotive space requires a consideration of high-impact factors that specifically affect the junction to case resistance. Some of these are outlined as follows:

1. Temperature limitations: A high ambient temperature reduces the overall psi-ja budget and increases current leakage, through an incremental Tj, negatively affecting psi-jc. Increased Tjmax parameters will result in aggressive degradation within the package [13].
2. Extended reliability requirements: Harsh environment constraints for packages require operation beyond traditional consumer electronics [14]. The required survivability parameters of automotive systems result in an accelerated degradation of the components within as compared to their use in traditional segments.
3. Power distribution parameters: The effect of local power density is significant in the context of the automotive application. This is because a reduced thermal budget is directly correlated to package resistance having a higher proportional impact. A die power distribution with a high nonuniformity will result in a hot spot augmentation increasing the psi-jc metric.
4. Geometric constraints: Height related form factor constraints reduce opportunities to optimize lid thickness for thermal spreading. A reduction in package and die footprints will likely result in an increased power density.

The main objective in package thermal design is to increase temperature uniformity prior to reaching the TIM2 layer. This is largely achieved by focusing upon the thermal bottleneck of TIM1. Automotive use conditions such as ADAS require that the package withstand a lifetime of >10 years under an aggressive thermal environment [15] reaching temperatures of up to 125 °C. While cost is a driving factor in the design of ECUs, it is necessary that TIM1 selection be made with consideration of performance at these extended requirements. Reliability modeling is employed to predict thermal degradation of the material focusing upon areas of high risk, such as die edges and corners.

When establishing psi-jc requirements, thermal performance of TIM1 must be coupled with the distribution of power within the die to determine the location and value of maximum temperature. If the areas of TIM degradation coincide with regions of high-power density in the die, the overall resistance of the package will be maximally increased, reducing the resistance budget. As such it is advantageous to optimize the die floorplan early in the design stage with the purpose of reducing thermal resistance. The level of detail in package modeling is adjusted to account for both local power density and TIM1 degradation effects.

Additional steps may be taken during the package design stage such that psi-jc is minimized. Lid thickness can be maximized for increased spreading, without violating any z-height constraints. Package size may also be increased such that the surface area in contact with the system is maximized as to enhance the effect of the cooling solution. Overall, the factors described in this section must be carefully taken into consideration when approaching the design of a package in for automotive applications. A balance must be achieved between project constraints (e.g. cost and size) and the thermal resistance budgets set by the segment.

3.3 Vehicle Operating Environment

Given a world-wide distribution of vehicle operating conditions, where most of the population may be operating a vehicle, the ambient operating range is −40 to + 55 °C. At the cold end of the operating range, −40 °C is the minimum "cold start" condition. ADAS controllers are expected to "boot" once cold-soaked at −40 °C. Similarly, the maximum operating ambient can be as high as 55 °C as seen in low elevation desert environments. However, the vehicle internal temperature can be 20–30 °C hotter than the ambient temperature due to solar loading where the vehicle color, window exposure, orientation, etc., all contribute to the internal temperature rise above ambient.

ADAS controllers are typically located either in the vehicle cabin or in the trunk (boot). Given the relatively complex component packaging of ADAS controllers, the "under the hood" environment is too harsh, especially for vehicles that utilize internal combustion engines where air temperatures can reach >150 °C. Compared to the "under the hood" environment, the vehicle cabin and trunk environment is relatively controlled. The cabin has an element of environmental control through the HVAC system for passenger comfort. However, control units are placed "out of sight" as to avoid passenger contact and other potential in-cabin hazards such spilled drinks and the like. As such, control units are placed under seats, behind the dashboard or behind the A-pillar kick panels for example (Fig. 19).

The in-cabin operating temperature range is typically from −40 to +65 °C. The hot end of the temperature range can occur in both the summer and winter months. A hot-soaked vehicle can have an internal ambient reaching 65 °C. Although the vehicle

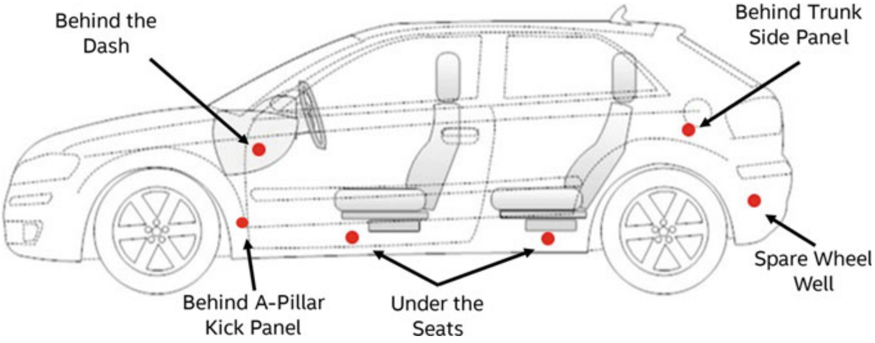
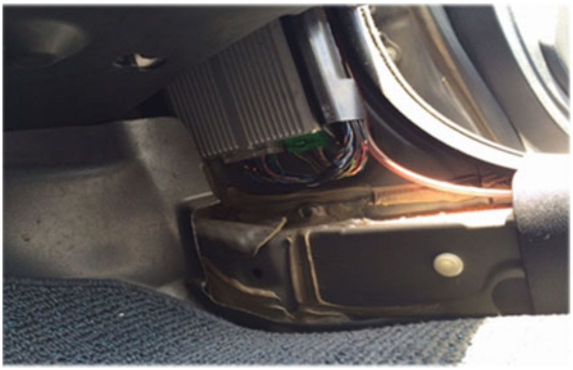


Fig. 19 Typical ADAS control unit locations

Fig. 20 ADAS controller behind passenger a-pillar kick panel



HVAC will likely be on once the vehicle starts, controllers located under seats or behind kick panels are not directly exposed to conditioned air. In winter, a sustained high temperature of 65 °C may occur when the vehicle heat is on HIGH, for example where a control unit is located near the HVAC return ducts behind the dash (Fig. 20).

For ADAS controllers located in the trunk, the operating temperature range is typically from −40 to +85 °C. The hot end of this temperature range is about 20 °C hotter than the cabin due to the lack of vehicle HVAC control as well as mutual heating effects from other nearby equipment like an audio amp in the same compartment. Controllers are placed behind panels, again “out of sight” to avoid any damage from items loaded into the trunk as well as items moving around the trunk unconstrained during travel, not to mention safety considerations to protect consumers from coming in direct contact with surfaces that may have a surface temperature approaching 100 °C.

3.4 ADAS ECU Thermal Management

There are several options employed today for modern passenger vehicles for ECU thermal management. Depending on system power dissipation, thermal management may be achieved by:

- Natural convection (fanless)
- Forced convection (integrated fan/blower or system airflow)
- Liquid cooling

Fanless Example/Discussion

A fanless design is preferred as this the best option for long term reliability (since fan reliability is not an influencing factor), quiet operation, and relatively simple vehicle integration. Most ADAS controllers employ a fanless design for thermal management. To operate in a fanless mode when the maximum operating ambient can be as high as 85 °C requires auto-grade components with high operating specifications, e.g. processor Tj-max of 125 °C, low system power dissipation, and an enclosure design with ample surface area for effective heat transfer to the environment by convection and radiation

Forced Air Example/Discussion

If the ECU requires forced convection cooling, there are limits on fan noise that must be observed (typically less than 30 dBA) and provisions made in the vehicle for some means of airflow ducting to avoid recirculation. Another forced convection approach is to employ the vehicle HVAC to provide both airflow and lower air temperature. This approach requires a significant integration effort.

Liquid Cooling Example/Discussion

Lastly, liquid cooling is required for high heat density and/or high-power devices and is usually applied for power transistors (hybrids, EV) and controllers for autonomous vehicles. This requires significant integration between the ECU and vehicle cooling loop.

To provide long term reliability in an automotive environment, the ADAS controller must be robust to withstand and function despite vehicle shock, vibration, humidity, temperature cycling, and other environmental constraints. As such, ADAS controller enclosures are designed to be ruggedized, sealed systems to minimize environmental exposure. The housings are typically designed to a specified ingress protection rating (IP-rating), such as IP54 which provides partial dust protection and protection against water splashes. To achieve this gasketing is used between housing halves and the I/O connectors are selected that have sealing capability (Fig. 21).

Internally, there is typically one printed circuit board (PCB) with all key components and connectors on one side. With this kind of PCB topography component thermal management is simpler: all "hot" components interfacing to one housing

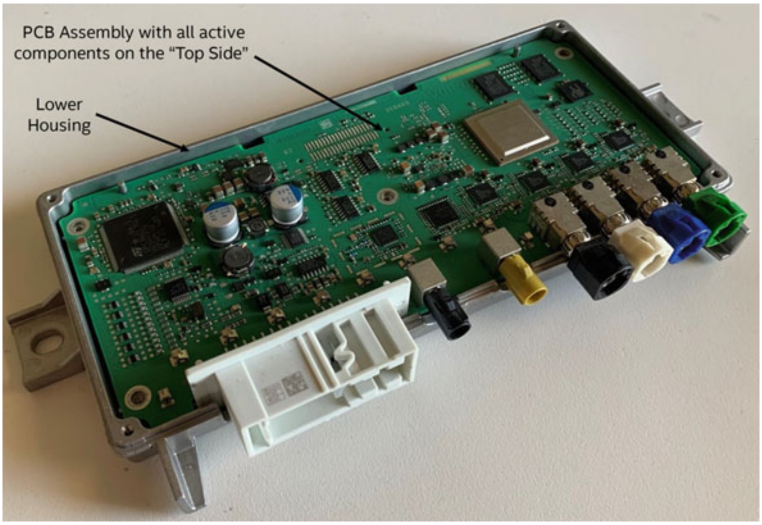


Fig. 21 Example of an automotive controller with components and connectors on PCB primary side

wall. Component sockets (processor and memory DIMM) and daughter-card connectors (M.2 and HDD) are typically not used to avoid potential loss of connectivity due to contact fretting over a lifetime of vibration exposure. If sockets and connectors are used, some means of staking is used to minimize relative motion at the socket/connector.

A thermal interface material (TIM) is required between the component and housing wall since there will be a necessary gap to account for component and assembly tolerances. As such, gap pad or gap filler materials are used to complete the conduction path. These thermal interface materials are selected for their thermal performance, material composition (e.g. silicone-free) and ease of assembly.

The mounting points of the PCB are defined to not only properly secure the PCB but also to ensure resonant frequencies and harmonics do not adversely affect component connections. Similarly, under shock events, the mounting points and to some extent the gap filler connections work to minimize damaging PCB accelerations (Fig. 22).

However, due to tolerances there must be a gap between the pedestal and component. To complete the conduction path from the component to the pedestal a thermal material is placed in the gap, known as a gap pad or gap filler. These materials have thermal conductivities in the 3–15 W/mK range, much less than the cast aluminum conductivity (~100 W/mK), so they present a significant thermal resistance. To minimize the resistance the gap is minimized as much as possible through a "gap" tolerance analysis given component, feature, and assembly tolerances (Fig. 23).

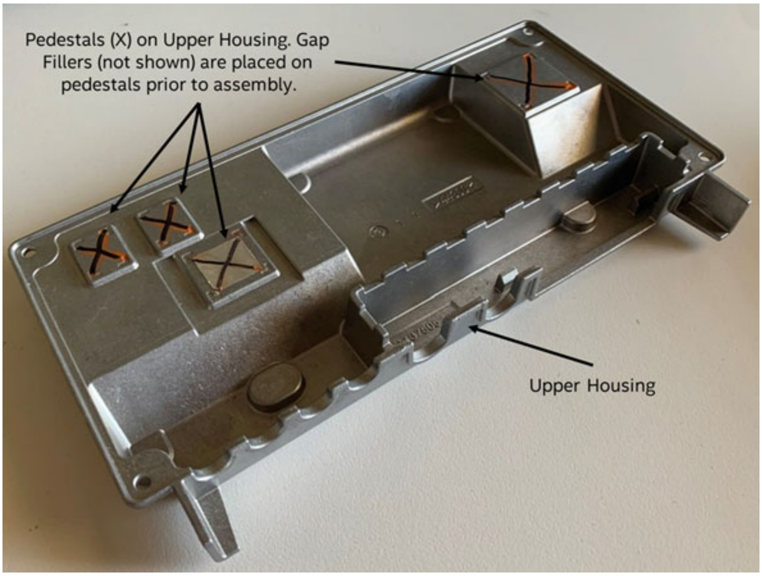
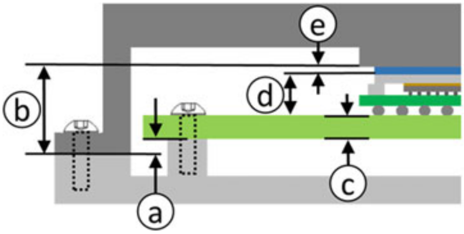

Pedestals (X) on Upper Housing. Gap Fillers (not shown) are placed on pedestals prior to assembly.
Upper Housing

Fig. 22 Cast aluminum housing with pedestal features-marked "X"



Symbol	Component	Tolerance [mm]
a	Die cast metal frame	0.13
b	Die cast metal heat sink	0.13
c	PCB	0.183
d	Package	0.167
e	Gap Filler (calculated)	0.31

Fig. 23 Example of a tolerance analysis to determine gap filler tolerance

To provide ample surface area for heat transfer to the environment, extended surfaces or fins are cast into the exterior of the housing. The fin geometry can be optimized for thermal performance, but also needs to follow the constraints imposed by the enclosure overall size limit, weight, and casting design rules.

When operating in a fanless mode, a significant part of the overall heat transfer from the controller to the environment occurs by radiation, approximately 30%, while the remaining heat transfer occurs by natural convection. The effectiveness of radiation heat transfer depends on the enclosure surface properties, particularly emissivity, which can vary between 0.1 and 0.2 for an "as cast" finish to 0.8 to 0.9 for a painted finish.

The enclosure mounting points can also provide a means of conduction heat transfer if the vehicle mounting surface is a sheet metal panel for example. However, this is not always the case where a molded plastic panel may be used, or the controller “snaps” into a plastic carrier where vehicle assembly time is minimized. Normally, controller thermal analysis assumes that there is no conduction heat transfer, just convection and radiation.

4 ADAS Product Reliability Requirements

4.1 Qualification Requirements

It is important to understand how the advanced driver assisted systems (ADAS) are leveraging the latest semiconductor technology prior to a discussion on ADAS quality and reliability requirements. As described previously, the four functional block concept makes ADAS concept possible: namely acquisition, perception, cognition, and action. Perception/Cognition block, to translate the sensed data to action is computationally intense and requires time sensitive (high performance and low latency) for safe and reliable operation. Automated systems require significantly higher computational power than the typical in-vehicle infotainment (IVI) systems. Continuous data integration from multiple sensors to assess environmental threats drives the need for continuous analytics and decision making. For decades, electronic components for cars have been qualified based on the AEC-Q100 standard [16], a summary of those requirements is shown in Table 2. Automotive Tier-1s have successfully relied on AECQ100 qualification approach to deliver products with almost perfect quality (generally <10 DPM at 0 km) and high reliability over the typical automotive life [17]. This was partly since lower performance products could rely on older, more mature silicon process technologies that have achieved low defect densities.

As illustrated earlier in Fig. 6, from a performance perspective, a typical 32-bit microcontroller with 100 s DMIPS capability may be sufficient to operate electromechanical systems of an automobile, an IVI chip however, requires 1000s of DMIPS to deliver intended function. To attain autonomy, an automobile will have computational needs in the 100–1000 k DMIPS range [18]. This computational power can be achieved in multiple ways; either use of higher performance semiconductors with high transistor count on the latest technology node, or from a system with distributed workload accommodated by multiple chips of lower computational capability. Either one of these approaches requires rigorous qualification method to ensure extremely high quality and reliability is achieved. The scenario is complicated further if the system must be compliant to the functional safety ISO 26262 standard necessary for ADAS type systems [19]. Functional safety compliance can be understood as predictable functional operation of the product preventing unacceptable harm to people. It defines how to manage random & systematic faults through specifically designed & certified HW & SW such that the automobile can be operated safely and predictably. For the purposes of this discussion, the authors will

Table 2 Summary of typical AECQ100 requirements for product qualification for Grade 1–3 (not an exhaustive list) along with comparison with PC Product

Stress name	Automotive requirement	PC product
Unbiased highly Accelerated stress test (UHAST)	Unbiased 110 °C/85%RH for 264 h	Unbiased 110 °C/85%RH for ~300 h with multiple reaouts
	77u/lot; 3 lots	77u/lot; 3 lots
Biased highly accelerated stress test (BHAST)	Biased 110 °C/85%RH 264 h	Biased 110C/85%RH ~300 h with multiple readouts
	77u/lot; 3 lots	77u/lot; 3 lots
Bake (HTSL)	+150 °C for 500 h	150 °C for upto 1000 h with multiple readouts
	45u; 1 lot	Sample size variable
Temp cycle	500 cycles of temp cycle B	Temp cycle B or other variations
	(−55 to +125 °C)	(−40 to 125 °C)
	77u/lot; 3 lots	Variable
HTOL	Vmax @ 110C; 1000 h	Stressed at higher voltage, with system stress, multiple readouts
	77u/lot; 3 lots	Variable
ELFR	Vmax @ 110; 48 h	Equivalent infant mortality Evaluation—stressed at elevated voltage & T
	800u/lot; 3 lots	Variable

assume that the functional safety needs are independent of reliability and outside this chapter's scope [20].

One of the challenges of using semiconductors that are qualified for general PC uses is the difference in quality expectations of those products. As an example, Fig. 24 shows the relative difference in quality and reliability expectations and the calendar life of a product qualified for a laptop relative to an IVI automotive product. Not only is the expected calendar life longer for an auto chip, the quality expectations are also ~10× more stringent. Furthermore, typical consumer electronic components are generally qualified for a shorter functional life span (3–4 years) with expectation of multiple software/firmware updates to manage product quality and security. Both elements are contrary to the automotive expectations where product life is 10+ years and software updates are infrequent and features generally fixed for the life of the automobile. Even the qualification requirements for typical PC products that have a functional life of less than 5 years are different than that of automotive products.

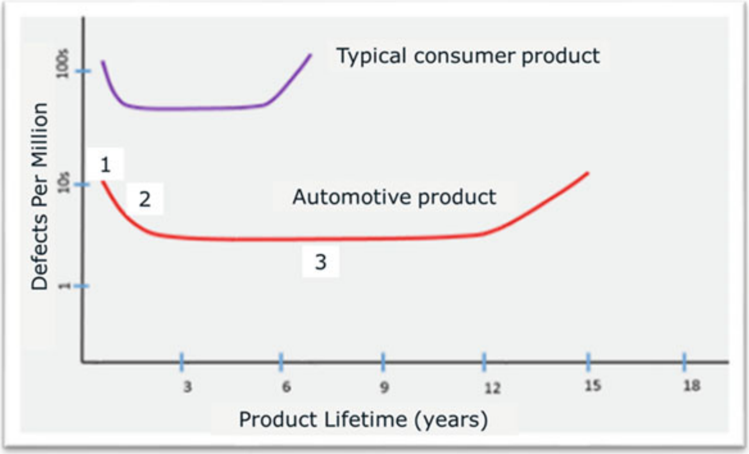


Fig. 24 Qualitative comparison of quality requirements for a computer chip used in a laptop compared to an automobile. Region 1 refers to quality (time = 0, region 2 refers to early life failure (infant mortality) and region 3 refers to reliability concepts

4.2 ADAS Performance Requirements and Implications—ADAS Mission Profile

One fundamental difference in ADAS, relative to traditional IVI systems, is the aggressive ADAS mission profile (usage model or use conditions), as shown in Table 3. Even within the ADAS family, the concepts of *transportation as a service (TaaS)* drive different set of requirements relative to a single owner, personal ADAS vehicle. Table 3 highlights operating hours for TaaS as ~3.5× more than the IVI segment. Even though the life expectation in TaaS may be 5 years, the CPU will be in continuous operation for those 5 years with ~85% residency in C0 state [21]. This usage model drives tremendous voltage, temperature, and thermo-mechanical stresses on the semiconductor product. Any successful qualification methodology will have to consider these stringent usage models to manufacture a reliable product.

Today, ADAS qualification requirements are not directly addressed by current AECQ standards which drives each Tier1 or OEM to define their own requirements for autonomous vehicles. Some customers have simply doubled the traditional AECQ100 requirements (HTOL and ELFR stress readout,[1] for example, increased by 2× of the IVI requirements) as a proxy for ADAS product qualification. This extended stress read out approach, without any due consideration for correlation to actual use models will likely provide inaccurate assessment of product reliability. For e.g., how should a single failure at 2000 h of HTOL readout be interpreted for infield failure rates? Such extended read outs serve as rough measures of wear out

[1] HTOL = High temperature operating life, ELFR = Early life failure rate.

Table 3 Basic usage model (use condition, mission profile) parameters to demonstrate differences in IVI, ADAS and transportation as a service (TaaS) segment.

Variable	IVI	ADAS	TaaS
Operating time (hrs./yr.)	1000	5000	8766
Power cycles per day	4	6–10	NA
Years of expected life (yr.)	12	5–15	5
CPU time in C0 state (%)	85	~100	~100
Non-operating time (hrs.)	93,210	Variable	NA
Time = 0 quality expectations (defects per million)	<50	<10	<10
Reliability expectations (% cumulative failure at end of life)	<0.5%	Variable	Variable

The parameter values, especially in ADAS use cases, are best guess estimates since these are fledgling concepts [22]

probability assessment and can disproportionately capture behavior of some mechanisms over others due to varying stress acceleration factors without providing any meaningful information on actual life of a product. Furthermore, if a product is based on the leading-edge silicon process that has not reached maturity or a predictable defect density, 77 units/lot for 3 lots, as stipulated in AECQ100 will not be sufficient to capture the entire process variation.

An alternative approach is to better understand the hardware and functional needs of the automated systems and utilize that knowledge to characterize known silicon and package failure mechanisms in terms of the variables (V, T, time, mechanical stress, and current) that accelerate those mechanisms. Once a technology is well characterized, modeling of these the mechanisms to suit different ADAS use conditions (mission profile) becomes relatively easy. This approach is described in the Jedec based Knowledge Based qualification standard where use conditions in terms of different variables (CPU use time, voltage, temperature, %RH, thermo-mechanical requirements, and vibrational) and operational life needs are considered. This allows better understanding of both the stimuli that stress the system, and the system response from a functional and reliability perspective, as described in the JEDEC Standard [23].

Finally, given that higher performance for these products is likely achieved using the latest technology (silicon and package), thorough characterization becomes necessary to ensure no latent failure modes lurk in the background. This requires investment in characterization techniques and a relatively large volume of data collection. For highest level of autonomous driving, building in functional redundancy and self-checking architecture to predict catastrophic failures and other such measures are necessary to achieve high reliability and safety.

4.3 Failure Regimes—Quality and Wear-Out Failures

A comprehensive product qualification approach should address the three distinct failure regions illustrated in Fig. 24. The semiconductors that fail during OEM system/vehicle tests get categorized as 0 km failures, generally referred to as quality failures. Products that fail during early vehicle operation (e.g., within a year) are categorized as *early life failures (also known as infant mortality).* Products that fail after some reasonable use due to wear out mechanisms are classified as reliability (or wear out) failures. For both regions 1 and 2 of the timescale shown in Fig. 24, proper product design, silicon process optimization and control and final product testing is critical to screening defective parts to deliver high quality products. Screening of defects is accomplished by backend testing of the final product, as illustrated in Fig. 25. Wafer level (sort) and product functional testing (ATE hot/cold and system) ensures that any circuit or process marginalities, while burn-in removes latent defects. These backend tests are becoming more crucial as the transistor and design complexities increase and the silicon process becomes more challenging with each new generation. Unfortunately, wear out failure (reliability and region 3) minimization cannot be accomplished by backend tests and requires correct by construction mentality and proper design engineering to ensure IP utilized, and the circuits constructed, are functional at the appropriate power, voltage and temperature corners for the duration of product life. Modeling of wear out mechanisms, based on empirical evidence, is critical to avoiding failures during the useful operating life of the product and is part of any thorough qualification process.

The underlying motivation for knowledge-based approach is to stress silicon, package, and interconnect combinations to better understand the intrinsic and extrinsic mechanisms that a product may succumb to in its life during operation. These reliability assessments are completed using experiments based on accelerated stresses such as temperature cycle (TC), power cycle, highly accelerated stress test (HAST), high temperature operating life (HTOL) and early life failure rate (ELFR) to tease out the potential failure mechanisms. Table 4 highlights the summary of the failure mechanisms prevalent in the traditional CMOS silicon-based semiconductors, and the variables that accelerate those mechanisms. Details of how KBQ approach can elucidate failure mechanisms can be found in published literature [24,

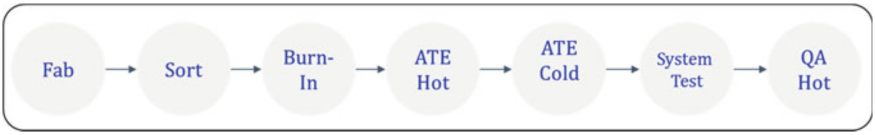


Fig. 25 Generic product testing flow to manage product quality, early life failure (ELF) and customer requirements/standards. Sort refers to testing at wafer level. Burn-in helps eliminate latent defects to prevent ELF. Testing from wafer sort to system tests helps screen out functionally defective parts as measured by structural, functional or system content

Table 4 Data showing typical silicon failure mechanisms and stimuli that accelerate the failures

User	Stimuli	Root cause	Failure modes
Const	Radiation	*n* & α *e-h* pair creation	Soft fail
Const	V, I	Latch-up and ESD	Electrical overstress
IM	N/A	Coverage of test	Logic fail
IM	V, T	External defects	Infant mortality fail
WO	V, T, RH	Electrochemical induced	Corrosion
WO	I, T	E-wind induced	Electromigration
WO	T	Voids, metal diffusion	Stress migration
WO	ΔT	Stress induced	Solder joint cracking
WO	ΔT	Stress induced	Interlayer cracking
WO	V, T	Gate dielectric leakage	Die break down
WO	V, T	Gate dielectric damage	Bias-temp instability
WO	V, I	Electron Ionization	Hot carrier
N/A	V	EOS caused by process	Process charging

25]. As Moore's Law drives miniaturization of transistors to a point where pitch-to-critical transistor-dimension ratio approaches unity, the likely hood of spatially sensitive marginalities increases. Given the latent nature of these marginalities there is an increased risk of field failures unless the "defects" are characterized properly and screened using burn-in or other methods.

As the power and current density increases in performance hungry ADAS applications, failure mechanisms such as biased temperature instability and electromigration can create risks of wear-out driven field failures if the technology is not characterized appropriately. A standards-based approach does not afford that detailed characterization of new mechanisms since the experiments are generally carried out at single V, T conditions for a predefined time, thus limiting the ability to predict behavior outside those chosen conditions. Another example of KBQ use to assess package interconnect failure risks to better understand the impact of temperature dwell time on lead-free solder joints can be found in an article authored by Xuejun Fan et al. [26]. The authors used various temperature shock experiments to better understand the impact of higher/lower temperature dwell time on the SnAgCu alloy-based solder. They demonstrated that higher dwell times at a given temperature led to lower solder joint life, demonstrating the value of detailed characterization of systems at various conditions to better predict life.

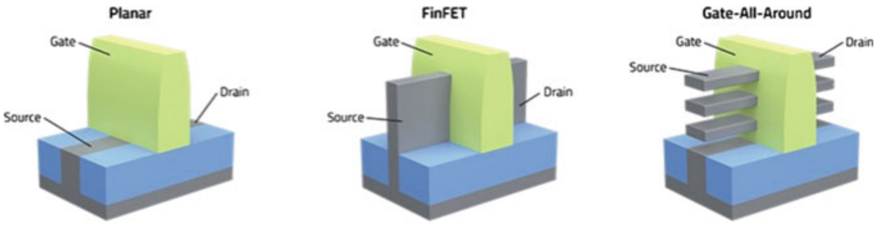


Fig. 26 Cartoon representation of transistor progressions from planar to FinFET to Gate-all-around. *Source* Https:/blog.lamrserach.com/finfets-give-way-to-gate-all-around

Extensive characterization of products is even more important when dealing with new technologies for new applications. The latest transistor evolution likely to be used in automotive products in ~5 years is called the *gate all around* transistor, as shown in Fig. 26. Autonomous systems will benefit from extensive failure mode characterization of these new transistor types to avoid any surprises of premature wear out mechanisms in the field during use.

IM infant mortality, *WO* wear out, Const constant, *V* voltage, *I* current, *T* temperature, *RH* relative humidity. See reference [12] for details.

4.4 Package Reliability Challenges

The key challenge facing the packaging community is to deliver highly functional packages, while ensuring near perfect quality and reliability. With Grade 0 requirements being most stringent, Fig. 27 presents the qualitative comparison of AEC-Q100 grade qualification requirements vs the current capability of packaging technologies. It ranges from traditional low I/O count QFN, QFPs', and FCBGA to the "state of the art" packaging technologies such as 2.5D/3D/SIP.[2] Figure 27 highlights the thermal challenges for high functional packaging technologies in AD applications. Significant R&D efforts and investments are required to optimize materials, design, and assembly/test processes for the delivery of higher functionality packaging technologies to reliability landing zone provided by AEC-Q standards.

Like silicon failure modes highlighted above, packages are also susceptible to a wide range of quality and wear out mechanisms. Figure 28 shows a few examples of failure modes that impact yield, quality (Time 0) or field performance. Most time zero package and assembly failures can be screened at the manufacturer during test. Reliability or wear-out failures during field operation need to be avoided by both proper design and characterization of package technology. Time 0 failures impact product yield, thus there is a financial incentive to eliminate these fails through process fixes. For instance, Fig. 28 under the "Yield title" displays "white bumps"

[2] QFP = Quad flat package, QFN = quad flat no-lead package, FCBGA = Flip chip BGA.

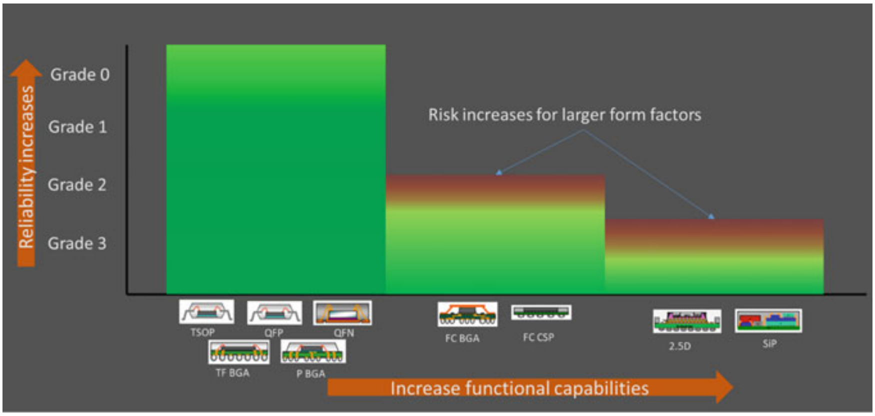


Fig. 27 Qualitative thermal capability reference with increasing package functionality with respect to automotive grade thermal requirements (grade 0 wider T range than grade 3)

(the contrast is from the scanning acoustic microscopy) which indicates delamination in silicon Low-K ILD layers. It is due to silicon-package interactions during the assembly processes. A comprehensive understanding of the failure modes and associated mechanisms allows for process optimization to eliminate the failure mode. In case of wear out mechanisms (shown as "reliability" title in Fig. 28) a detailed understanding of the mechanism and the reliability model to project the failure mechanism in the product lifetime needs to be developed by application of appropriate accelerated tests. Even cosmetic defects have the potential to be latent failures in the field. Figure 28 under the title "Quality" shows scratching on the silicon during assembly or testing. While they have no immediate functional impact on the product, customers reject units with these cosmetic defects because of latent failure possibilities. Developing appropriate metrologies for sub-par quality screening is crucial to meet product quality goals.

The ADAS concept in general is in a fledgling state and exploration in the next decade will better highlight the necessary workloads, features and customer expectations that need to be met for semiconductors to successfully enable safe and reliable vehicles. In absence of that maturity, a methodology that allows empirical characterization of failure modes is essential to tailor products to specific usage models or mission profiles since it allows higher confidence risk assessments. An equally important, but dynamic, component of ADAS platforms is software that is used to both integrate all aspects of autonomy, as well as unleash hardware functionality. Designers and system engineers realize that given the evolving nature of the use case and the peripherals in ADAS space, the system has to be flexible to be modified post-production, and at times *over the air* to manage functionality of the CPU.[3] This includes HW and SW security considerations to harden systems for vehicle safety.

[3] *Over the air* refers to the ability to provide software updates to the car (via cellular or WIFI) without a physical connection.

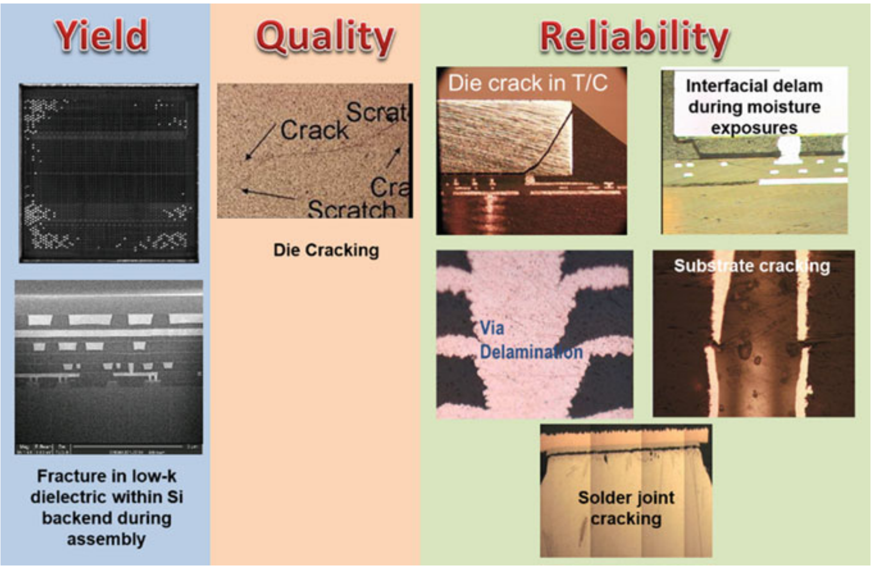


Fig. 28 Images of package failure mode examples categorized as yield, quality (time = 0) and reliability (wearout)

Regardless of the qualification processes used, SW updates will become frequent to manage safety, functionality, and security of the overall systems. This is yet another difference in how the IVI usage model has been implemented in the past, where SW is generally frozen once the systems are qualified.

In summary, given the overall ADAS complexity, and the lack of mature mission profiles that define truly autonomous systems, thorough characterization, and qualification of semiconductor components to understand their limitations (functional and reliability) is essential to design safe and reliable autonomous systems. Although standards-based systems can still provide value once technologies are mature, other methods such as knowledge-based qualification that allow empirical characterization of new technologies will prove to be valuable to deliver high quality and reliability automotive experience.

5 Summary

Automotive market is at the cusp where a traditional mechanical vehicle is becoming more like a *server on wheels* given the demands of a fully autonomous vehicle. The ADAS concept requires significant computational power to integrate the devices and features that are necessary to create a safe and reliable functional vehicle. Advanced packaging technologies such as EMIB, CoWoS, and Foveros bring significant integration advantages and will become mainstream in ADAS segment, however, they come with challenges. The chapter reviewed some basic concept of packaging technology, thermal challenges that will need to be managed, along with different approach to enabling highly reliable products for the stringent automotive needs. Furthermore, a more integrated approach will have to be considered to ensure component level challenges (e.g., thermal) are properly accounted for at the platform level to enable reliable, functional systems. This provides significant opportunities to the packaging community to develop innovative materials, processes, and designs to enable advanced technologies to meet the demands and requirements of the autonomous systems without compromising safety or quality.

Acknowledgements Authors would like to thank Dr. Ravi Mahajan for his guidance and feedback. The editors would like to thank Ms. E. Jan Vardaman from TechSearch International, Inc. for her dedicated review of this chapter.

References

1. https://www.pwc.com/gx/en/technology/publications/assets/pwc-semiconductor-survey-interactive.pdf
2. https://www.pwc.com/gx/en/industries/automotive/publications/eascy.html
3. Jacob W. Ward, et al., “The Impact of Uber and Lyft on Vehicle Ownership, fuel economy, and transit across U.S. Cities”, iScience 24, 101933, January 22, 2020
4. Deloitte. Autonomous driving moonshot project with quantum leap from hardware to software & ai focus (2019). https://www2.deloitte.com/content/dam/Deloitte/be/Documents/Deloitte_Autonomous-Driving.pdf. Accessed 30 January 2022
5. Computing System for Autonomous Driving: State-of-the-Art and Challenges, L. Liu, et al., Distributed, Parallel and Cluster Computing, Technical report: CAR-TR-2020-009, arXiv: 2009.14349
6. Computational requirements based on data gathered by Intel from OEM discussions (understanding of landscape circa 2018); Self-driving cars involve processing of high-resolution data from several sensors.
7. Rao R. Tummala, “Fundamentals of Device and Systems Packaging: Technologies and Applications”, second edition, McGraw Hill Publication.
8. https://eps.ieee.org/technology/heterogeneous-integration-roadmap/2021-edition.html
9. John Lau, Evolution, Challenge, and outlook of TSV, 3D IC Integration and 3D Silicon Integration, International Symposium of Advanced Packaging Materials (PAM), 2011
10. A general overview of reliability of single or multi-chip packages can be found in the *Heterogeneous Integration Roadmap*, Chapter 8, 2019 Ed., Shubhada H Sahasrabudhe et al., published by IEEE. [Http://eps.ieee.org/hir]
11. More discussion on next generation packages can be found in the *Heterogeneous Integration Roadmap*, Chapter 5, 2021 Ed., Sandeep Sane et al., published by IEEE. [Heterogeneous Integration Roadmap, 2021 Version (ieee.org)]
12. B. A. Myers, G. Eesley and D. Ihms, "Electronics Cooling in the Automotive Environment," Electronics Cooling, vol. 16, no. 1, pp. 16-21, 2010.
13. W. R. Johnson, J. L. Evans, P. Jacobsen and J. R. C. M. Thompson, “The Changing Automotive Environment: High-Temperature Electronics,” IEEE Transactions on Electronics Packaging Manufacturing, Vol. 27, No. 3, pp. 164-176, 2004.
14. M. Ohadi and Jianwei Qi, "Thermal management of harsh-environment electronics," Twentieth Annual IEEE Semiconductor Thermal Measurement and Management Symposium (IEEE Cat. No.04CH37545), 2004, pp. 231–240, doi: https://doi.org/10.1109/STHERM.2004.1291329.

15. N. Chang *et al*., "Emerging ADAS Thermal Reliability Needs and Solutions," in *IEEE Micro*, vol. 38, no. 1, pp. 66–81, January/February 2018, doi: https://doi.org/10.1109/MM.2018.112130058.
16. AEC-Q100 Standards is governed by the Automotive Electronics Council and can be obtained from http://aecouncil.com/AECDocuments.html.
17. Even though automotive life can be 12–15 yrs., the ICs within the automobile are operational for a fraction of that life. Typical operation expectation is 1000 hours ON time /year.
18. DMIPS stands for Dhrystone Million Instructions per second and is a relative measure of performance.
19. ISO 26262 FuSA related standard can be accessed via this ISO site: https://www.dnv.com/services/functional-safety-for-automotive-iso-26262--82719,
20. Functional Safety for automotive needs is governed by the ISO 26262 standard that can be accessed here; https://www.iso.org/standard/43464.html.
21. Given the absence of any standards, the time-in-state assessments, along with product life for TaaS segments are determined based on discussions with European automotive Tier 1 manufacturers.
22. These parameters are based on engineering discussions between Intel & European Tier 1 customers.
23. Knowledge based qualification methodology is described in the Jedec standards, *JESD94.01*.
24. Setting use conditions for reliability modeling, R. Kwasnick, P. Polasam and A. Lucero, *2014 IEEE International Reliability Physics Symposium*, 2014, pp. PR.5.1–PR.5.2.
25. Telemetry for System Reliability, R. Kwasnick, 2017, International Reliability Physics Symposium
26. Effects of Dwell Time and Ramp rate on Lead-Free Solder Joints in FCBGA packages", Xuejun Fan, George Raiser, Vasu S. Vasudevan, ECTC

Flash Memory and NAND

Zengtao Tony Liu

Abstract This chapter is focused on the fundamentals of Flash memory and NAND in particular. The explosion of data with the rise of the Internet and mobile computing in the past 30 years has made NAND Flash the most successful nonvolatile memory technology in the world. Compared to other alternatives, the low-cost-per-bit, high Read/Write speed, lightweight and compact form factor, and higher reliability over conventional hard disk drive (HDD) make NAND Flash the storage medium of choice for numerous applications. In the first section of this chapter, the basic floating gate memory cell structure is introduced to illustrate the fundamental physical characteristics that make NAND amenable for device scaling and well suited for data storage applications. Then, the historical evolution of NAND Flash is reviewed. In the second section, NAND fundamentals, including basic operations, memory architecture, and manufacturing processes, etc., will be discussed in detail. Starting around 2014, the NAND industry went through a major architecture transition from 2D to 3D NAND. The third section is devoted to the unique technology and design challenges of 3D NAND. It is followed by a discussion of various reliability issues, before the final conclusion on the 3D NAND future outlook.

1 NAND Flash—The Perfect Storage Medium

1.1 What is NAND Flash

NAND Flash is a kind of nonvolatile semiconductor memory based on the floating gate (FG) metal–oxide–semiconductor-field-effect-transistor (MOSFET) technologies. Here, the word "nonvolatile" refers to the ability of the memory to hold its data when the power is off. By contrast, another popular kind of memory used in virtually all computing systems, dynamic random access memory (DRAM), is volatile, because it requires active power to hold its data (the content will be lost within a few

Z. T. Liu (✉)
NAND-DTM, Intel Corporation, Dalian, Liaoning Province, China
e-mail: tony.zengtao.liu@intel.com

Y. Li and H. Shi (eds.), *Advanced Driver Assistance Systems and Autonomous Vehicles*,
https://doi.org/10.1007/978-981-19-5053-7_5

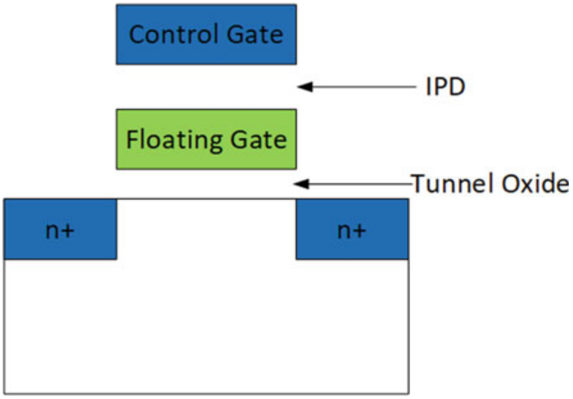


Fig. 1 Basic FG MOSFET memory element

minutes after the power is off). The basic memory element of NAND Flash is a FG MOSFET, which encodes the data by the amount of charge stored on the floating gate. “NAND” in the name refers to the way these memory elements are organized into arrays, while “Flash” refers to the fact that the stored data can only be erased in very large chunks at a time (“in a flash,” so to speak) [1].

Figure 1 shows the basic FG MOSFET memory element of NAND Flash. Compared to regular MOSFET, where the control gate (CG) electrode is directly above the conduction channel between the Source and Drain electrodes, FG MOSFET has an additional floating gate sandwiched in between. The floating gate is completely surrounded by insulating dielectrics, therefore is capable of holding electric charge for a long time. The dielectric between FG and the channel, often made of SiO_2, is commonly referred to as the tunnel oxide (more on that name later), while the dielectric between the FG and CG is called inter-poly-dielectric (IPD).

The FG MOSFET works very much in the same way as a regular MOSFET, except for the fact that now the CG modulates the channel potential through a capacitive network including the FG. In this capacitive network, the channel potential is a function of the CG bias as well as the amount of net charge on the FG. The more negative charges on the FG, the harder it is for the CG to raise the channel potential and turn on the transistor. In another word, the higher is the threshold voltage (V_T).

This relationship between the charge state of FG and transistor V_T provides a very effective way for nonvolatile data storage. Effectively the “0” state is defined by the negatively charged FG and associated with high V_T, and the “1” state is defined by neutral or positively charged FG and associated with low V_T. The Program (or Write) operation involves injecting electrons into the FG and putting it into the negatively charged “0” state, while the Erase operation involves removing electrons from the FG and putting it into the neutral or positively charged “1” state. And, the read operation is simply the determination of the transistor V_T to detect its corresponding data state. That, in a nutshell, is the working principle of Flash memories based on FG MOSFET.

1.2 NOR Versus NAND

To build a working Flash memory chip, many FG MOSFET memory elements have to be organized into arrays in a cost-effective way. There are two fundamental requirements on the memory organization.

1. Each memory element has to be independently addressable for Program and Read operations. Erase operation does not have to be done for each element individually, which is the hallmark of Flash.
2. The layout of the memory array needs to be as compact as possible because it is a major factor determining the bit cost of the memory chip.

Two types of Flash memory organizations enjoyed widespread commercial successes throughout the evolution of Flash memories, NOR and NAND. The basic array construction of NOR Flash is illustrated in Fig. 2. In NOR Flash, the memory elements are connected in parallel, with each FG MOSFET connecting to the source line (SL) and bit line (BL) directly. As a result, in the layout shown in Fig. 2b, a BL contact has to be included in the memory bit cell design, limiting the word line (WL) pitch. In this configuration, the memory elements connected to the same BL perform the “NOR” logic operation—the BL output is pulled low when any of the WL input is high.

The advantage of NOR Flash is its fast random access time and bit/byte alterability, which make it ideal for code storage. The programming operation of NOR Flash is usually achieved by biasing both the selected WL and BL high to pass a large current from the BL to the SL. The large current combined with the large Drain-to-Source bias (V_{DS}) generates a lot of “hot electrons”—electrons with enough energy to overcome the energy barrier at the oxide-silicon interface. These channel hot electrons are trapped in the FG and cause the cell to enter into the high V_T “0” state. The Erase operation is performed by applying a positive bias on the substrate and negative bias on the WL to induce a strong electric field in the tunnel oxide, which causes

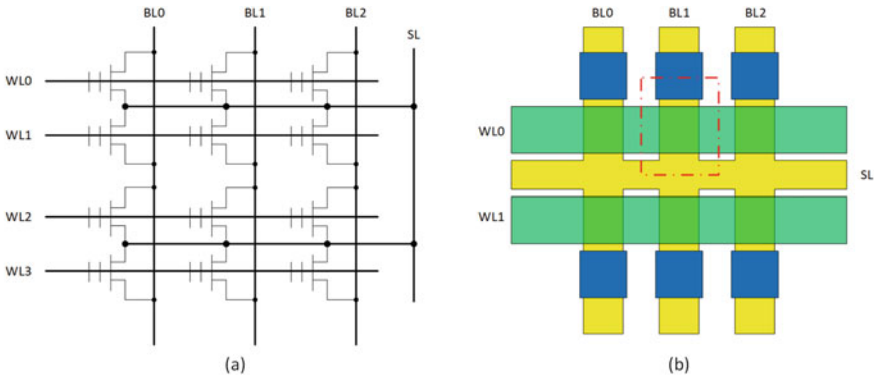


Fig. 2 NOR flash **a** schematic, **b** layout. The red rectangle in **b** indicates the unit cell

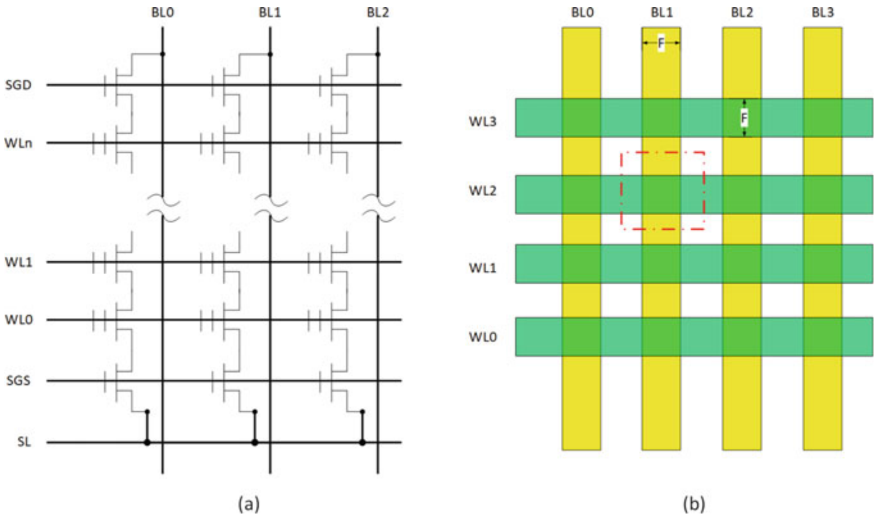


Fig. 3 NAND flash **a** schematic, **b** layout. The red rectangle in **b** indicates the unit cell

the electrons in the FG to escape through a process called Fowler–Nordheim (FN) tunneling.

NAND Flash, on the other hand, organizes FG MOSFETs into NAND strings, as shown in Fig. 3. In each NAND string, there are two Select Gate (SG) transistors, one on the Source end (SGS) and the other on the Drain end (SGD). Between SGS and SGD, many FG MOSFETs are connected in series. The serial connection of the NAND cells forms a logical NAND operation—the BL output is pulled low only when all the WL inputs are high.

The most obvious advantage of NAND Flash is its compact layout, as shown in Fig. 3b. The BLs/AA and WLs are all simple straight-line patterns orthogonal to each other, forming the NAND cell at their intersections and resulting in a cell size of $4F^2$, where F is the minimum feature size of the technology. The serial connection of NAND cells makes it difficult to access individual NAND cells at a time. So in NAND Flash, data are organized into pages—cells connected to the same WL, and Read/Program operations are performed on a per-page basis. Even though the operations tend to be slower compared to NOR Flash, because one page consists of many bits, the parallelism results in a very high data throughput, which makes NAND Flash more suitable for data storage.

1.3 Evolution of NAND Flash

NAND Flash memory was invented by Fujio Masuoka while working at Toshiba Corporation, Japan, in 1987 [2], and Toshiba brought NAND Flash into commercial production in 1992 on a 0.7 um technology [3]. NAND Flash did not receive wide adoption right away. Even though the NAND Flash cell is much smaller than the NOR Flash, the benefit did not translate into a smaller die size because of the more complicated peripheral circuits required to support the architecture. Another advantage of NOR over NAND Flash is its fast random reads, since each memory cell is directly accessible owing to its parallel organization.

All of that starts to change in the late 1990s. Two underlying forces gradually drive NAND Flash to become the more dominant nonvolatile memory technology over its NOR cousin and to the forefront of consumer electronics. The first is technical and common to most of the semiconductor technologies. Gordon Moore, a co-founder of Intel Corp, observes in a 1965 paper in Electronics Magazine that the number of transistors on an IC chip has been doubling every 1.5 ~ 2 yrs. This phenomenon, later dubbed "Moore's law," has stayed more or less true for over half a century. Though not a fundamental "law," it highlights the potential of semiconductor technologies driven by the collective innovations of countless researchers and engineers working in the field. The reason why more and more transistors can be packed into a single chip is that each individual transistor can be made smaller and smaller through a process called scaling. By scaling the critical dimensions of the transistors, more of them can be crammed into the same footprint while maintaining the basic functionality of each individual transistor, sometimes improving their performance or power efficiency in the process. NAND Flash is perfectly suited for this kind of scaling because of its simple array configuration. As the memory density (the number of bits that can be stored on a single die) increases, the overhead of complicated peripheral circuits matters less and less, so NAND Flash gradually becomes more efficient in die size and costs less to produce than NOR Flash.

The second force is the rise of mobile computing, first in the form of consumer electronics, then through the rise of smart phones and mobile Internet. Though not good for fast random read, the large page size of NAND Flash makes it an ideal data storage medium. Therefore, since its inception, NAND has been envisioned as a potential replacement for magnetic storage technologies (think tapes and hard drives). By the late 1990s, NAND Flash is still many times more expensive than the tapes or hard drives, but its cost has dropped to a point to justify adoption for mobile applications due to its lightweight, compact size, and superior reliability. The rise of digital photography in this time period fueled the growth of NAND Flash market. Soon after, NAND is also used in digital music players, voice recorders, USB sticks, and many other portable consumer products.

The real explosion of the NAND Flash follows the introduction of the modern smart phones. It is not just the hardware itself, which uses NAND Flash for data storage. It is the fundamental shift in how people interact with the Internet through their mobile devices—the rise of social networks, the ubiquity of e-commerce, and the ease of producing and consuming content with the smart phones (and tablets), all of which generating ever growing amount of data and demanding ever higher storage capacities. Today, NAND Flash is not only used in the smart phones but also made into Solid-State Drives (SSDs) that go into both personal computers and powerful servers that power the Internet and all kinds of cloud applications.

Along with the NAND Flash market explosion, to meet the ever-increasing need for higher memory storage capacity, many technological innovations are developed. Two of them are especially worth mentioning. The first one is the multi-bit-per-cell technology. At the inception, each Flash memory cell can only store 1 bit, which is encoded by two different V_T states. Later on it is only natural for people to device schemes to store more bits per cell by differentiating more than two different V_T states. To encode 2 bits, 4 different V_T states are needed; to encode 3 bits, 8 different V_T states are needed; so on and so forth (1-, 2-, 3-, and 4-bit/cell technologies are commonly referred to as SLC, MLC, TLC, and QLC technologies, which will be explained in later sections). With multi-bit-per-cell, higher storage capacity can be achieved without actually adding more memory cells, effectively lowering the per-bit cost. However, more complicated circuits and operations are needed to differentiate more V_T states. And there is significant performance degradation associated with it as well. We will cover the subject in more details in the next section. As of today (2021), both 3-bit and 4-bit-per-cell technologies are commonplace in commercial products.

The second major technological innovation is the transition from 2D to 3D NAND. The basic Program/Erase operations of NAND Flash involve quite high voltages and strong electric field, which prevented the scaling down of certain cell geometries or packing them too close to each other. Once below the 20 nm node, scaling the 2D NAND to smaller dimensions becomes increasingly difficult. Another fundamental limit of 2D NAND comes from the capacitive coupling nature of NAND device operations. With smaller devices packed closer to each other, on the one hand, the individual capacitance of each FG node becomes smaller. On the other hand, the capacitance between adjacent memory cells becomes larger relative to the total capacitance. The smaller FG capacitance means it takes less charge to induce the same amount of V_T shift. At <20 nm node, different V_T states are separated by just a few dozen electrons on the FG. With such few electrons, random variation (which scales with $1/\sqrt{n}$) becomes a huge issue and is hard to control. At the same time, the larger cell–cell coupling means that the V_T sate of one cell is more easily influenced by the charges on the FGs of the adjacent memory cells (or their V_T states), leading to greater interference.

3D NAND offers very effective solutions to the scaling issues of 2D NAND. In 3D NAND, the NAND string is rotated 90° and no longer resides inside the Si substrate. It now consists of vertical transistors built on top of the Si substrate with a poly-Si channel. There are two main families of 3D NAND technologies on the market, Floating Gate (FG) and Replacement Gate (RG). In FG NAND, the charge is stored in discrete floating gates, just like in 2D NAND. In RG NAND, however, the charge is stored in a dielectric film continuous throughout the NAND string. The film (usually SiN) is engineered to have a lot of electron traps for effective charge storage. Compared with sub-20 nm 2D NAND cells, both FG and RG 3D NAND cells are much bigger and have less cell–cell coupling. Therefore, they demonstrate superior electrical performance. The unique 3D architecture also allows many cells to be formed at the same time without adding too many process steps, therefore lowering the cost-per-bit. Figure 4 shows the scaling trend of NAND technologies as

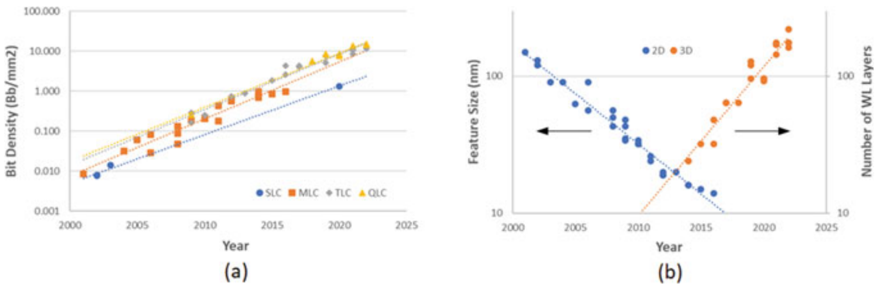


Fig. 4 NAND flash scaling trend **a** gross bit storage density (GBSD) scaling, **b** 2D NAND feature size and 3D NAND tier count scaling

reported in the International Solid-State Circuit Conference (ISSCC) by the major NAND Flash manufacturers from 2001 to 2022 [4–30, 42–47]. It could be seen that the industry has kept the pace of Moore's law for almost 20 years. And, both the geometric scaling (in terms of smaller feature size in 2D NAND and increasing number of layers in 3D NAND, respectively) and the migration toward higher number of bits per cell play important roles in the evolution of the storage density.

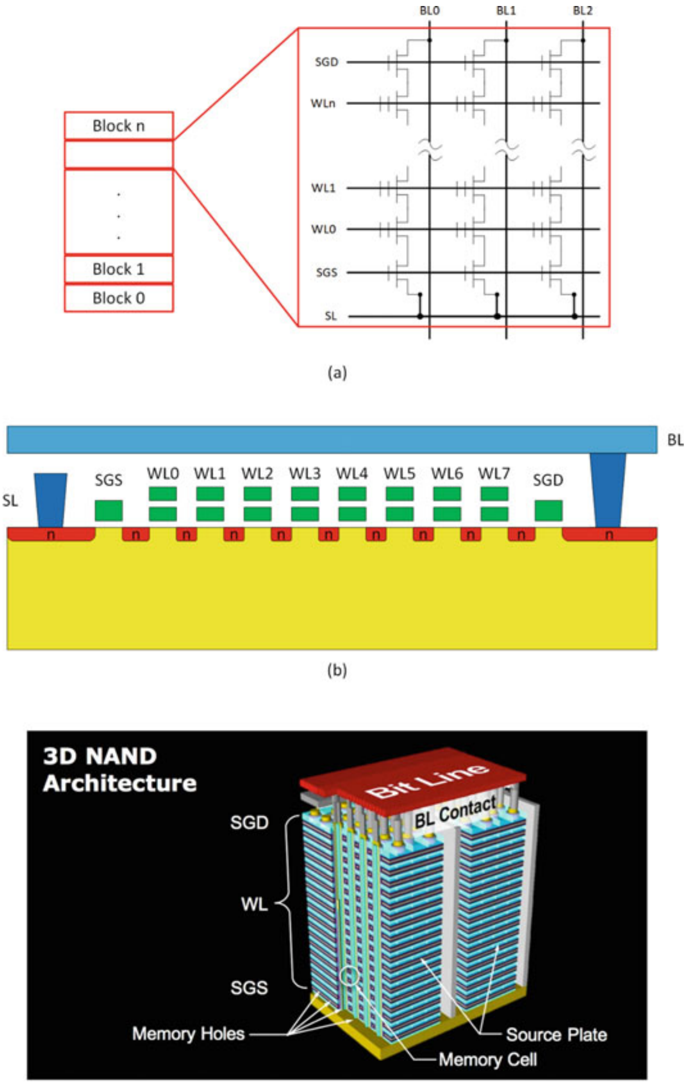
2 NAND Fundamentals

2.1 NAND Arrays in 2D and 3D

As mentioned in Sect. 1, NAND Flash memory cells are organized into NAND strings, with each string consisting of an SGD, an SGS, and many NAND cells connected in series. One end of the strings is connected to the BLs, while the other end of the strings is connected to a common node called the source line (SL). The SGD/SGS/WLs are running orthogonal to the BLs, and all the NAND strings sharing the same set of SGD/SGS/WLs form a block. Many blocks are placed next to each other, sharing the same set of BLs, and form a continuous NAND array. The schematic is illustrated in Fig. 5a, while Fig. 5b shows a cross section of its 2D implementation, and Fig. 5c shows a 3D implementation.

In the NAND arrays, a block is the smallest unit for the Erase operation, while the Read/Program operations are performed on a per-page basis. A page refers to all the cells on a particular WL within a block.[1] In modern NAND arrays, a common page size is 16 KB, which means 16 KB info is written into or read from the NAND

[1] This is true for the ABL (All-BL) architecture, which is adopted by virtually all 3D NAND technologies. In 2D NAND, there is an alternative SBL (Shielded-BL) architecture, where the cells on a particular WL are organized into 2 pages. The cells on the Even BLs belong to 1 page, while the cells on the Odd BLs belong to another page.



(a)

(b)

(c)

Fig. 5 NAND array **a** schematic, **b** 2D implementation, **c** 3D implementation

array after a single Program/Read operation. It is this high level of parallelism that makes NAND uniquely suited for nonvolatile data storage.

2.2 *Basic NAND Operations*

2.2.1 Read

All the Read/Write operations of NAND Flash are performed on many cells in parallel (for the whole page or even block). Because of the random variations of the manufacturing process and the stochastic nature of the electron tunneling across a dielectric layer, those cells are not going to show identical electrical characteristics. This will manifest as a difference in their cell V_T. So if we plot the number of cells against their V_T value, we are going to get a distribution for both the programmed "0" state and the erased "1" state, as shown in Fig. 6a. To be able to distinguish the "1" bits from the "0" bits, the two distributions have to have a large enough separation to allow us to place a read voltage (V_R) in between.

Figure 6b shows the typical bias condition of the NAND Read operation. During Read, all the BLs are biased with a small voltage ~0.5 V, while the SL is biased to GND. The SGS/SGD of all the unselected blocks is also biased to GND to disconnect them from the BLs and the SL. For the selected block, the SGS/SGD is biased to ~5 V to turn them on. As mentioned earlier, the Read operation is done on a per-page basis. So, among all the WLs, only one WL is selected and applied a V_R, while the rest of the WLs are biased with a V_{Pass_R}. The choice of V_R and V_{Pass_R} voltages needs to ensure that V_R is between the "1" and "0" V_T distributions, while V_{Pass_R} is higher than the highest V_T of the "0" distribution. This way all the NAND cells along the unselected WLs will be turned on regardless of their V_T states. For any cell on the selected WL, if its V_T is higher than V_R, the cell will be off and no current can flow between the BL and the SL; if its V_T is lower than V_R, the cell will be on, therefore turning on the whole NAND string to allow current to flow from the BL

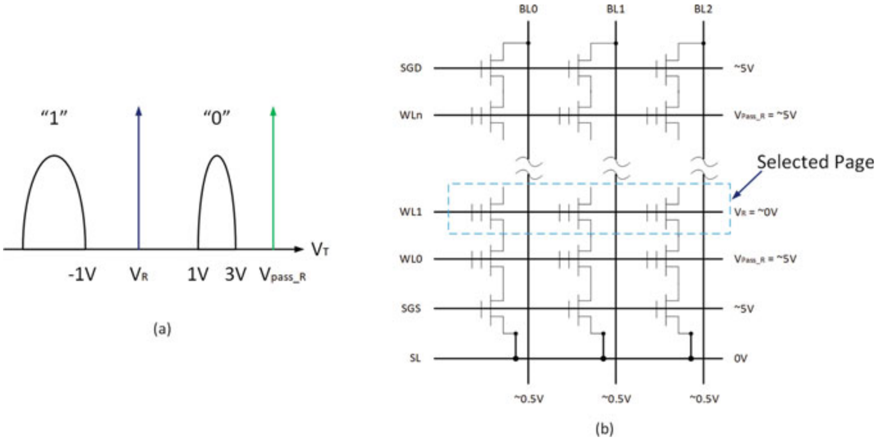


Fig. 6 NAND Read operation, **a** typical V_T distributions, **b** typical bias condition of the selected block during read

to SL. All the BLs are connected to sensing circuits that can detect if a current is flowing and store the results as a “1” (with current) or a “0” (no current) to output later.

2.2.2 Program

There are two types of Write operations in NAND Flash. The Program operation turns selected bits on a page from the “1” state to the “0” state, while the Erase operation turns the whole block into the “1” state. Figure 7a shows the typical bias condition of the Program operation. During the Program operation, for the cells to be programmed (inside the blue circles), a very high voltage (~20 V or higher) is applied on their gates (selected WL), while their channels are grounded through the BLs. This creates a very strong electric field in the tunnel oxide of those cells, which causes electrons to be injected from the channel into the FG through FN tunneling, increasing their V_T.

As shown in Fig. 7a, all the cells along the selected WL get the high V_{Pgm} applied on their gates. However, not all of them need to be programmed into the “0” state. Then what happens to the cells that are supposed to remain in the “1” state? How to prevent them from being programmed? This is achieved through a process called Inhibit. As shown in Fig. 7a, during programming, both SGS and SL are biased at GND, meaning SGS transistors are in the off state and does not participate in the Program operation. While the selected WLs are biased to V_{Pgm} (~20 V), all the unselected WLs are biased to an intermediate V_{Inh} (~10 V). The SGD is biased to V_{CC} (~2.5 V). The voltages of the BLs are applied based on the data to be written—if a bit is “0”, the corresponding BL is biased to GND, the cell gets programmed; if a bit is “1”, the corresponding BL is biased to VCC (~2.5 V), and the cell gets inhibited.

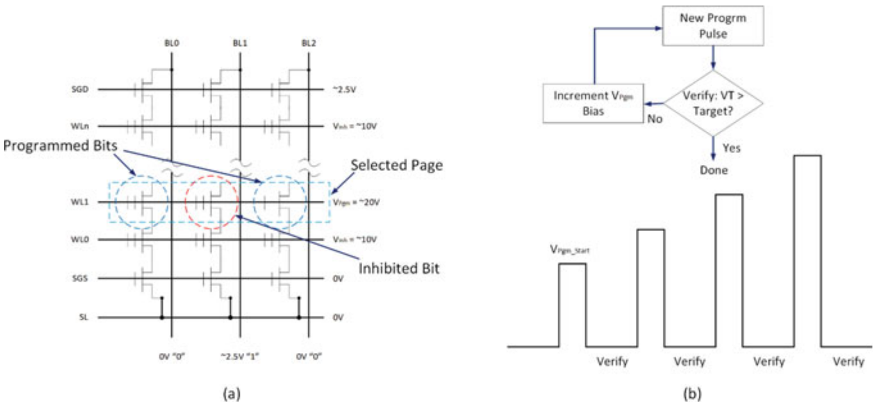


Fig. 7 NAND program operation **a** typical bias condition of the selected block during programming, **b** basic program algorithm and the selected WL pulses

To understand how Program Inhibit works, it is important to note that the BL and SGD biases are set BEFORE the Program pulse is given. This is to set the proper "seed" voltage inside the NAND string channels. For the BLs biased at GND, the V_{CC} on the SGD is high enough to pass the GND into the NAND string channels and leave the SGD transistor fully on, so the channels stay at GND throughout the Program pulse, keeping a large voltage difference between the selected WL and the channels (V_{Pgm}) to generate FN tunneling. For all the cells under the unselected WLs, they see an intermediate bias (V_{Inh}) that is not high enough to cause them to be programmed.

For the BLs biased at V_{CC} (~2.5 V), the same V_{CC} on the SGD will pass a voltage of V_{CC}-V_{T_SGD} into the NAND string channels and leave the SGD transistor only weakly on. When the V_{Pgm}/V_{Inh} pulses are applied on the selected/unselected WLs, respectively, because the SGD transistor is only weakly on, the NAND string channels will be capacitively coupled up to a higher voltage. As soon as they are coupled up, the SGD transistors are effectively turned off, allowing them to basically follow the V_{Inh}. Then, for the inhibited cells on the selected WL (red circle in Fig. 7a), the voltage difference between the WL and channels is ~V_{Pgm}-V_{Inh}, not large enough to cause them to be programmed.

To control the Programmed V_T distribution, the Program operation is usually done through successively increasing V_{Pgm} pulses according to the algorithm shown in Fig. 7b. After each V_{Pgm} pulse, a Program Verify (essentially a Read operation) is performed. If a bit passes Verify, meaning the cell V_T exceeds the target, it will be inhibited for the subsequent V_{Pgm} pulses. For each V_{Pgm} pulse following a Verify, V_{Pgm} is incremented by a fixed value to make it easier to program the rest of the "hard-to-program" bits. The process repeats until all the bits pass Verify.

2.2.3 Erase

Erase operation is essentially the opposite of the Program, where a high-voltage bias is applied between the WLs and the channel to generate a high enough electric field in the tunnel oxide and cause FN tunneling, except that the bias polarity is reversed and the operation is done for the entire selected block.

Figure 8 shows the typical Erase bias condition for both 2D and 3D NAND. For 2D NAND, a high voltage is applied on the P-well (PW) of the NAND array. Both the BLs and SL are left floating, which will be charged up by the PW through forward-biased junctions. For the unselected blocks, the WLs/SGD/SGS are all left floating as well, and they get coupled up to around the PW bias through capacitive coupling. So, no high electric field is generated in the unselected blocks. They simply retain the data they had prior to the Erase operation. For the Selected Block, however, all the WLs are biased to GND, generating a high electric field in the tunnel oxide of all the NAND cells inside, causing the electrons in the FG to tunnel into the PW, and reducing their V_T in turn. The full Erase operation is conducted in a manner similar to Program, with the Verify operation interlaced between a series of Erase pulses with increasing magnitude. The operation stops when all the bits pass Verify.

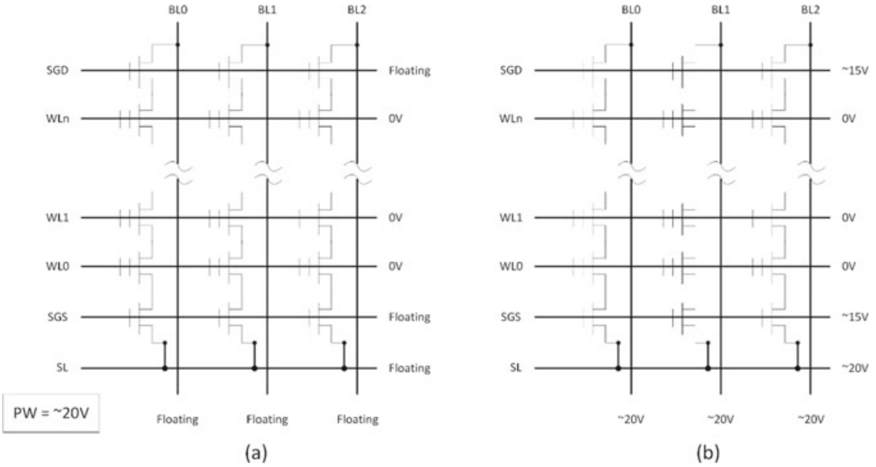


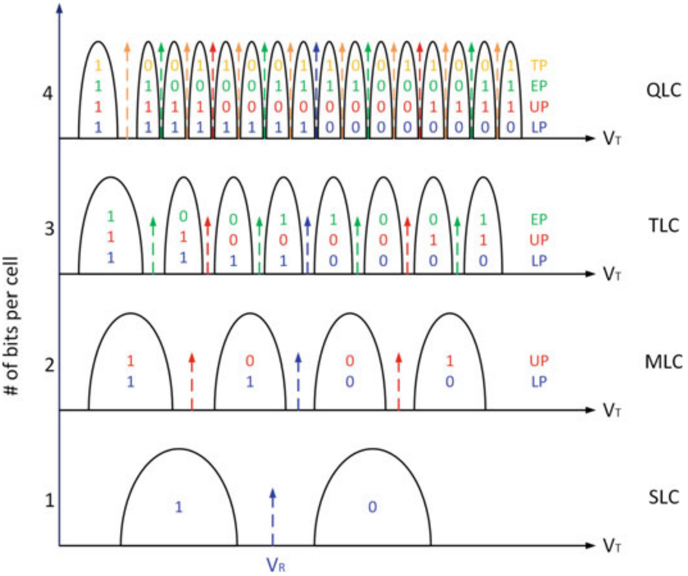
Fig. 8 Typical bias condition of a selected block during erase of **a** 2D NAND, and **b** 3D NAND

For 3D NAND, because the NAND channel is formed inside each individually etched high aspect-ratio (AR) hole, it does not have a direct access to any "P-well" node. So, for the Erase operation, a high bias of the NAND channel is supplied through the BLs and/or SL by the SGD/SGS Gate-Induced Drain Leakage (GIDL) current. As shown in Fig. 8b, the SGD/SGS of the selected block is biased ~5 V lower than the BLs/SL. This negative bias creates a high electric field inside the SGD/SGS transistors near the Drain edge, which causes a significant leakage current large enough to charge up the channel of the NAND strings, despite that SGD/SGS is technically in the "off" state. For the unselected blocks (not shown in Fig. 8b), all the WLs/SGD/SGS are floating and they get coupled up capacitively to near the BLs/SL voltage, similar to the 2D NAND case.

2.3 Multi-Bit-Per-Cell Technologies

Figure 6a in the previous section shows the simplest representation of two unique V_T states representing logic "1" and "0", respectively, separated by the Read voltage V_R. With two V_T states, each cell stores 1 bit of information. This is called Single-Level-Cell (SLC) within the industry, which is an obvious misnomer since two V_T levels are involved here. So, the "single" in the name more accurately describes the number of bits instead of the number of V_T levels. With the relentless drive to reduce the bit cost of NAND Flash, methods of storing more bits per cell by introducing more V_T states have been developed as an extra vector in addition to scaling down the cell dimensions.

Figure 9 compares the V_T distributions of SLC (1-bit, 2 V_T sates), MLC (Multi-Level-Cell, 2-bit, 4 V_T states), TLC (Triple-Level-Cell, 3-bit, 8 V_T sates), and QLC



# of bits per cell
4
1 0 0 1 1 0 0 1 1 0 0 1 1 0 0 1 TP
1 1 0 0 0 0 1 1 1 1 0 0 0 0 1 1 EP
1 1 1 1 0 0 0 0 0 0 0 0 1 1 1 1 UP
1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 LP
QLC
V_T
3
1 0 0 1 1 0 0 1 EP
1 1 0 0 0 0 1 1 UP
1 1 1 1 0 0 0 0 LP
TLC
V_T
2
1 0 0 1 UP
1 1 0 0 LP
MLC
V_T
1
1
0
SLC
V_T
V_R

Fig. 9 V_T distributions of SLC, MLC, TLC, and QLC NAND

(Quad-Level-Cell, 4-bit, 16 V_T states technologies. For the multi-bit-per-cell technologies, the extra bits encoded by the additional V_T levels are organized into separate pages. So, for MLC, the NAND cells making up one physical page store two logical pages, one is called the Lower Page (LP) and the other the Upper Page (UP). For TLC, an Extra Page (EP) is introduced on top of LP and UP. For QLC, another page, Top Page (TP) is added.

A few observations can be made from the V_T distributions in Fig. 9. First, to accommodate the exponentially increasing number of V_T states required to store the extra bits per cell, two things are necessary.

1. The overall V_T window (loosely defined as the difference between the highest Programmed V_T and the highest Erased V_T) has to be enlarged to accommodate the increasing number of V_T states.
2. The width of each programmed V_T state has to be reduced to allow sufficient separation between them.

To achieve them, more stringent control of the manufacturing process and more advanced Program algorithms are usually required. The former can lead to additional wafer cost, offsetting the cost benefit of multi-bit-per-cell technologies somewhat, while the latter leads to longer Program time, reducing performance.

The arrows in Fig. 9 indicate the Read voltages (V_R) needed for each logic page. It can be seen that while only 1 V_R is needed for the LP, 2 are required for UP, 4 for EP, and 8 for TP. It does not necessarily mean that eight separate Read operations are needed to read out the TP data for QLC, but it is true that with more bits per cell, generally more reads are needed to decode the stored data on average. In addition to the number of reads, packing so many V_T states within the limited V_T window means that their separation becomes smaller by necessity. So the read margin (difference between V_R and the edge of adjacent V_T states) becomes smaller as well. This leads to more read errors, which require either advanced read algorithms or more powerful Error-Correct-Code (ECC) to correct. All of those lead to lower Read performance.

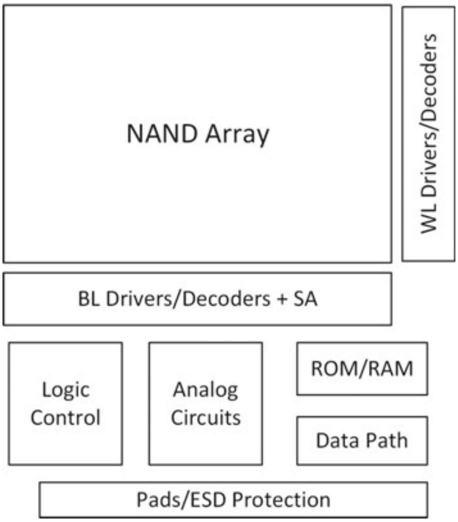
While the cost benefit of multi-bit-per-cell technologies is quite clear, it does carry some penalties in performance. Up to now, the cost benefit still wins out and the industry is even pushing for PLC (Penta-Level-Cell, 5-bit, 32 V_T states) technology. It is most likely impractical to go beyond PLC due to the increased technical challenges, diminishing return on cost, and the excessive performance degradation.

2.4 Anatomy of a NAND Product

So far, the discussion has been focused on the NAND array and its operations. To build a functional NAND product, it is necessary to have enough supporting circuitry as well, so all the necessary array operations can be carried out and data communicated with the outside world. To realize those supporting circuitry, additional CMOS devices are needed other than the NAND memory elements. Figure 10 illustrates the basic components of a typical NAND die, which can also be generalized to other nonvolatile memories as well.

NAND array is of course the most important part of the NAND die, being what the customers actually pay for. To access the memory elements inside the NAND array, both WL and BL drivers/decoders are needed. WL drivers/decoders are needed to select the active block/page and set the correct voltages on all the WLs, while the BL

Fig. 10 Building blocks of a NAND die



drivers/decoders set the BL voltages for Program/Erase and connect the BLs to Sense Amplifiers (SA) during Read. In modern NAND products, there are multiple latches built into the BL drivers/decoders as well to handle the logic operations needed to support the multiple Read operations associated with multi-bit-per-cell technologies. The whole BL circuits are commonly referred to as the Static Page Buffers (SPB).

Other than the circuits directly interfacing with the NAND array, there need to be additional digital circuits on the die to process customer commands, translate them into the correct sequences of memory operations, and facilitate the completion of those operations. Some analog circuits are needed to generate the various internal voltages/currents in support of all the memory operations. Then, there is ROM/RAM which is needed to store pre-programmed codes and specialized data path circuits to handle the high-speed data communication between the SPB and the Data Queue (DQ) pads. All the NAND pads are usually located on one side of the die to make it easier to stack multiple dies into a single package, increasing the memory density of the final product.

To realize all the functional blocks mentioned above, it usually requires multiple flavors of CMOS devices. The data path needs high-speed CMOS devices to support >1Gbps I/O speed at relatively low voltage, while the WL drivers/decoders and charge pumps need high-voltage MOSFETs that can support >20 V operations. At the end of the day, all these CMOS devices have to be integrated together cheaply to keep the overall wafer cost low.

2.5 3D NAND Technology Basics

2.5.1 FG Versus RG NAND

There are two main types of 3D NAND technologies in today's high-volume commercial products—FG NAND developed jointly by Micron Technology and Intel, and RG NAND first developed by Samsung and later adopted by the other major NAND Flash producers as well. Both FG and RG NAND follow the basic memory architecture illustrated in Fig. 5c, having NAND strings formed inside the high aspect-ratio holes etched into stacks of oxide/WL films. A continuous film of polysilicon forms the conductive channel of the NAND strings. The key difference between the two technologies is in the construction of the basic 3D NAND cells. In FG NAND, the WLs are made of polysilicon and each NAND cell has a discrete FG between the WLs (CG) and the conductive channel. In RG NAND, the WLs are made of W and, instead of the discrete FGs, a continuous nitride film along the NAND string is sandwiched between two oxide layers between the WLs and the conductive channel. Charges (electrons/holes) can be trapped inside the nitride layer, modulating the cell V_T in the same way as the FG.

Figure 11 (adapted from [31]) illustrates the basic cell formation process of the FG NAND. First, an alternating oxide/poly film stack is deposited, before individual 3D memory holes are patterned and etched into the film stack. Then, a CG recess is performed through wet etch of the poly WLs. IPD films are formed on the inside of the recessed WLs, followed by the FG poly-deposition. Another poly-wet etch is performed to isolate the FGs. Finally, the cell is completed by tunnel oxide formation and the deposition of another poly-layer as the conductive channel.

The RG NAND cell formation process is somewhat similar to FG NAND, except that instead of an oxide/polyfilm stack it starts with an oxide/nitride stack. After the cell hole etch, oxide/nitride/oxide/polyfilms are deposited inside the holes, forming both the charge storage layer (nitride) and the conductive channel (poly). Afterward, a unique WL cut and replacement process is performed as illustrated in Fig. 12 (adapted from Fig. 4 of [32]). After the WLs are cut into smaller segments, a wet

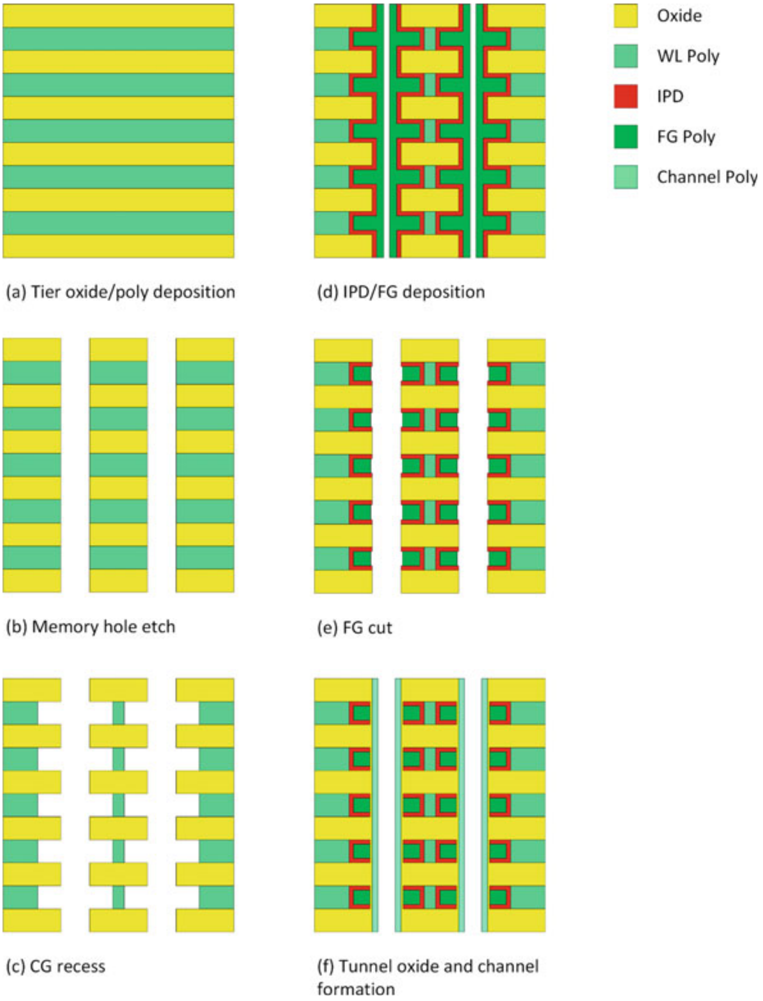
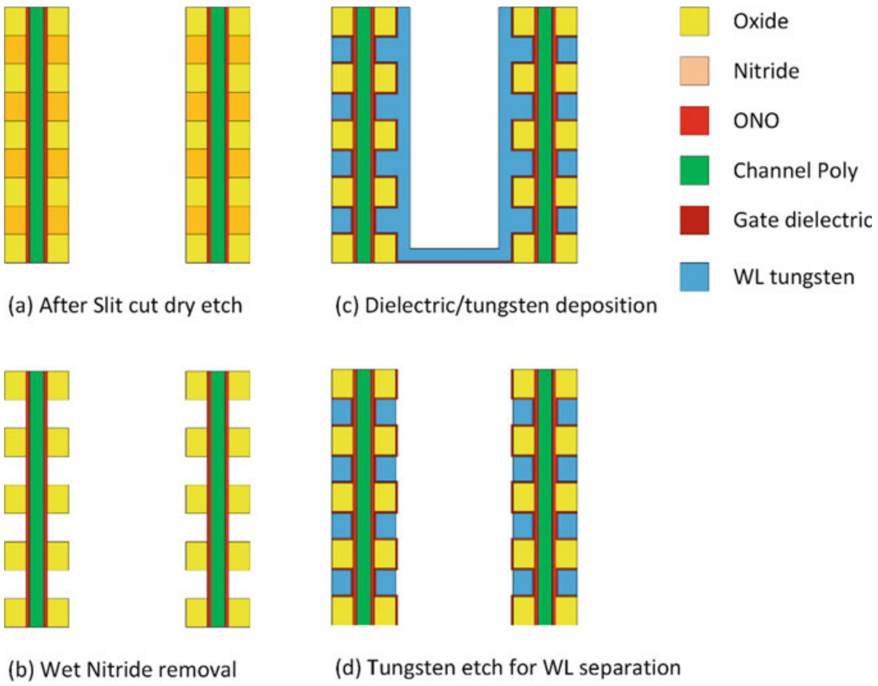


Fig. 11 FG NAND cell formation process



(a) After Slit cut dry etch

(c) Dielectric/tungsten deposition

(b) Wet Nitride removal

(d) Tungsten etch for WL separation

Fig. 12 Gate replacement process of RG NAND

etch is performed to remove all the nitride film from the cut opening. Then, a dielectric film (part of IPD) is deposited, followed by W deposition, replacing the nitride as the WL conductor material (therefore the RG moniker). Another W etch is needed to separate the WLs.

The biggest advantage of RG NAND over FG NAND is the much lower resistivity of *W* compared to poly. This allows RG NAND to support much longer WLs at comparable WL RC delay than FG NAND. On the other hand, it has a couple of drawbacks too. First, it is hard to integrate the "gate replacement" process with the FG cell structure, so RG NAND has to use a charge trap layer (generally silicon nitride) as the charge storage medium for the NAND cells, which tend to have smaller V_T window and worse data retention properties. Second, to perform the "gate replacement," it requires a slit for WL cut between every few holes, effectively increasing the unit cell size. A larger cell size directly translates into a higher cost/bit.

2.5.2 3D NAND Unique Features

A. Closely packed cells

Figure 13 shows the typical 3D NAND cell layout. Each NAND cell is a collection of concentric circles consisting of oxide filler at the center, a poly channel, tunnel oxide, FG or charge trap layer, and IPD. The space between the outer IPD of each cell is filled with the CG (poly for FG NAND and *W* for RG NAND) material forming the WLs. To maximize the cell density, they are packed into a honeycomb pattern, with every three adjacent cells forming an equilateral triangle. The cell pitch along each side of the triangle is *d*, and it has to be big enough to accommodate all the components outlined above. With such close packing, the cell size is $<d^2$ ($0.866d^2$, to be exact). Then considering the fact that with 3D NAND, each cell in the layout is actually *N* cells overlaid on top of each other (*N* is the total number of WL tiers), the effective cell size is actually $0.866d^2/N$. So, by scaling either cell pitch *d* or the tier count *N*, 3D NAND can reduce the effective cell size from generation to generation, achieving ever lower cost/bit.

Another feature of 3D NAND is the mismatch between the BL pitch *p* and cell pitch *d*. As shown in Fig. 13b, the BL pitch is only a quarter of the cell pitch, allowing 4 BLs to run through between two adjacent cells. By employing off centered plugs to connect the BLs and the NAND string channels, adjacent BLs are tapping into cells in different rows. This allows multiple rows of the 3D NAND strings to be organized into a single page, effectively reducing the WL length for a given page size. The shorter the WL, the better the RC and Read/Write performance.

B. Staircase for WL connections

As shown in Fig. 11, one key element for 3D NAND to realize tremendous process cost saving is the ability to form memory cells in multiple tiers of WLs simultaneously. However, to operate NAND strings, all those WLs have to be biased independently, which means that they all need their own drivers. And there has to be an economical way to form WL contacts that can tap into every single tier of WLs and connect them to their drivers. Given that today's state-of-the-art 3D NAND technologies employ hundreds of tiers, it would have been economically prohibitive to use a dedicated lithography layer to define the contact region for every tier of WLs. An ingenuous way to solve this challenge is to form a "staircase" of the WLs with a single lithography step, as shown in Fig. 14 (adapted from Fig. 8 of [33]).

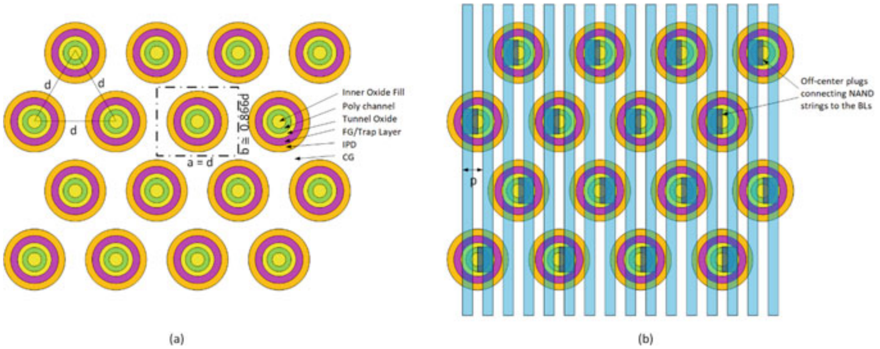


Fig. 13 3D NAND cell layout **a** closely packed cells, **b** cell-BL connections

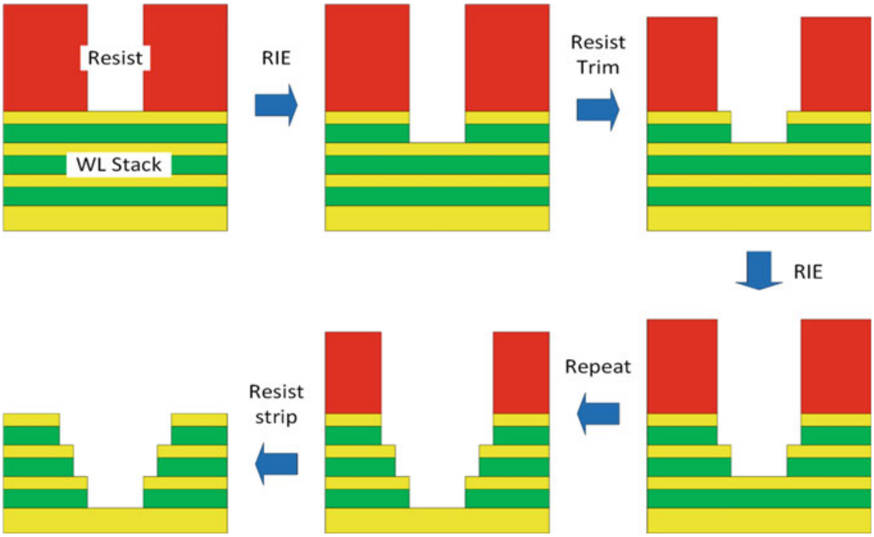


Fig. 14 Scheme to form WL "staircase" with a single lithography layer

Starting with a very thick resist layer, a trench (or line) is formed in the resist with conventional photolithography. Then an anisotropic Reactive-Ion Etching (RIE) is performed to etch away one tier of WL, stopping on the oxide layer between tiers. Afterward, an isotropic resist "slimming" is done with very high selectivity to the WL stacks. The slimming process shrinks the resist edges and exposes some new area that did not get etched in the first WL RIE step. Then, another WL RIE step is performed to etch away 1 oxide and 1 WL layer that are exposed. By repeating the anisotropic WL RIE + isotropic resist slimming steps as many times as needed, a WL staircase is formed, leaving every WL tier exposed. Then, a single WL contact level can be implemented to connect to all the WL tiers simultaneously. Of course, this is a challenging process to develop, because the contact height varies greatly between the top and the bottom tiers. The contact etch has to be very selective to the WLs to guarantee correct connection without shorting different WL tiers together.

C. CMOS under the array

As mentioned in Section A, because the NAND cells in all the WL tiers are stacked on top of each other, the cell size is effectively amortized between all the tiers, resulting in a much smaller effective cell size that is inversely proportional to the tier count. In addition to the continued cell size reduction with ever-increasing WL tier count, another way to further reduce the NAND die size is to tuck the periphery CMOS control circuits under the array as well.

In a conventional NAND product, the periphery circuits can easily take up 30 ~ 40% of the overall die size. So putting them under the array can generate significant die size savings. To realize the cost benefit (smaller die size translates directly into lower cost/bit), two things are necessary,

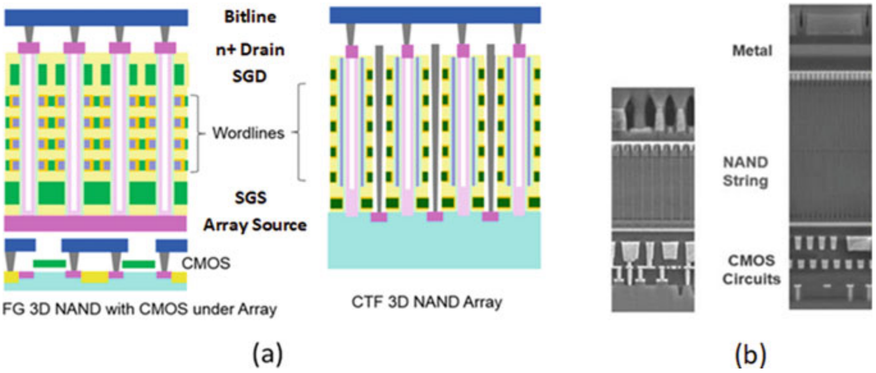


Fig. 15 **a** Comparison between CMOS under Array (CUA) architecture and non-CUA, **b** cross-sectional SEM images of Intel/Micron's 64-layer 3D NAND product with CUA (adapted from Figs. 3 and 4, respectively, of [34])

a. The NAND memory formation does not need direct access to the substrate.
b. There has to be additional interconnect layers connecting the CMOS circuits below and the NAND array above.

Figure 15a shows a comparison between a FG 3D NAND with CMOS under Array (CUA) architecture and a charge trap flash (CTF) 3D NAND without CUA [34]. The FG NAND array employs an n+ polysilicon source plate above the Si substrate while the CTF NAND array has both the SGS and the source line formed in the Si substrate. The polysilicon source plate enables the placement of the CMOS circuits under the array without any impact on the NAND array functionality. The additional under-the-array interconnect layers would increase the wafer cost somewhat, offsetting the cost benefit of CUA slightly. But the overall cost/bit benefit is still significant. Figure 15b shows the cross-sectional SEM images of Intel/Micron's 64-layer 3D NAND product with CUA.

3 3D NAND Technology and Design Challenges

3.1 Cost-Performance-Reliability Tradeoffs

To develop a successful nonvolatile memory product, multiple metrics have to be balanced—cost, Read/Write speed, reliability, power, and form factors, etc. Of them, the cost is the most important. It is the relentless cost/bit reduction of NAND Flash in the past 3 decades that opened it up to numerous applications. For other metrics, there are niche applications where some of them are highly valued. But for most products, the requirements can be relaxed somewhat as long as a minimum can be met.

The cost/bit can be calculated quite simply by the following formula, where c is the cost/bit in \$, WC is the manufacturing wafer cost in \$, M is memory density in # of bits, and N is the # of yielding dies per wafer.

$$c = \frac{\mathrm{WC}}{(M \times N)}$$

This calculation excludes additional costs related to the testing and packaging of the dies because they are usually much smaller compared to the manufacturing wafer cost. The manufacturing wafer cost is dominated by the depreciation of the tools needed to complete the fabrication process—dry etch, wet process, photolithography, deposition, and Chemical–Mechanical-Polishing (CMP), etc. As the technology moves to a more advanced node (marked by a smaller geometry for 2D or a larger tier count for 3D), more sophisticated tools are usually required, driving *WC* higher. Then to realize a lower cost/bit, $(M \times N)$ has to be maximized. While M is generally chosen based on the target applications, to maximize N, the die size has to be minimized during the design stage.

Either technology or design choices that can lower the cost/bit usually have an impact on the performance or reliability of the memory product. For example, for the most straightforward tier count increase in 3D NAND, it reduces cost/bit by reducing the effective cell size, which is inversely proportional to WL tier count N. With a higher tier count, the NAND string gets longer (more transistors in series). On the one hand, the longer string results in a lower string current, reducing Read performance. At the same time, more WLs along the NAND string means each cell is stressed more by the operations of the other cells along the same string, degrading the memory reliability. So corrective measures have to be taken to recover the loss in performance and reliability to meet the product spec, some of which may increase the wafer cost slightly. At the end of the day, each successively more advanced node has to deliver a significant enough cost/bit saving while meeting the product performance and reliability specs.

3.2 3D NAND Technology Challenges

The primary scaling path of 3D NAND is the increase in tier count as evidenced by the latter half of Fig. 4b. As of July 2021, multiple manufacturers have announced products up to 176-tiers [35, 36], and even higher tier counts are in the works at the R&D stage. This skyward march is indispensable for the continued bit cost reduction and storage capacity increase. However, it also brings with it a tremendous amount of technological challenges, which are discussed briefly here.

3.2.1 Process Scaling Challenges

One key element enabling low-cost 3D NAND technologies is the Bit Cost Scalable (BiCS) concept [37]. As shown in Fig. 16 (adapted from Fig. 1 of [37]), with BiCS, all the film stack needed for the NAND cells are first deposited, and then, holes are patterned and etched through the whole film stack. Subsequent process steps form all the NAND storage elements at the same time. With this scheme, when the tier count increases, all the memory cells can still be formed with the same number of process steps in principle, therefore holding the wafer process cost relatively constant and reducing the cost/bit.

However, the reality is more complicated. As the tier count increases, the hole etch becomes more and more challenging. To maintain the bit cost scaling, the physical cell size of 3D NAND is kept at a constant. The more tiers there are, the thicker the overall film stack, and the higher the AR of the holes needed to form the NAND cells. AR is defined as the ratio of the height and the diameter of the holes. In state-of-the-art 3D NAND technologies, the stack height of 100 + memory tiers is usually several μm, while the diameter of the holes is ~100 nm, leading to ARs greater than 50 or even 100. It is also important to maintain a very straight profile for the sidewall of the holes. With such high AR, even a small deviation from the right angle leads to a large Critical Dimension (CD, in this case the hole diameter) change from top to bottom.

To realize such high AR etch at an extremely low defect rate, very high power plasma and advanced etch tools are required. The challenges to be solved are not just the etch process itself. For such higher power and high AR etch, a hard mask strong enough to withstand it on top of the stack, a solid etch stop layer at the bottom to protect the films under the stack, and the prevention of the plasma attack on the sidewall all need to be addressed. The high power etch tends to amplify any tiny imperfections of the hole patterns. It is also sensitive to local loading effect and can be perturbed by any variations of the hole density. So the design of the hole pattern,

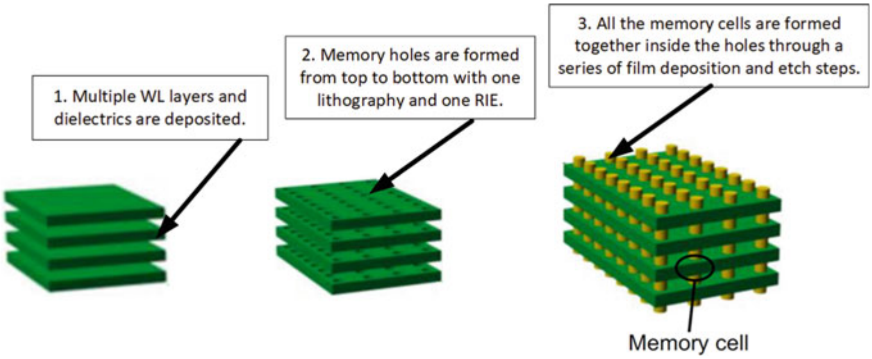

1. Multiple WL layers and dielectrics are deposited.
2. Memory holes are formed from top to bottom with one lithography and one RIE.
3. All the memory cells are formed together inside the holes through a series of film deposition and etch steps.
Memory cell

Fig. 16 Basic concept of BiCS technology

especially the array/periphery interfaces where a uniform hole pattern cannot be maintained, and the lithography process needs to be carefully optimized as well.

Memory hole etch is not the only high AR etch step of the 3D NAND technologies. As tier count increases, the AR of the WL staircases, the WL contacts, and the periphery contacts between the above-array metal routing layers and the below-array CMOS devices all increase along with it. While their ARs are significantly lower than that of the memory holes, they have their own set of requirements and can be quite challenging as well.

Another process scaling challenge is the management of the film stresses. Most of the films used in 3D NAND are not stress-free. The film stress also changes with temperature, and the wafer temperature can vary by hundreds of degree C throughout the manufacturing processes. The more tiers of the memory cells, the higher the aggregate stress of the whole film stack. The film stress can lead to significant wafer warpage and cause inaccuracies in lithography and overlay errors between different layers. In the worst case, the warpage can prevent correct wafer handling by the manufacturing tools, or even break the wafer. In addition to engineering the film stress for the memory tiers, one common solution is to deposit another stressful film at the backside of the wafer to balance out the stress on the front side, reducing the overall warpage.

3.2.2 Memory Cell Optimization

Increasing tier count not only makes the manufacturing process more challenging, it also carries electrical implications to the 3D NAND memory operations. When NAND technologies transitioned from 2D to 3D, the conduction channel material is changed from single-crystalline Si to polycrystalline Si (poly). Poly has significantly lower electron mobility compared to single-crystalline Si because of the excess scattering at the grain boundaries. As a result, the string current is reduced significantly. Because the cells in a NAND string are connected in series, as the tier count increases, the NAND string gets longer with more cells and the string current drops even further.

The reading of the NAND memory state is performed by sensing the string current. If the current is too low, both the read speed and accuracy (error rate) can suffer. Innovations have to be in place to recover the string current with increasing tier count. It usually involves increasing the poly channel grain size and reducing the amount of interface states.

When there are more cells in the NAND strings, the variations from cell to cell become a bigger problem that has to be carefully controlled. The variations contain both systematic and random components. For example, the high AR hole etch will produce a systematic diameter variation from top to bottom, with the top CD usually larger than that at the bottom. The deposition of various films to construct the 3D NAND cells tends to produce a systematic thickness variation from top to bottom as well. These systematic variations can be compensated with other process vectors (intentional doping/thickness changes from top to bottom), or addressed by Design through WL-specific biasing schemes.

Random variations in CD, film thicknesses, poly channel grain structures, etc., lead to variations in 3D NAND cell electrical characteristics, and generally result in wider V_T distributions and smaller read windows. Both process optimizations and advanced Program/Erase algorithms are needed to manage the random variations.

3.3 3D NAND Design Challenges

To design a competitive 3D NAND product, the most important metric is the die size, because it directly translates into the overall cost-per-bit—the smaller the die size, the more total bit output per wafer, therefore the lower cost-per-bit. At the same time, it is also important to maintain or improve the performance, usually characterized by t_R, t_{Prog}, and I/O bandwidth, as defined below.

t_R—Read time measured in μs is the time it takes to complete a Read operation of 1 page data.

t_{Prog}—Program time measured in μs is the time it takes to complete a Program operation of 1 page data. For multi-bit-per-cell technologies, different logic pages on the same WL have vastly different t_{Prog} due to the significant increase in Program operation complexities when programming the higher pages. At the product level, t_{Prog} usually refers to the average Program time of all the logic pages.

I/O bandwidth—Measured in GT/s refers to the speed at which one I/O pad can transmit or receive data.

As the tier count increases for each successive generation of 3D NAND technology, simplistically speaking, there are two ways to design the new product. One way is to keep the memory capacity constant, in which case the memory organization and the periphery circuitry can stay largely unchanged. However, with more tiers in the string, each NAND block contains more pages, so the total number of blocks is reduced compared to the previous generation. The key drawback with reduced number of blocks is that they take up a smaller die area, so all the periphery circuitry is occupying a relatively larger percentage of the die. This reduced array efficiency (percent of the die size utilized by the actual memory array) drags down the effectiveness of the cost scaling.

Alternatively, the new product can be designed with a larger memory capacity so the new array takes up as much (or even larger) die area as in the previous generation to maintain a comparable or better array efficiency. This approach produces efficient cost scaling but somewhat compromises the performance at the system level. To understand why, one has to recognize that modern storage systems usually employ multiple NAND dies and operate them in parallel to maximize the Read/Write performance. With each individual die having a larger capacity with the same Read/Write performance, the systems built with newer generation of chips need fewer dies to reach the same storage size. With fewer dies working in parallel, the new system Read/Write performance would actually degrade compared to the old ones. With this effect in mind, the system builders tend to demand the 3D NAND chipmakers to scale their technologies with a constant performance per GB, meaning if the individual die capacity increases, the die level Read/Write/IO performance should scale up proportionally.

Regardless of the design choice in die capacity, it is critical to optimize all the components that can help improve the array efficiency. Here, the WL connections will be discussed as an example. As illustrated in Fig. 14, 3D NAND requires a rather elaborate staircase to connect all the WLs. No active cells are present in this connection area so it counts as an array overhead. To minimize this overhead, several things have to be considered,

1. The frequency of the staircase or the length of the WL—To put it simply, for the same size of the staircase, the longer the WL, the smaller it takes up as an overall percentage. However, the length of the WL is usually limited by the WL RC. In NAND, ramping the WLs to the needed voltage is one of the limiting factors of the Read/Write speed. That puts a limit on the WL length.
2. The placement of the staircase—Given the WL RC limitation, it would be more economical to place the staircase at the center of WLs, effectively doubling WL length per staircase at the same WL RC.
3. The size of the steps—The smaller the steps of the staircase, the smaller the overall footprint. However, the steps are where the WL contacts land on each individual WL; shrinking the step size limits the WL contact size and reduces the contact landing margin. Misaligned WL contacts can lead to yield loss or reliability problems, which require more advanced process control to manage.
4. Staircase design—The staircase does not have to be a one-dimensional design as shown in Fig. 14. Across the width of the NAND block, there can be multiple zones, each having a staircase for a unique set of WLs. The arrangement of different zones can be optimized to reduce the overall staircase footprint.
5. WL contact routing—In addition of the size of the staircase itself, the WL contacts have to be routed to their individual driver transistors. The routing can take up quite a bit of the die area as well and needs to be carefully optimized.

4 NAND Reliability Issues

The unique operating mechanisms bring many challenges to NAND array reliability. Because the Program/Erase operations involve a very high electric field in the tunnel oxide in excess of 10 MV/cm, the large electrical stress degrades the oxide quality through the generation of electron and hole traps inside it. These traps cause extra leakage (commonly referred to as the Stress-Induced Leakage Current, of SILK) and weaken the ability of NAND cells to hold their V_T states. To manage such degradation, NAND Flash devices need to limit the number of Program/Erase cycles each memory cell can perform, which is called the array endurance. Within the specified number, some basic reliability characteristics are guaranteed, such as the minimum data retention time under certain temperature conditions, and the maximum error rate when reading the data. Beyond the endurance spec, the NAND Flash is no longer guaranteed to be nonvolatile [4]. It is very important for the product designers to keep this fundamental limit in mind, as highlighted in a most recent Tesla recall [40]. If too many Program/Erase operations are performed on the NAND Flash beyond its specification, device can potentially malfunction. As NAND Flash is more widely integrated into different applications beyond the conventional media storage, e.g., mission critical advanced driver assistance systems or autonomous driving systems [41], the reliability issues demand elevated attention to ensure overall system quality and public safety.

According to [4], the NAND reliability issues can be broadly classified into three big categories—V_T placement errors during programming, or write errors; V_T changes induced by other array operations, or disturb errors; and the V_T changes due to inherent time-dependent instability, or data retention errors. This classification is based on the underlying mechanisms. In reality, all of them are affecting the V_T states of the memory cells in various degrees, depending on many factors, e.g., the exact operating sequences they go through, the data patterns being written, and the algorithms used for Read/Program operations.

4.1 Write Errors

Write operations of NAND Flash are usually accompanied by the Verify operations, which is a special read to determine if the cell V_T exceeds the target V_T. So the most common write error is overprogramming during the last Program pulse. Overprogramming can happen due to the stochastic nature of the FN tunneling, especially for extremely scaled memory cells where the number of electrons involved is small [38]. It can also occur when there are defects in the tunnel oxide, or due to damages caused by the previous Program/Erase cycles in the form of extra electron/hole traps in the tunnel oxide. With excessive cycling damages, sometimes underprogramming can occur as well due to “instant charge loss” post-program verify. To mitigate write errors, it is desirable to ramp up the programming voltages in finer steps, or employ more advanced programming algorithms.

4.2 Disturb Errors

In a sense, disturb is a phenomenon inherent to closely packed memory arrays. When the memory cells are packed close to each other, the operation of one cell will affect the V_T state of the adjacent cells through the parasitic coupling between them. For NAND, the serial connection of NAND cells triggers additional mechanisms for their coupling because any operations on one memory cell necessarily involve all the cells in the same NAND string.

There are three different kinds of disturbs to be considered for the NAND operations. The first is the Program interference. The NAND control gates (WLs), floating gates (FG), and the conduction channel form a capacitive network, through which the NAND string is turned on or off by the WL voltages. However, there is additional coupling between adjacent FGs and between FG and the neighboring channels, as shown in Fig. 17. The effect of these unwanted coupling is that the V_T of each individual NAND cell is not just a function of the electrical charge stored on its own FG but can also be influenced by the amount of charge on the neighboring FGs. For example, if an erased cell is surrounded by programmed cells in both the WL and BL directions, its V_T can increase substantially and appear as programmed. Basically, the V_T of the "victim cell" is interfered with by the programming of the adjacent "aggressor cells." The program interference leads to a loss of read window or even data corruption.

As shown in Fig. 17, for 2D NAND, FG-FG coupling can be quite significant in both the WL and BL directions. And, it gets progressively worse when the space between the FGs gets smaller in more advanced nodes. In 3D NAND, on the other hand, FG-FG coupling is an issue only in the vertical direction (along the channel)

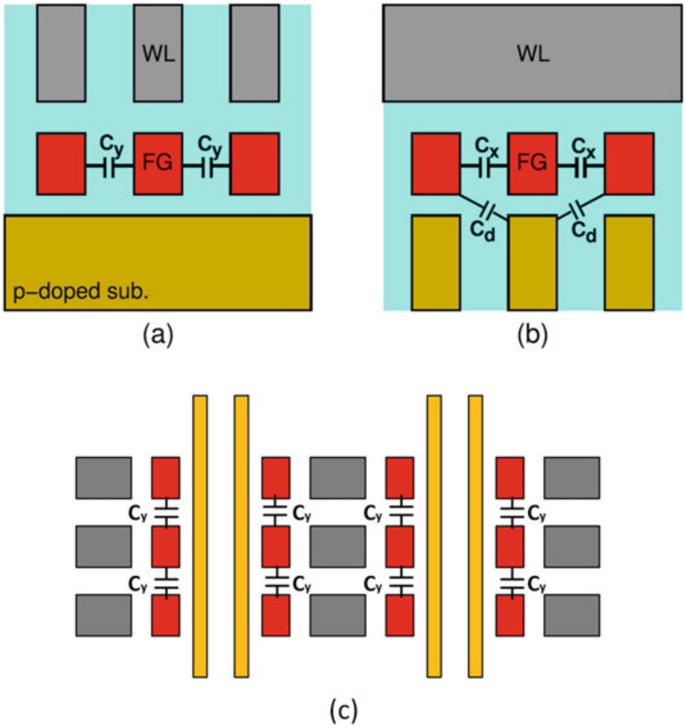


Fig. 17 Illustration of the capacitive coupling between the floating gates in 2D and 3D NAND arrays giving rise to cell-to-cell interference **a** Coupling along the 2D NAND string, **b** coupling between the 2D NAND string, **c** coupling between 3D NAND cells

between the WLs, because the CG fully wraps around the FG and provides effective shielding between the FGs within the same WL. This is another architectural advantage of 3D NAND over 2D.

To mitigate Program interference, on the process side, the cell construction can be engineered to minimize the FG-FG coupling, e.g., reducing the FG height, creating a WL-wrap-around of FG as a shield, or introducing an airgap between the FGs to reduce their coupling. On the design side, the programming sequence of different WLs (pages) and the programming algorithms can be optimized to reduce the program interference.

The second kind is Program Disturb. As the name suggests, it refers to the unwanted cell V_T changes during programming. During NAND Program operation, high voltage between the selected WL and the grounded channels of the selected (grounded) BLs create intended V_T changes for the selected cells (Fig. 7a). When we examine the voltage difference between the WLs and the channels of all the cells in the selected block, they can be divided into four categories.

1. Cells between the selected WL and selected BLs see the full programming voltage V_{Pgm}.
2. Cells between the selected WL and the inhibited BLs see the voltage difference between V_{Pgm} and the V_{Inh}, ~ V_{Pgm}-V_{Inh}.
3. Cells between the unselected WLs and the selected BLs see the full Inhibit voltage VInh.
4. Cells between the unselected WLs and the inhibited BLs see ~ 0 V because the channels are coupled up to ~ V_{Inh}.

Of the four, categories 2 and 3 both have moderately high voltage across the tunnel oxide. Even though the resulting electric field is not strong enough to program the cells during the Program pulses, there can be appreciable FN tunneling current (or SILK current if the cells have gone through many Program/Erase cycles) to induce a small amount of V_T shift. When different pages are programmed in the same block, the disturb effect is cumulative and can lead to read failures.

The last kind is the Read Disturb. During the Read operation, because of the serial nature of the NAND cell connection along the same string, all the unselected WLs have to have a relatively high V_{Pass} applied on them to ensure those cells being on regardless of their V_T states. Similar to the Program Disturb, V_{Pass} also generates a moderately high electric field inside the tunnel oxide of the unselected cells and cause their V_T to increase. This effect gets worse after the tunnel oxide is degraded by the Program/Erase stress after many cycles. And because NAND Flash needs to perform many more Read operations than Program operations, Read Disturb can be a more serious reliability concern.

4.3 Data Retention Errors

Data retention errors refer to the cell V_T changes without going through any memory operations. There are two main mechanisms causing data retention errors—SILK-induced charge loss from the storage node (FG) and the electron detrapping from the tunnel oxide.

SILK can be a serious issue when there are a large enough number of defects in the tunnel oxide. These defects can either come from the manufacturing process or be induced by the Program/Erase stress after many cycles. Usually only a small number of cells have significant SILK leakage, so it manifests in NAND V_T distributions as a tail (mostly a low V_T tail for the programmed cells, but can be a high V_T tail for the Erased cells as well).

Even when the number of defects in the tunnel oxide is small, they can trap electrons during the NAND Program/Erase operations. These electrons can escape later (or get detrapped) and cause the cell V_T distributions to widen and shift lower. The electron trapping/detrapping is sensitive to the conditions of both the Program/Erase cycling and the data retention bake [39]. When the device is baked at a higher temperature, there will be more electron detrapping due to the stronger thermal activation. If the device is cycled at a higher temperature or with longer delay time in between, however, the electron trapping/detrapping will be less because some traps can be annealed out between the cycles.

5 3D NAND Future Outlook

Since the 2D to 3D NAND transition around 2014, 3D NAND Flash has come a long way and become the most dominant nonvolatile memory technology in the market. Its success was built on the relentless scaling of the cost/bit by increasing the number of WL tiers ever higher. This trend will no doubt continue despite the process and design challenges associated with high AR memory hole and contact etches, stress management, and reduced string current, etc. Currently 3D NAND technologies with hundreds of tiers are already in the works at the major NAND Flash manufacturers. Various process and design innovations are being developed and deployed to address those challenges and keep delivering higher performances at both the product and system levels.

Beyond a few hundred tiers, the scaling trend is expected to slow down. However, 3D NAND Flash is likely to remain the dominant nonvolatile memory technology for the foreseeable future. When the process technology scaling slows down due to the ever-increasing difficulties and diminishing return in the cost/bit benefit, it will push the innovation front to the memory architecture, packaging technologies, and the product design. As long as the explosion of the data storage needs continues (which is showing no sign of subsiding), 3D NAND Flash will keep pushing forward to meet the demand.

Acknowledgements The author sincerely thanks Dr. Zhen Wei and Dr. Xi Qian for their detailed review.

References

1. F. Masuoka *et al*, "A New Flash E2PROM Cell Using Triple Polysilicon Technology," *IEDM Tech. Dig.*, Dec. 1984, pp. 464–467.
2. F. Masuoka *et al*, "New Ultra High Density EPROM and Flash EEPROM with NAND Structure Cell," *IEDM Tech. Dig.*, Dec. 1987, pp. 552–555.
3. S. Aritome, "NAND Flash Memory Technologies," *IEEE Press*, 2016.
4. Monzio Compagnoni *et al*, "Reliability of NAND Flash Arrays: A Review of What the 2-D to 3-D Transition Meant," *IEEE Trans. Electron Devices*, Vol. 66, No. 11, pp. 4504-4516, Nov. 2019.
5. Y. Tseng *et al*, "A 34MB/s-program-throughput 16Gb MLC NAND with All-bitline Architecture in 56nm," *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2008, pp. 420–421.
6. Y. Li *et al*, "A 16Gb 3b/cell NAND Flash Memory in 56nm with 8MB/s Write Rate," *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2008, pp. 506–507.
7. R. Zeng *et al*, "A 172mm^2 32Gb MLC NAND Flash Memory in 34nm CMOS," *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2009, pp. 236–237.
8. C. Trinh *et al*, "A 5.6MB/s 64Gb 4b/cell NAND Flash Memory in 43nm CMOS," *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2009, pp. 246–247.
9. T. Futatsuyama *et al*, "A 113mm^2 32Gb 3b/cell NAND Flash Memory," *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2009, pp. 242–243.
10. C. Lee *et al*, "A 32Gb MLC NAND-flash Memory with Vth-endurance-enhancing Schemes in 32nm CMOS," *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb, 2010, pp. 446–447.
11. G. Marotta *et al*, "A 3bit/cell 32Gb NAND Flash Memory at 34nm with 6MB/s Program Throughput and with Dynamic 2b/cell Bocks Configuration Mode for A Program Throughput Increase up to 13MB/s," *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2010, pp. 444–445.
12. H. Kim *et al*, "A 159mm^2 32Gb MLC NAND-flash Memory with 200MB/s Asynchronous DDR Interface," *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2010, pp. 442–443.
13. K-T. Park *et al*, "A 7MB/s 64Gb 3-bit/cell DDR NAND Flash Memory in 20nm-node Technology," *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2011, pp. 212–213.
14. K. Fukuda *et al*, "A 151mm^2 64Gb MLC NAND Flash Memory in 24nm CMOS Technology," *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2011, pp. 198–199.
15. Lee *et al*, "A 64Gb 533Mb/s DDR Interface MLC NAND Flash in Sub-20nm Technology," *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2012, pp. 430–431.
16. Naso *et al*, "A 128Gb 3b/cell NAND Flash Design Using 20nm Planar-cell Technology," *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2013, pp. 218–219.
17. K-T. Park *et al*, "Three-dimensional 128Gb MLC Vertical NAND Flash-memory with 24-WL Stacked Layers and 50MB/s High-speed Programming," *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2014, pp. 334–335.
18. M. Helm *et al*, "A 128Gb MLC NAND-flash Device Using 16nm Planar Cell," *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2014, pp. 326–327.
19. S. Choi *et al*, "A 93.4mm^2 64Gb MLC NAND-flash Memory with 16nm CMOS Technology," *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2014, pp. 328–329.

20. M. Sako *et al*, “A Low-Power 64Gb MLC NAND-flash Memory in 15nm CMOS Technology,” *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2015, pp. 128–129.
21. J-W. Im *et al*, “A 128Gb 3b/cell V-NAND Flash Memory with 1Gb/s I/O Rate,” *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2015, pp. 130–131.
22. T. Tanaka *et al*, “A 768Gb 3b/cell 3D-floating-gate NAND Flash Memory,” *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Jan./Feb. 2016, pp. 142–143.
23. D. Kang *et al*, “246Gb 3b/cell V-NAND Flash Memory with 48 Stacked WL Layers,” *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Jan./Feb. 2016, pp. 130–131.
24. R. Yamashita *et al*, “A 512Gb 3b/cell Flash Memory on 64-word-line-layer BiCS Technology,” *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2017, pp. 196–197.
25. C. Kim *et al*, “A 512Gb 3b/cell 64-stached WL 3D V=NAND Flash Memory,” *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2017, pp. 202–203.
26. S. Lee *et al*, “A 1Tb 4b/cell 64-stacked-WL 3D NAND Flash Memory with 12MB/s Program Throughput,” *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2018, pp. 340–341.
27. Maejima *et al*, “A 512Gb 3b/cell 3D Flash Memory on a 96-word-line-layer Technology,” *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 336–337.
28. N. Shibata *et al*, “A 1.33Tb 4-bit/cell 3D-flash Memory on a 96-word-line-layer Technology,” *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 210–211.
29. C. Siau *et al*, “A 512Gb 3-bit/cell 3D Flash Memory on 128-wordline-layer with 132MB/s Write Prformance Featuing Circuit-under-array Technology,” *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 218–219.
30. Goda, “3-D NAND Technology Achievements and Future Scaling Perspectives,” *IEEE Trans. Electron Devices*, vol. 67, No. 4, pp. 1373–1381, April 2020.
31. K. Parat *et al*, “A Floating Gate Based 3D NAND Technology with CMOS Under Array,” *IEDM Tech. Dig.*, Dec. 2015, pp. 48–51.
32. J. Jang *et al*, “Vertical Cell Array using TCAT (Terabit Cell Array Transistor) Technology for Ultra High Density NAND Flash Memory”, *VLSI Symp. Tech. Dig.*, Jun. 2009, pp. 192–193.
33. Tanaka *et al*, “Bit Cost Scalable Technology with Punch and Plug Process for Ultra High Density Flash Memory,” *VLSI Symp. Tech. Dig.*, Jun. 2007, pp. 14–15.
34. K. Parat *et al*, “Scaling Trend in NAND Flash,” *IEDM Tech. Dig.*, Dec. 2018, pp. 27–30.
35. Micron Technology Press Release, https://investors.micron.com/news-releases/news-release-details/micron-ships-worlds-first-176-layer-nand-delivering-breakthrough, Nov. 2020.
36. J- W. Park *et al*, “A 176-Stacked 512Gb 3b/Cell 3D-NAND Flash with 10.8Gb/mm2 Density with a Peripheral Circuit Under Cell Array Architecture,” *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2021, pp. 422–423.
37. Nitayama *et al*, “Bit Cost Scalable (BiCS) Technology for Future Ultra High Density Memories,” *VLSI Symp. Tech. Dig.*, Jun. 2013.
38. Monzio Compagnoni *et al*, “Analytical Model for the Electron-injection Statistics during Programming of Nanoscale NAND Flash Memories,” *IEEE Trans. Electron Devices*, vol. 55, no. 11, pp. 3192-3199, Nov. 2008.
39. N. Mielke *et al*, “Flash EEPROM Threshold Instabilities due to Charge Trapping during Program/Erase Cycling,” *IEEE Tran. Device and Materials Reliability*, vol. 4, no. 3, pp. 335-344, 2004.
40. Tesla recall information, https://www.tesla.com/support/8gb-emmc-recall-frequently-asked-questions, Dec. 2021.
41. M. Jung, *et al*, “Driving into the Memory Wall,” MEMSYS, Oct. 2018.
42. Park *et al*, “A 176-Stacked 512Gb 3b/Cell 3D-NAND Flash with 10.8Gb/mm^2 Density with a Peripheral Circuit Under Cell Array Architecture,” *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2021, pp. 422–423.
43. Khakifirooz *et al*, “A 1Tb 4b/Cell 144-Tier Floating-Gate 3D-NAND Flash Memory with 40MB/s Program Throughput and 13.8Gb/mm^2 Bit Density,” *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2021, pp. 424–426.

44. Cho *et al*, “A 512Gb 3/Cell 7th-Generation 3D-NAND Flash Memory with 184MB/s Write Throughput and 2.0Gb/s Interface,” *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2021, pp. 426–428.
45. Yuh *et al*, “A 1-Tb 4b/Cell 4-Plane 162-Layer 3D Flash Memory With a 2.4-Gb/s I/O Speed Interface,” *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2022, pp. 130–132.
46. W. Cho *et al*, “A 1-Tb, 4b/Cell, 176-Stacked-WL 3D-NAND Flash Memory with Improved Read Latency and a 14.8Gb/mm^2 Density,” *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2022, pp. 134–135.
47. Kim *et al*, “A 1Tb 3b/Cell 8th-Generaion 3D-NAND Flash Memory with 164MB/s Write Throughput and a 2.4G/s Interface,” *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2022, pp. 136–137.

Interconnect

Yongjun Huo, Yingxia Liu, and Fan-Yi Ouyang

Abstract Interconnect could provide physical and logical connection between two electronic devices with interconnect with affordable quality and reliability, and thus interconnect is critical for the development of Advanced Driver Assistance Systems (ADAS). In this chapter, interconnects and solder joint technology for applications under the hood are reviewed. In addition, the bonding techniques by using Cu-Cu direct bonding and hybrid bonding are presented.

1 Interconnects for Applications Under the Hood

Power electronics are frequently used in the automobile applications, where the power electronics component must serve in the high-working temperature environment under the hood. The peak ambient temperature can reach up to 150°C for the automotive engine control electronics or the electric/hybrid vehicle power management and distribution (PMAD) system [1]. Therefore, the interconnecting technology or the die-attachment method for the power electronics chip must have a high-working temperature (>200 °C) to serve this purpose. In addition, the power electronics need to regulate the high current and high voltage, where the electrochemical migration (ECM) phenomenon [2] often cause serious reliability issues. Therefore, in this session of the chapter, we will introduce and compare two mainstream interconnection technologies for the power electronics packaging, namely, nanoparticle (NP) sintering method and transient liquid phase (TLP) bonding technology. Both interconnecting methods above shared one important technological

Y. Huo
School of Material Science and Engineering, Beijing Institute of Technology, Beijing, China

Y. Liu
Department of Advanced Design and System Engineering, City University of Hong Kong, Hong Kong SAR, China

F.-Y. Ouyang (✉)
Department of Engineering and System Science, National Tsing Hua University, Hsinchu, Taiwan
e-mail: fyouyang@ess.nthu.edu.tw

Y. Li and H. Shi (eds.), *Advanced Driver Assistance Systems and Autonomous Vehicles*,
https://doi.org/10.1007/978-981-19-5053-7_6

advantage, i.e., achieving a high-working temperature bonding joint with a lower process temperature. In this chapter of the book, we would introduce the basic principles of nanoparticle sintering method and transient liquid phase bonding technology, and the electrochemical migration behaviors will be discussed for the long-term reliability of high-power electronics packaging.

1.1 Nanoparticle Sintering Method

The usage of metal nanoparticles as the low-temperature bonding medium has been extensively studied in detail by the electronics packaging community. Different kinds of metals, such as gold [3], silver [4], copper [5], nickel [6], and tin-based solders [7], can be prepared into the nanoscale particles with proper organic coatings and dispersion solvent. Among them, silver has the highest electrical and thermal conductivity, and possesses good mechanical properties, chemical stability at a reasonable manufacturing cost. Therefore, the low-temperature bonding technology with silver nanoparticles (AgNP) has received extra attention from the academic and industrial communities of electronic packaging and has already realized the industrial-level mass production [4], especially in the die-attachment process of power electronics for the automobile applications. Therefore, in this session, we would mainly focus on the AgNP sintering technology as a die-attachment method to discuss its suitability for the automobile applications.

When the size of metal particles reaches the nanometer scale (<100 nm), the ratio of surface area to volume of the material is extremely large, rendering the particles themselves in a high energy and unstable state. According to the second law of thermodynamics, to stabilize the system, the driving force to reduce the internal energy would lead to spontaneous agglomerations between the metallic nanoparticles. By using organic molecular coatings to encapsulate the metallic nanoparticles, the nanometer size of the particles can be maintained within the organic dispersion solvent. In the nanosintering process, only a small amount of thermal energy is required to decompose and evaporate the organic molecular coatings and the dispersion solvent, and then the metal particles in nanoscale would be spontaneously interconnected to each other by the agglomeration effect. In addition, due to the enhanced surface diffusion mechanism for the nanoscale particles, a molecular dynamics simulation study has shown that the melting point of the surface atoms on the nanoparticles will be significantly lowered [8]. Therefore, the nanoparticle interconnection process can be also assisted from a sufficient surface diffusion mechanism, whereby finishing the bonding process at a temperature much lower than the melting point of its corresponding bulk metal material. The sintered bonding joint retains its bulk material melting temperature, which is 962 °C for pure Ag. Therefore, AgNP sintering process has a significant technical advantage as a die-attachment technology, recognized as a low-temperature process for the high-temperature application. At relatively low temperature (<200 °C), the shear stress strength of AgNP bonding joints are ranging from 15 to 30 MPa [9], while the AgNP bonding joint possesses the capability to resist a certain level of thermal fatigue [10] because of the good ductility of pure Ag.

On the other hand, during the AgNP sintering process, it must decompose and evaporate the organic molecular groups encapsulating the nanoparticles, together with the organic dispersion solvent. Thus, the AgNP interconnected structure would always retain the outgassing channels formed during the escape of organic gas molecules, resulting in a porous microstructure of AgNP bonding joint, as demonstrated in Fig. 1. This porous microstructure is well-known as the inherent nature of the nanoparticle sintering method for the die-attachment. Due to this porous nature, AgNP bonding joint is not hermetic, which further leads to oxidation issues during the high-power operation of a device in the atmosphere environment. When power device operates at elevated temperatures, the porous AgNP bonding joint cannot prevent the connected Cu substrate from oxidation, which in turn leads to continuous degradation of the device performance. In the severe cases, it would even cause the peeling of the interconnected interface, resulting in the ultimate failure of the electronics device [11]. Accordingly, extensive research activities in the field have been dedicated in mitigating the negative effects of this porous nature. During the nano-sintering process, a certain level of applied pressure would help to reduce the porosity of the bonding joint. Usually, increasing the process temperature and bonding pressure within a certain range would lead to a proportional increase on the interfacial strength of the resulting bonding joint [9], through the reduction of the porosity level. Altering the size [11] and shape [12] of the metallic nanoparticles would also have a certain degree of influence on the interface quality of the AgNP bonding joint. With the persisting optimization and technical improvement, albeit with its porous nature, the AgNPs sintering method would still be a good candidate in the die-attachment application of high-temperature power electronics.

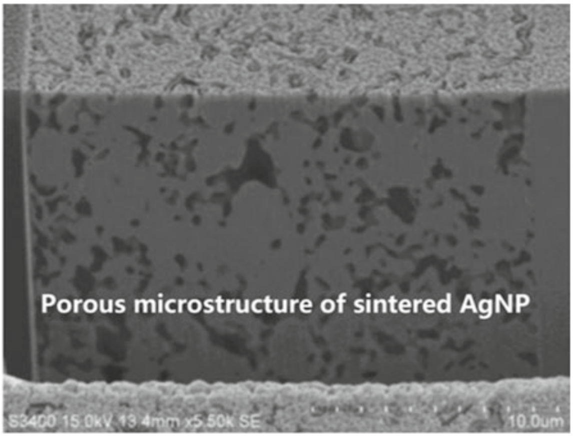

Porous microstructure of sintered AgNP
S3400 15.0kV 13.4mm x5.50k SE
10.0um

Fig. 1 The SEM images of the typical bonding joint produced by the AgNP sintering method: a representative porous microstructure

1.2 Transient Liquid Phase Bonding Technology

Transient Liquid Phase (TLP) bonding, also known as solid–liquid interdiffusion bonding (SLID), usually requires the design of a “high melting point-low melting point” metallurgical combination [13], whereby accomplishing the low-temperature bonding process through the interdiffusion between a solid phase and a temporary generated liquid phase. At a relatively low process temperature, the metal with a low melting point (low T_m) would turn into the molten phase and wet the surface of high melting point (high T_m) metal at the beginning stage of the bonding process. During the subsequent solid–liquid interdiffusion, the high T_m metal in its solid-state would continuously react with its counterpart low T_m metal in the liquid phase. At this stage, the ratio of high and low melting point metal elements at the bonding joint interface would continuously change, eventually, leading to the occurrence of corresponding phase transformation. Under isothermal conditions, the solidification process of the liquid phase would be finished, thereby forming a metallurgical bonding joint. At the final stage, the new phases formed at the bonding joint interface can be fully homogenized and stabilized through the sufficient diffusion process, whose melting points would be much higher than the original low T_m metal. Therefore, the transient liquid phase bonding technology has been remarked with the technical advantage of “a low-temperature process for the high-temperature application” as well.

Typically, the high melting point elements can be chosen from gold, silver, copper, and nickel, whereas the low melting point elements can be selected from tin, indium, and bismuth, for the design of a TLP bonding technology. In the literature, various binary systems of “high T_m -low T_m” metallic materials have been extensively studied in the utility of TLP bonding technology, such as gold-tin (Au–Sn) [14], gold-indium (Au-In) [15], silver-indium (Ag-In) [16], silver-tin (Ag-Sn) [17], copper-tin (Cu-Sn) [18], nickel-bismuth (Ni–Sn) [19], and etc. Among various kinds of TLP bonding technologies, the Au–Sn binary system has been widely recognized and applied in the electronics packaging industry because of its good mechanical properties, chemical stability, and long-term reliability. Similar to the AuSn eutectic method, the AuSn bonding with the TLP scheme could have a lower process temperature, not necessarily reaching to the eutectic point of Au–Sn system (280 °C), but rendering a high-strength bonding joint (>100 MPa) with identical compositions at the bonding interface [20]. The Au-In TLP bonding is another promising low-temperature die-attachment method for high-temperature power electronics. However, the Au-In binary system would produce an indium-rich phase ($AuIn_2$) as the dominant component in the bonding joint, which might raise long-term reliability concerns regarding to the indium oxidation and thermal migration issues. Importantly, the raw material cost of the gold-based TLP method was simply too high, so that it is affordable to the high-end power electronics products with a large profit margin. The high manufacturing cost would also put a limitation when broadening the range of applications with the economic considerations. Therefore, it is attractive to find alternative TLP bonding methods with equally good or even better performances compared to the Au-based TLP methods. Other types of tin-based TLP methods, such as Ag-Sn, Cu-Sn, and Ni–Sn TLP bonding systems, are less expensive, but their bonding joints were all composed by extremely brittle intermetallic compounds (IMCs), such as Ag_3Sn, Cu_3Sn, Cu_6Sn_5, and Ni_3Sn_4. The brittleness of the IMCs would lead to mechanical failures when the high-temperature power electronics devices experiencing the impacts of mechanical shock or thermo-mechanical shock [21] during the automobile applications.

The silver-indium (Ag-In) binary system has been proven to be a good candidate for the development of transient liquid phase technology. The Ag-In TLP bonding technique is usually conducted at low process temperatures (∼180 °C), forming a sandwich-like (Ag)-In/IMC/(Ag)-In bonding joint structure, where (Ag)-In represents the silver-indium solid solution. It has been demonstrated that the Ag-In bonding joint has a good temperature stability (>300 °C), a relatively high shear strength (>45 MPa) [22], and a great thermal fatigue resistance (> 5000 thermal cycles from − 40 °C to 200 °C) [23]. It has been found that the silver-indium solid solution is a highly ductile phase (tensile elongation >110%) with an excellent ultimate tensile strength (UTS >450 MPa) [24], whose tensile elongation and UTS value are both more three times higher than that of pure silver. The underlying mechanism for this plastic-toughness enhanced performance of (Ag)-In has been revealed as the nano-twinning microstructure induced plasticity and strengthening [25]. Therefore, the Ag-In TLP bonding technique is a good candidate as low-temperature die-attachment method for the high-temperature power electronics application, in terms of the capability of absorbing and handling the CTE-induced thermo-mechanical stress through a proper amount of plastic deformation of the bonding joint. For a better thermal management design, a TLP bonding technology with an ultrathin (<5 μm) bonding joint would be attractive to the high-power application. However, it is not an easy task to scale down the critical dimension of the Ag-In TLP bonding joint. Due to the large value of the interdiffusion coefficient in the Ag-In diffusing couple [26], the silver-indium TLP bonding couples would undergo a significant solid–solid interdiffusion and associated phase transformation before the bonding process. During the scaling down the Ag-In TLP bonding joint thickness, this would lead to a major problem known as undersupply of the liquid phase [27], which further causes the generation of interfacial voids, as shown in Fig. 2. Previously, it has been shown that the minimum critical dimension of the Ag-In TLP bonding joint using an electroplating-based process is around 30 μm [28].

Recently, an important technological breakthrough has been made in scaling down the critical dimension of the Ag-In TLP bonding joint, with a multilayer thin film structure design and a physical vapor deposition (PVD) process. This improved Ag-In TLP process was designated as the ultrathin Ag-In TLP bonding technology [29], in which the thickness of the bonding joint is only 3 μm. More importantly, the Ag-In spinodal decomposition phenomenon was firstly discovered within the ultrathin Ag-In TLP bonding joint. The spinodal modulated nanostructure is expected to share the advantages guided by the decomposition strengthening mechanism. We would expect that Ag-In spinodal nanocomposite can overcome the brittleness drawback of the intermediate IMC layer at the TLP bonding joints. If the Ag-In spinodal decomposition mechanism can be effectively controlled and utilized, in the near

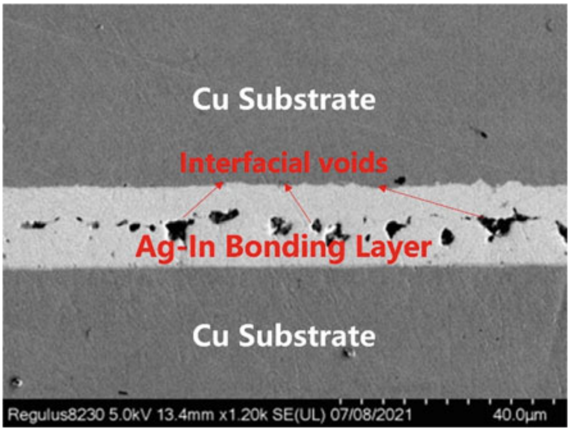


Fig. 2 A typical Ag-In transient liquid phase bonding joint with interfacial voids, caused by the undersupply of liquid phase issue

future this technology may open an entire new category of low-temperature bonding technology, namely, Spinodal Nanostructured Bonding (SNB), for the automobile applications.

1.3 Electrochemical Migration Phenomenon

Silver-based alloys are ideal materials to serve as a bonding joint in the die-attachment process for the automobile application. They usually possess good electrical conductance and thermal conductivities, are not susceptible to surface oxidation, and having a much lower raw material cost than gold (1:80). However, during the development of silver-based low-temperature bonding technologies, the ECM issue of silver-based materials must be seriously taken into account during the high-voltage. As known, silver-based materials are very prone to the impact of the electrochemical migration phenomenon [30], which can lead to the formation of Ag + ion channels under the driving force of a high electric field, eventually leading to the serious electrical short-circuit failure due to the dendritic growth of the silver whisker, as depicted in Fig. 3.

Typically, ECM process can be divided into four individual steps: 1. electrolyte layer formation; 2. dissolution of metal; 3. ion transport; 4. deposition of metal ions. More surface area would increase the capability for the metal dissolution process during the ECM. Therefore, due to the porous nature of AgNP bonding joint, the ECM process of AgNP would be easier to take place under the driving force of a high-voltage bias, compared to a pure Ag bonding joint with a compact microstructure. Thus, it should be aware of this issue when applying AgNP sintering method as interconnection in the automobile application under a high-voltage bias. In comparison, a recent research has shown that Ag-In alloys possess a good resistance to the electrochemical migration, whereas the Ag_9In_4 IMC exhibited a complete resistance

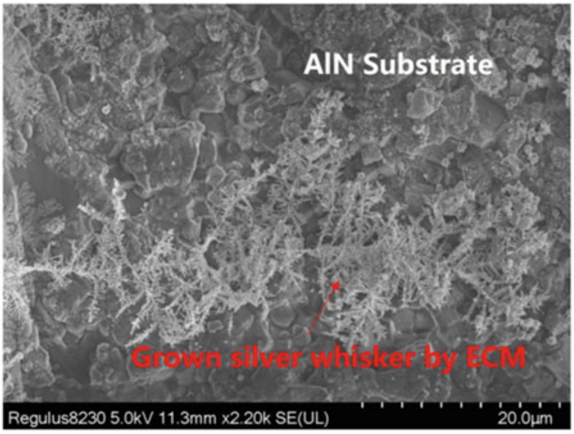

AlN Substrate
Grown silver whisker by ECM
Regulus8230 5.0kV 11.3mm x2.20k SE(UL)
20.0μm

Fig. 3 A typical grown silver whisker from electrochemical migration process under humidity and high-voltage bias, which may lead to short-circuit failure during the automobile application

to the electrochemical migration issue [31]. The anti-electrochemical migration of Ag-In alloys was still unclear, but it may originate from either from surface passivation mechanism or 4d-valance band electronic structure alternation. We can also add indium into the AgNP sintering method, in order to reduce the porous level and suppress ECM issues at the same time. In conclusion, with the utility of Ag-In alloys or other types of anti-electrochemical migration Ag alloys, it might eventually resolve and overcome the ECM reliability issue of silver-based materials for the automobile application under high-voltage bias.

2 Solder Joint Technology for Applications Under the Hood

To accomplish specific functions into car structure, there is a range of the operating temperatures for the automotive electronics [32]. As shown in Fig. 4, the maximum operating temperatures can reach values between 358.15 K (85 °C) and 478.15 K (205 °C) [33]. Therefore, for different applications in automotive electronics, there is a demand for solder joints with different melting points and assembly temperatures. We need the high, middle, and low melting point solders to work together in automotive electronics.

For high melting point solders, we have Pb5Sn and Pb5Sn2.5Ag, which are currently one of the die-attach materials in power semiconductors packages, especially for large-sized dies [34–36]. The microstructures of the high-Pb solders are generally quite stable and they do not change much during long-term aging [37, 38]. These high melting point and high-Pb alloys which combine the high-temperature stability with desirable mechanical and electrical performance for most applications are widely used in a range of applications including the packaging of high-power modules [38, 39].

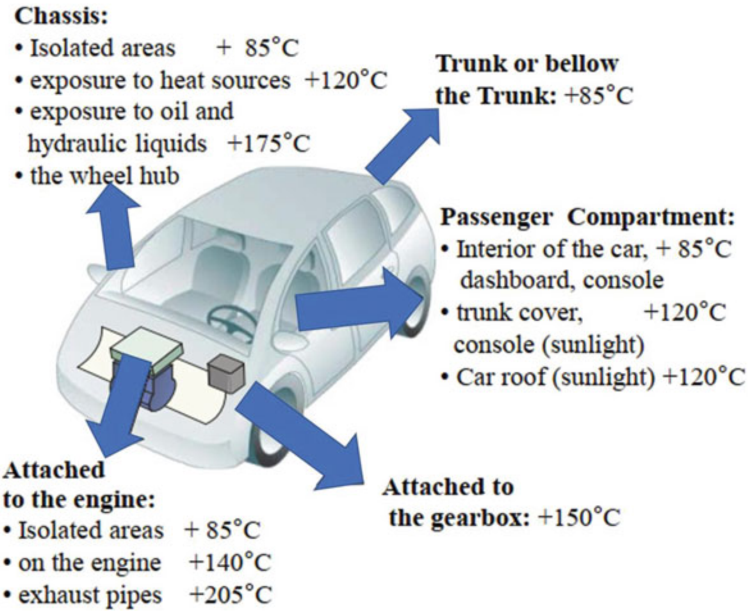

Chassis:
• Isolated areas + 85°C
• exposure to heat sources +120°C
• exposure to oil and hydraulic liquids +175°C
• the wheel hub
Trunk or bellow the Trunk: +85°C
Passenger Compartment:
• Interior of the car, + 85°C dashboard, console
• trunk cover, +120°C console (sunlight)
• Car roof (sunlight) +120°C
Attached to the engine:
• Isolated areas + 85°C
• on the engine +140°C
• exhaust pipes +205°C
Attached to the gearbox: +150°C

Fig. 4 Maximum operating temperatures for the automotive electronics (Adapted form Ref. [34])

For middle melting point solders, we usually use eutectic SnAg or SnAgCu solders. The solders have good wettability and reliability, which have been applied in a wide range of applications in electronics for decades [40, 41]. Considering the requirement of production costs reduction, there is a trend toward low melting point alloys that will enable energy savings and cost reductions during the manufacturing processes in automotive electronics production [34]. However, for low melting point solders, so far there are no ideal solders in industry. In this section, we will review several low melting point solders and discuss the methodology of applying low melting point solder paste to achieve low-temperature assembly.

2.1 Low Melting Point Solders

The major reason that we need eutectic alloys for solder is to have a single melting point. Thus, thousands of solder joints can melt and solidify at the same time, ensuring good yield and reliability performances of packaging structures. But there are not many candidates in Sn-based binary or ternary eutectic alloys, and after decades of research, those candidates have been well studied. Two of the most widely recognized low melting point solders are eutectic Sn-Bi and Sn-In solder. However, eutectic Sn-Bi solder is too brittle and Sn-In solder is too soft. Both of these two solder alloys have limitations in low-temperature assembly applications. Here we will briefly discuss the properties of these two alloys.

1. Eutectic Sn-Bi solder

The eutectic point is around 139 °C with a composition of 58 wt.% of Bi [42]. Apart from low melting point, the eutectic SnBi has a low CTE value (15 ppm/ °C) which is 1/3 lower than eutectic SnPb [43]. The low CTE value will lead to better fatigue performance on low CTE substrates, such as Alloy 42 (42Ni58Fe) [44]. On other substrates, usually, the fatigue life of eutectic SnBi is shorter than eutectic SnPb when a large shear strain, e.g., 10%, is applied; but comparable at smaller strain [45]. Eutectic SnBi shows a slower creep rate, and smaller rupture strain for the same loading stress compared with eutectic SnPb [44]. The aging of SnBi solder joints will change their mechanical properties greatly. The aging of SnBi solder joints at 80 °C can increase both the strength and the strain at failure of the joints. The ductility is low and independent of strain rate after 3 days of aging, but the ductility will increase remarkably for those joints after aging for 30 days. After annealing, the joints' shear strength decreases with strain rate. After being aged for three days, the fracture is between the interface of solder and the intermetallic compound. However, the fracture occurs in the bulk solder adjacent to the interface after being aged for 30 days [46]. The drop tests show that the performance of eutectic SnBi solder joints becomes poor as the aging time increases [47]. The shear test results present brittle fracture characteristics compared to other traditional solder joints [34].

Research concentrating on improving the mechanical properties of eutectic SnBi was reported. We summarize below in Table 1 the effects of trace amounts addition for the third or fourth element to eutectic Sn-Bi solder. The effects include the wetting properties, interfacial reactions, and mechanical performances. In this table, we evaluate the relative changes in properties of the value before and after addition [48].

2. Eutectic Sn-In solder

The composition of the eutectic of Sn-In alloy is 51.7 at. % In and 48.3 at. % Sn, with a melting point of 118 °C. Indium-base solders are soft and ductile, which is a common characteristic for them. The reason has been explained by the fast diffusion of the atoms in the solder as the solder operates at a high homologous temperature. The fast diffusion of atoms in the solder affects the mechanical properties to be soft and ductile [49].

Compared to Sn-37Pb solder, In-48Sn solder has a reported 3–4 times higher elongation [50]. Figure 5 shows the stress–strain curve of the eutectic Sn-In solder joints, where work softening rather than work hardening is exhibited as the solder joints deform [51]. It is because when the alloy operates at high homologous temperatures, the recovery and recrystallization would dominate over hardening.

Although indium solders are soft and ductile, they can have a superior lifetime when designed properly taking advantage of the creep characteristics. Shimizu et al. [50] report indium-based solder joints show much better thermal shock performance than traditional SnPb solder joints and have a longer fatigue life when the thermal

Table 1 Relevant changes in properties of eutectic Sn-Bi solder after element addition [48]

Element	Alloy composition	Melting point (^{o}C)	Wettability	IMC growth rate	Tensile strength (%)	Elongation (%)	Shear strength (MPa)
Cu	Sn-58Bi-1Cu		Slightly	Slightly			
	Sn-58Bi-0.5Cu		Slightly		Slightly	+ 140	
	Sn-40Bi-0.1Cu	132.2			+ 12.6		
	Sn-40Bi-0.1Cu-2Zn	136.3			+ 21.9		
In	Sn-58Bi-2In	129.7		Slightly	+ 6.5		+ 8.0
	Sn-58Bi-2.5In				Slightly	+ 104.5	
Al	Sn-58Bi-2Al	142	Worse	Slightly			
Zn	Sn-xBi-3Zn (x = 37,39,41,43)	136.7–138.0	Depends	Depends			
	Sn-38Bi-xZn (x = 0,2,3,4)		Depends				
	Sn-58Bi-0.7Zn	136.3		Depends	+ 7		
	Sn-58Bi-0.5Zn				Weaken	Slightly	
Ag	Sn-58Bi-1Ag	Little effect	Better	Decrease			
	Sn-58Bi-0.5Ag	Little effect	Better	Increase	+ 13.2	+ 80	+ 7 after aging
	Sn-58Bi-2Ag	139.0			+ 12.1		+ 6.7
	Sn-58Bi-2In-2Ag	133.6			+ 8.1		+ 12.9
Sb	Sn-58Bi-2Sb	147	Better	Increase			-15.2
	Sn-58Bi-0.5Sb				Slightly	+ 320	

(continued)

Table 1 (continued)

Element	Alloy composition	Melting point (°C)	Wettability	IMC growth rate	Tensile strength (%)	Elongation (%)	Shear strength (MPa)
RE	Sn-58Bi-0.1RE	Little effect	Superior	Increase			+ 14 after aging
	Sn-58Bi-0.5La	137.8					Decrease
Co	Sn-58Bi-0.5Co	Slightly		Increase			
Ga	Sn-58Bi-xGa	Little effect		Decrease			

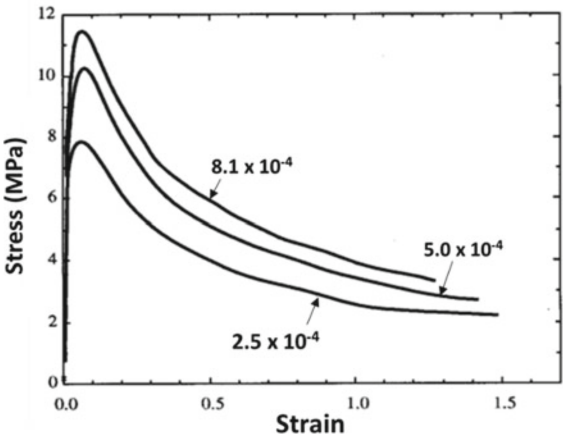


Fig. 5 The strain–stress correlation at 40 °C for the eutectic Sn-In joint (Adapted from Ref. [51])

shock temperature is between liquid nitrogen and room temperatures, as shown in Fig. 6. However, Seyyedi et al. [52] reports under the thermal cycling condition of −10 °C to 70 °C range, eutectic Sn-In solder shows a shorter fatigue lifetime than eutectic Sn-Bi and Sn–Pb solder.

The tensile strength of pure In is about the same as other indium-based solders but is less than half of that in pure Sn. [53]. The values of Youngs modulus, E, shear modulus, μ, and bulk modulus, B, of eutectic Sn-In alloy, are measured to be 8.123 GPa, 2.899 Gpa, and 13.675 Gpa, respectively. While Sn95Ag5 has measured E, μ, and B, as 31.0275 Gpa, 11.42 Gpa, and 36.289 Gpa, respectively [54]. Similar to eutectic SnBi solder, eutectic SnIn solder is sensitive to the shear speed in shear tests. [55].

Research on improving the machinal properties of eutectic Sn-In solder by minor alloy addition has been carried out. The studies on the trace addition to eutectic Sn-In solder are not as much as to eutectic Sn-Bi solder. We summarize the effect of Au, Ag, and Zn addition to eutectic Sn-In solder in Table 2. Again, the data is evaluated as the relative changes compared with eutectic Sn-In solder [48].

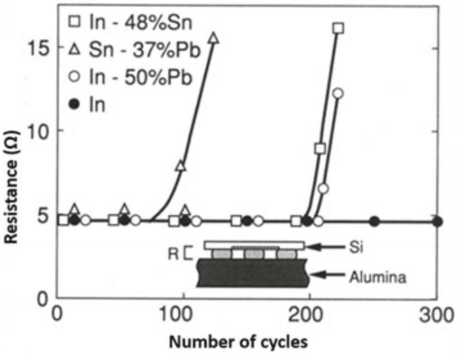


Fig. 6 Thermal shock test for flip-chip solder joints between liquid nitrogen and room temperatures (Adapted from Ref. [50])

Table 2 Properties of eutectic Sn-In solder by trace amount of third element addition [48]

Element	Alloy composition	Melting point (°C)	IMC growth rate	Tensile strength (%)	Elongation (%)
Au	In-48Sn-20Au	152			
	In-48Sn-20Au-5Ag	133			
Ag	In-48Sn-0.5Ag	113 ~ 117		Increase	
	In-48Sn-1.5Ag				+100
Zn	In-44Sn-6Zn	108	Increase	No change	+100
	In-20Sn-2Zn		Decrease		

Although eutectic Sn-Bi solder is too brittle and Sn-In solder is too soft, and both of them are not ideal in applications, an Sn-rich In-containing solder paste—Durafuse™ LT was designed by Indium Cooperation [56]. This new solder paste may imply the future trend of developing a low melting point solder methodology. The solder paste is reflowable at 200 °C and above, and the solder paste has excellent drop-shock performance with acceptable thermal fatigue behavior. Durafuse™ LT is a mixed solder powder paste, in which the In-containing powder will melt first around 118 °C to spread and wet. The Sn-rich powder (melting point > 217 °C) dissolves into the molten solder during reflow. The solder joint has a melting temperature of around 189 °C after the dissolution accomplishes. The drop performance of this mixed solder joint is much better than that of BiSnAg eutectic solder, and the failure number is around two-orders-of magnitudes higher. The drop-shock performance of the mixed solder is even better than SAC305 when the reflow profile is optimized. This low melting point solder is promising to be applied in portable electronics.

2.2 *Low-Temperature Assembly*

Although both eutectic Sn-Bi or Sn-In alloys are not ideal to be applied inside the chip, we can try to make reliable solder joints by the hybrid bonding of SAC305 solder ball and low melting temperature solder paste [57, 58]. Figure 7 is a schematic diagram to illustrate the hybrid bonding. With the hybrid bonding technology, it's possible we can achieve low-temperature assembly, while the uniformed solder joints may show better yield and reliability performance. Also, the solder joints can have a higher melting point and will be able to survive the working temperature for applications under the hood. Figure 8 shows the cross-section images of mixed SAC305 + SnBi solder paste solder joints [58].

The use of BiSn-based solder paste can substantially improve the BGA solder joint yield while achieving low-temperature bonding [58]. With SAC305 solder balls, the amount of Bi and In can be diluted in the hybrid bonding. In this way, the brittle nature of Bi and the soft nature of In can have a smaller effect on the whole joints. The characteristic life of the hybrid bumps is better than eutectic Sn-Bi solder joints

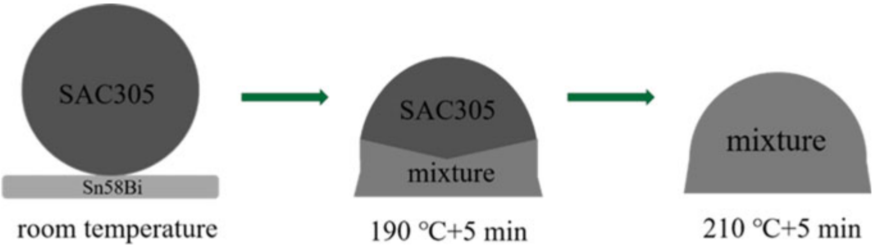


Fig. 7 Schematic diagram of the bonding by SAC305 solder ball and Sn58Bi solder paste

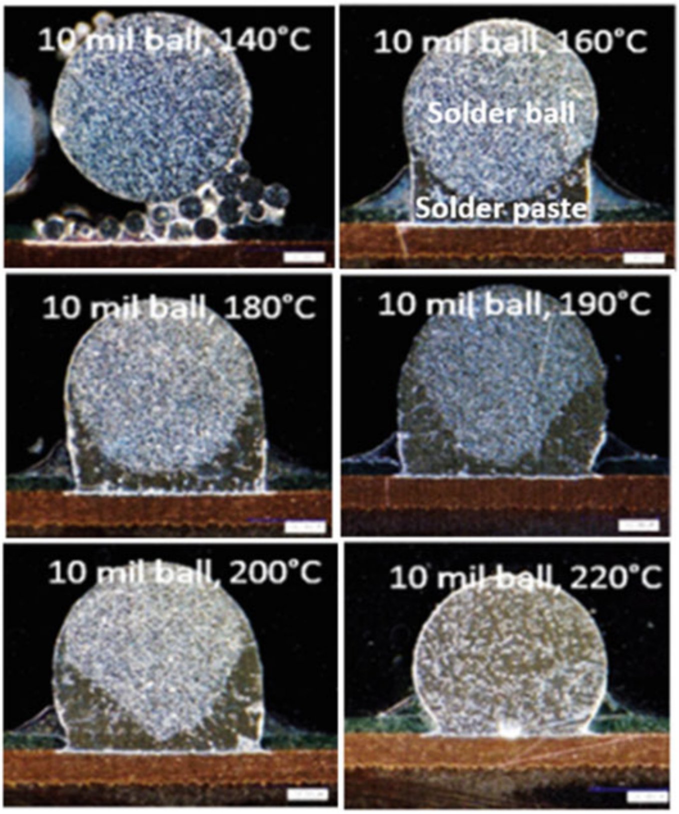


Fig. 8 Cross-section images of mixed SAC305 + SnBi solder paste solder joints after melting at the indicated temperatures (Adapted from Ref. [58])

under mechanical shock or drop tests. But is still lower than full SAC solder joints. Low-temperature soldering is effective in energy cost savings and can reduce carbon footprint significantly. It is estimated that the cost savings of the low-temperature operation can be $8749/oven/year, and the approximate carbon footprint reduction is 57.2 metric tons of CO_2 per oven per year. In this way, with the development of new low melting solder paste, it's promising we can finally develop cost-saving, yield, and reliability-enhanced low-temperature assembly technology.

3 Introduction for Low-Temperature Cu to Cu Direct Bonding

The integration intensity in large-scale integrate circuits (ICs) increases significantly and the feature size of transistor has approached the physical limit with the development of microelectronics technology in recent years. As a result, three-dimensional integrated circuits (3D ICs) packaging technology has been regarded as a prominent alternative technology due to its higher packing densities, smaller featured sizes, and better performances [59, 60]. In 3D IC technology, through-silicon-vias (TSV) and Pb-free-based solders are used to vertically integrate chips in one package. During the bonding process, the Pb-free-based solder would react with under bump metallization (UBM) to form intermetallic compounds (IMCs), which are brittle in nature. As the solder height decreases down to several micrometers for higher integration density, the volume fraction of the intermetallic compounds (IMCs) in microbumps increases significantly, causing higher electrical resistance and new reliability issues in solder-based interconnects [61, 62]. Furthermore, with the development of the emerging industry, such as artificial intelligence, electric vehicles, internet of things, and 5G communications, the application environment become harsher due to high power, high switching frequency, and high-temperature applications. Under such environments, solder-based interconnect may be no longer suitable and many new reliability issues have emerged for solder-based interconnect.

In order to eliminate the solder-based interconnect from the bonding interfaces, metal-to-metal direct bonding has been regarded as a promising technique to achieve heterogeneous integrations. Cu-Cu direct bonding has been regarded as most promising bonding method to achieve vertical integration because its superior electrical and thermal conductivity, high bonding strength, better electromigration resistance, and compatibility to current packaging fabrication [63]. However, Cu is easily oxidized and thus high bonding temperature larger than 400 °C is required to enhance interdiffusion of Cu [64]. High bonding temperature may degrade the performance of IC devices and cause some reliability and thermal-related issues, such as bonding alignment, wafer warpage, and compatibility with back-end-of-line process. As a result, the development of low-temperature Cu-Cu bonding is critical.

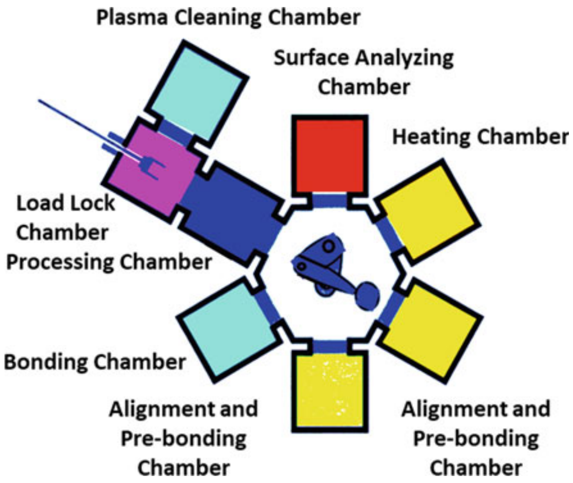

Plasma Cleaning Chamber
Surface Analyzing Chamber
Heating Chamber
Load Lock Chamber
Processing Chamber
Bonding Chamber
Alignment and Pre-bonding Chamber
Alignment and Pre-bonding Chamber

Fig. 9 Schematic view of surface-activated bonding machine (Adapted form Ref. [65])

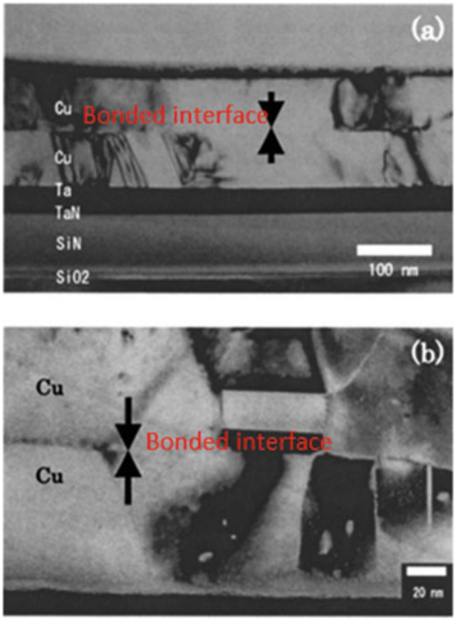
3.1 Cu-Cu Bonding by Surface-Activated Bonding Process

Surface-activated bonding (SAB) method was proposed to achieve Cu-Cu direct bonding by Suga et al. For SAB, energetic ions or fast atom beams (e.g., Ar-FAB) were used to remove the surface oxide in ultrahigh vacuum environment to activate the bonding surface [65]. Figure 9 shows the schematic setup for SAB process. Prior to the bonding process, the wafers are transferred into the surface-activated bonding machine and the surface activation process is conducted by an ion source with low energy (40–100 eV, acceleration voltage: 80 V) and high current (2.92 A). Continuous ion bombardment on the wafer surfaces for 60 s is required in order to remove the oxide and contaminations. During the surface activation process, the substrate was rotated to ensure the uniformity of activation through the whole surface area. After the surface activation process, two wafers are transferred into the preliminary bonding chamber and brought into contact to provide an initial bonding under a load of 50 kgf. Then, the pre-bonded sample is bonded under 1000 kgf load in the bonding chamber at room temperature and ultrahigh vacuum environment (~10^{-8} Torr). Figure 10 shows TEM image of the bonded Cu-Cu interface using SAB process, demonstrating good bonding interfaces. Although SAB process can provide high bonding strength and room bonding temperature, the requirement of the ultrahigh vacuum environment limits its industry application due to its high cost and low throughput.

3.2 Cu-Cu Bonding by Chemical Pretreatment

To enable Cu to Cu bonding at low temperature, effectively removing the oxide of Cu is necessary. Different kinds of chemical pretreatments, including hydrochloric acid [66, 67], citric acid [68], sulfuric acid [69], and acetic acid [70] have been proposed

Fig. 10 **a** Low magnification and **b** high-resolution TEM images of the bonded Cu-Cu sample (Adapted form Ref. [65])



to be used for the removal of the oxides and citric acid has been found to be the most effective one to decrease the bonding temperature to 300 °C. [71] In addition, through using different combination of different acidic solutions, such as HF/H_2SO4 and HCl/H_2O [72], the bonding temperature can be further reduced to 250 °C. Although chemical pretreatment is an easy, low-cost, and industrial-compatible process, it will simultaneously cause rougher surfaces, which is detrimental for bonding process.

3.3 Cu-Cu Bonding by Thermal Compressive Bonding

In addition to temperature, pressure is usually applied to facilitate metal-to-metal directing bonding, which is called thermal compressive bonding (TCB). During bonding process, the pressure can help to break the residual oxide as well. With the introduction of a compressive pressure, the diffusivity of Cu atoms during the bonding process would increase and the diffusivity under stress conditions can be expressed as follows [73]:

$$\ln\left(\frac{D}{D_0}\right) = (\sigma - \sigma_0)\frac{V^*}{kT} \tag{1}$$

where D and D_0 is, respectively the diffusivity of Cu atoms during and before bonding process, V^* is the atomic volume of Cu, σ_0 and σ, respectively represents the stresses before and during the bonding process, k is the Boltzmann constant and T is the bonding temperature. We can find that the diffusivity of Cu would significantly increase when a large compressive stress is applied, which could effectively shorten bonding temperature or time.

The mechanism of thermal compressive bonding is mainly caused by atomic surface diffusion along stress gradient at the bonded interfaces, which is similar to the Coble creep [74]. The bonding process is schematically presented in Fig. 11. Firstly, the native oxides on the surface of Cu is broken due to the presence of applied pressure. Secondly, when two interfaces are contacted during thermo-compression bonding process, surface roughness of Cu would cause the formation of boding interface between strained regions (contacted interfaces) and unstrained regions (formed cavities), as shown in Fig. 11a. Due to the stress difference between contacted interfaces (high compressive stress) and formed cavities (nearly stress-free state), a stress gradient is established between contacted regions and cavity regions at the bonded interface. This stress gradient would trigger an atomic flux dominated by surface diffusion of Cu. Once Cu atoms diffuse from the contacted region to the cavities to release large compressive stress at the contacted interfaces, the cavities would be filled and two interfaces finally merge together. Consequently, a new bonded interface is formed, as depicted in Fig. 11b.

A diffusion field should be used to analyze the kinetic during thermal compressive bonding process; however, the curved surface of cavity yields a complicated process. To simplify, the surface atomic flux during TCB process can be represented as following by using the Nabarro-Herring (or Coble) model of creep [74]

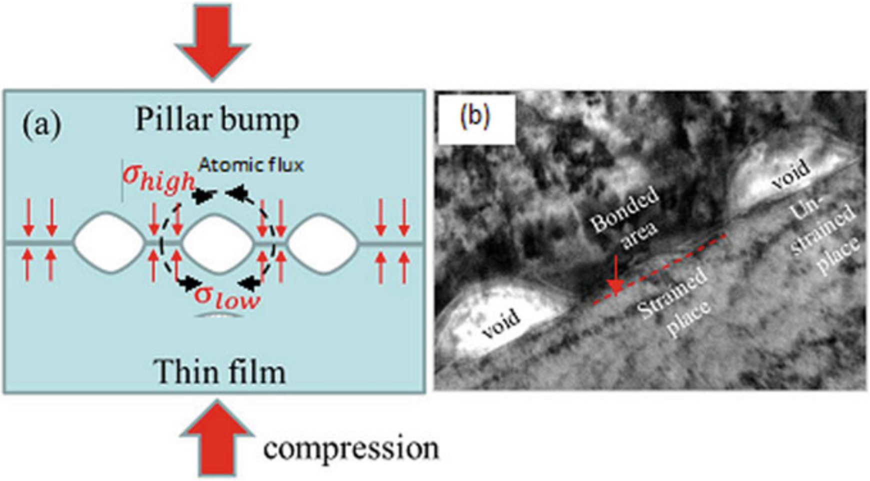


Fig. 11 Schematic diagram illustrating metal to metal bonding mechanism. **a** atoms diffuse along the contacted interface due to stress gradient, and **b** TEM image of the bonded Cu-Cu interface (Adapted form Ref. [75])

$$J_s = -C_s \frac{D_s}{kT}\left(-\frac{\Delta\mu}{\Delta x}\right) = C_s \frac{D_s}{kT}\frac{\sigma\Omega}{r} = \frac{D_s}{kT}\frac{\sigma}{r} \tag{2}$$

where C_s is the surface atomic concentration, D_s is the surface diffusivity, σ is the applied stress, k is the Boltzmann constant and T is the temperature, Ω is the atomic or vacancy volume, and r is the radius of the cavity and is regarded as the distance of surface diffusion, as shown in Fig. 11(b). If the unit of surface atomic flux is assumed to be #atoms/cm^2—sec instead of #atoms/cm—sec, then C_s can be taken as the number of atoms in an atomic layer on the surface, i.e., $C_s = 1/\sigma$ for a pure element.

It is noteworthy that the surface roughness of Cu plays a critical role in stress-induced migration during thermal compressive bonding process; the small surface roughness would provide higher surface atomic flux and form smaller cavities at the bonding interface. Consequently, it requires less time for Cu atoms to fill in the cavities. Thus, a chemical mechanical polishing (CMP) process is usually conducted before thermal compressive bonding.

3.4 Low-Temperature Cu-Cu Bonding by (111) Nanotwinned Structure

Based on the above-mentioned bonding mechanism, the surface diffusivity significantly affects the bonding temperature, time and quality since higher surface diffusivity would facilitate bonding process. Table 3 lists the surface diffusivity of different crystal orientations at different temperatures; we can find that (111) plane exhibits the fastest diffusivity. Liu et al. [74, 76] have proposed to use highly (111) oriented nanotwinned Cu structure for direct bonding because surface diffusivity of Cu(111) is 3–4 orders higher than other planes and they successfully reduce the bonding temperature to 200 °C. In addition, nanotwins are introduced in the electroplated Cu because of their unique properties, including low electrical resistivity, high mechanical property, and high thermal stability comparing to the nanocrystalline structure [78–84]. Figure 12 shows the results of plan-view Electron Backscatter Diffraction (EBSD) orientation image map for electroplated nanotwinned Cu films; almost 100% of (111)-oriented surface grains are observed for nanotwinned Cu films.

Figure 13 shows the results for nt-Cu pillar with 30 µm in diameter bonded with nt-Cu film by thermal compression bonding at 40.6 MPa for 20 min at various temperature gradients in N$_2$ ambient. During bonding process, recrystallization and grain growth with (100) and (110) orientations were observed. When temperature is higher than 350 ∘C, large grains with (211) and (100) orientations grew from the pillar bump to consume the (111) columnar grains in the Cu film. For Cu joints bonded at 200 °C/100 °C, the shear strength could reach 73.3 MPa, being larger than traditional SnAg solder joint.

Table 3 Calculated surface diffusivity of Cu and Ag on different planes at various temperatures. [76, 77]

Element	Temperature (°C)	$D_{\text{surf.}}$ (cm^2/s)		
		(111)	(100)	(110)
Cu	150	6.85×10^{-6}	2.15×10^{-10}	6.61×10^{-12}
	180	8.37×10^{-6}	6.31×10^{-10}	2.64×10^{-11}
	190	8.89×10^{-6}	8.47×10^{-10}	4.02×10^{-11}
	200	9.42×10^{-6}	1.19×10^{-9}	6.02×10^{-11}
	250	1.22×10^{-5}	4.74×10^{-9}	3.56×10^{-10}
	300	1.51×10^{-5}	1.48×10^{-8}	1.55×10^{-9}
Ag	180	1.08×10^{-5}	3.62×10^{-9}	7.13×10^{-10}
	190	1.14×10^{-5}	4.83×10^{-9}	9.88×10^{-10}
	200	1.21×10^{-5}	6.36×10^{-9}	1.35×10^{-9}

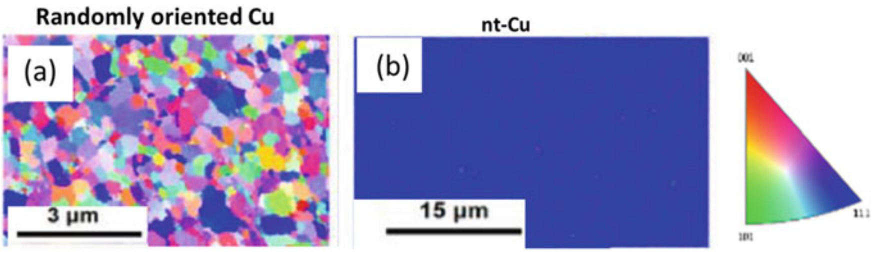


Fig. 12 EBSD orientation map on the surface of (**a**) randomly oriented Cu film, and (**b**) (111) nt-Cu film (Adapted form Ref. [74])

3.5 *Low-Temperature Cu to Cu Bonding with Ag Passivation Under Atmosphere*

Although the above-mentioned methods are used to reduce the bonding temperature, the low-oxygen environment is usually mandatory to enable Cu-Cu direct bonding in order to prevent the oxidation of Cu. To achieve the Cu-Cu bonding process without vacuum environment, Ag is regarded as a good passivation layer because Ag possesses prominent electrical and thermal conductivity and excellent oxidation resistance. Theoretical prediction indicates that the Ag_2O is possible to dissociation to solid Ag and O_2 at temperature above 145 °C under air atmosphere [86, 87]. In addition, based on in-situ thermogravimetric analysis (TGA) [77], no significant weight change of the pre-oxidized Ag samples at 175 °C was found during TGA test, inferring that Ag_2O may not effectively dissociate at 175 °C. However, when ambient temperature increases to 180 °C and 185 °C, the weight of the oxidized Ag samples decreases with time due to the dissociation of Ag oxides. Thus, an oxide-less

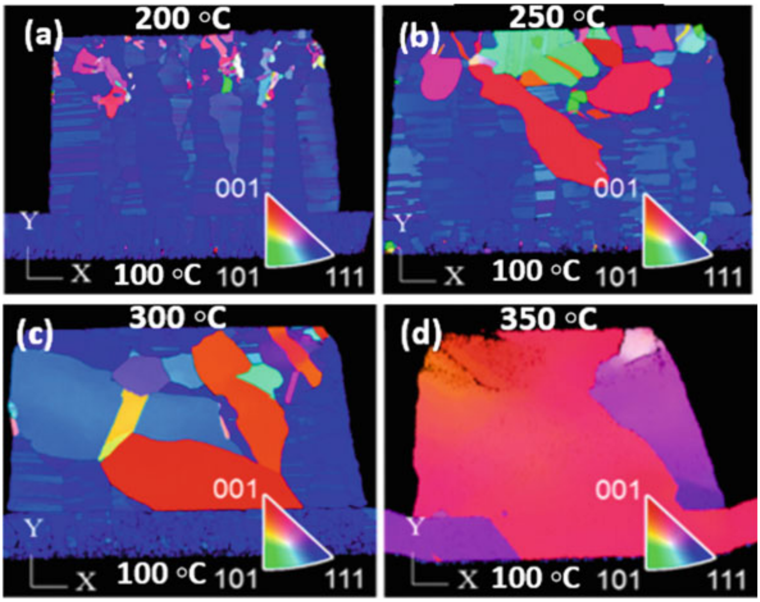


Fig. 13 Cross-sectional electron backscatter diffraction (EBSD) images on samples bonded for upper interface at **a** 200 °C, **b** 250 °C, **c** 300 °C, and **d** 350 °C. The temperature at lower interface is fixed at 100 °C (Adapted form Ref. [85])

surface is formed with active Ag atoms on the surface, which will facilitate bonding process, as shown in Fig. 14.

Figure 15 shows cross-sectional FIB micrograph, EBSD, and surface morphology of as-deposited films. The Ag thin film possesses microstructure with nearly equiaxed grains and some columnar structure with high-density nanotwins. In addition, only ~46% of the surface grains are (111)-oriented and the root mean square (RMS) surface roughness of the Ag thin film is 10.5 nm. Figure 16 shows bonding results of sputtered Ag films on the electroplated Cu. The bonding process is conducted at temperature ranging from 180 °C to 200 °C for 1 or 3 min in air atmosphere

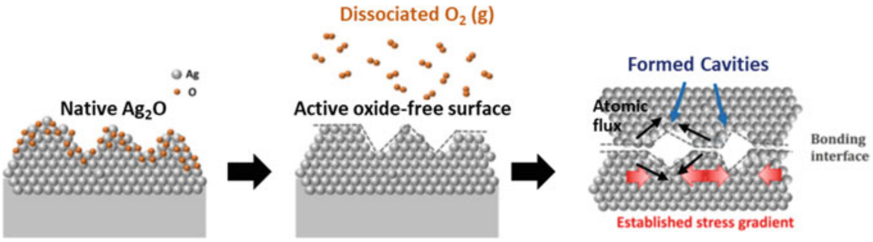


Fig. 14 Schematic diagram illustrating Ag–Ag direct bonding mechanism (Adapted form Ref. [77])

under compressive pressure of 30 MPa. Figure 16a represents the interface of Ag thin films bonded at 180 °C for 1 min in air atmosphere. We can find well-bonded interfaces in most regions and only a few small voids remained at the interface. Figure 16b shows cross-sectional microstructure of samples after bonding at 200 °C for 3 min under applied pressure of 30 MPa and further annealing at 200 °C for 30 min in air atmosphere. No voids can be found at the bonded interface and good bonding is achieved. Their corresponding TEM analysis is shown in Fig. 16(c). A zig-zag interface is formed and a lots of triple junctions can be found at the interface, suggesting efficient interdiffusion between two surfaces of Ag passivation. The shear strengths for samples are all over 15 MPa and the highest bonding strength can reach 58.92 MPa in average when samples were bonded at 200 °C for 3 min under applied pressure of 30 MPa and annealed for 30 min in air atmosphere. The above-mentioned results suggest using Ag as a passivation layer can successfully bond Cu in air atmosphere in short time at low temperature.

Figure 17 shows the results of resistance of test vehicles for 2236 bumps in total with the diameter of 40 µm after bonding and temperature cycling test (TCT). The bonding process is conducted at 200 °C for 3 min in air atmosphere with additional annealing process under 200 °C for 30 min and TCT was conducted from −55 °C to 125 °C for 1000 cycles with a dwelling time of 5 min for each cycle. The results indicate that the well-bonded structure is formed given that the electrical resistance of test

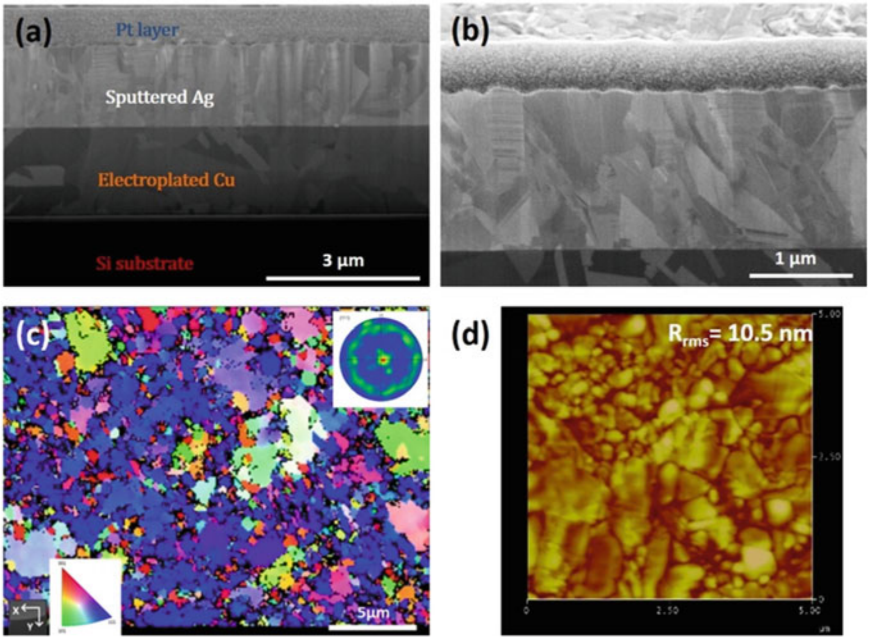


Fig. 15 **a**, **b** FIB cross-sectional images, **c** plan-view electron backscatter diffraction orientation image map, and **d** surface roughness measured by AFM for Ag passivation films on electroplated Cu

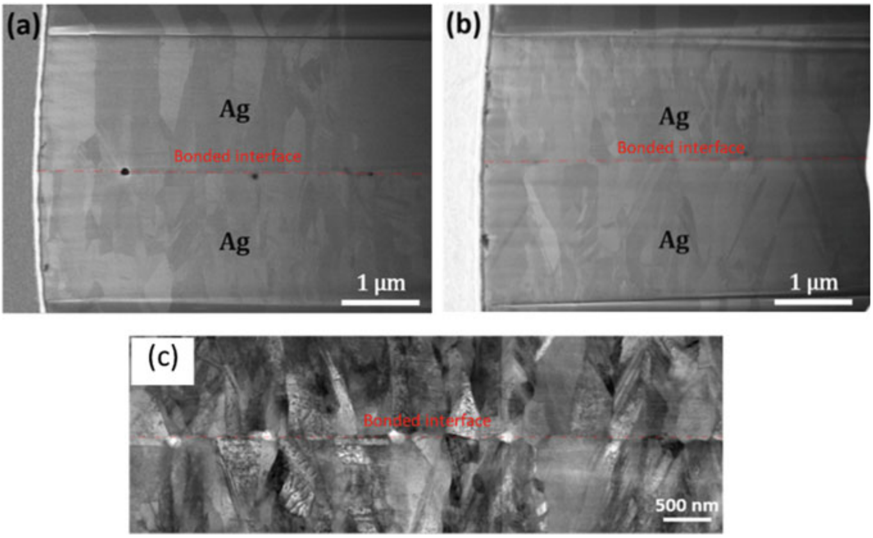


Fig. 16 FIB images of cross-sectional samples for **a** bonded at 180 °C for 1 min under pressure of 30 MPa in air atmosphere; **b** bonded at 200 °C for 3 min under pressure of 30 MPa in air atmosphere and then subject to further annealing for 30 min. **c** TEM bright filed image for sampled bonded at 200 °C for 3 min under pressure of 30 MPa in air atmosphere and then subject to further annealing for 30 min. The original bonding interface is indicated as red dash line (Adapted form Ref. [77])

vehicle increases linearly with increasing numbers of measured rows. Furthermore, the contact resistances of the test vehicle do not significantly change and decrease slightly after TCT because high temperature during TCT facilitates interdiffusion of Ag atoms and it further improved the bonded interface. The above-mentioned results demonstrate that using Ag as a capping layer for Cu to Cu direct bonding can be conducted at low temperature of 180 °C in air atmosphere and achieve high bonding strength and high reliability after TCT.

3.6 Hybrid Bonding

To enable high-density interconnect, short throughput, and high yield and reliability, hybrid bonding is regarded as an emerging approach for 3D integration. Hybrid bonding is the combination of metal-to-metal bonding and wafer-level bonding with adhesives and dielectrics. The advantage of using hybrid bonding is that dielectric or adhesives can act not only the bonding material but also underfill; thus the bonding strength can be effectively enhanced [88]. In addition, given that metal and dielectric or adhesives are bonded simultaneously, the process flow is simplified and process time is saved. Cu/SiO_2 and Cu/adhesive are acknowledged as the most promising

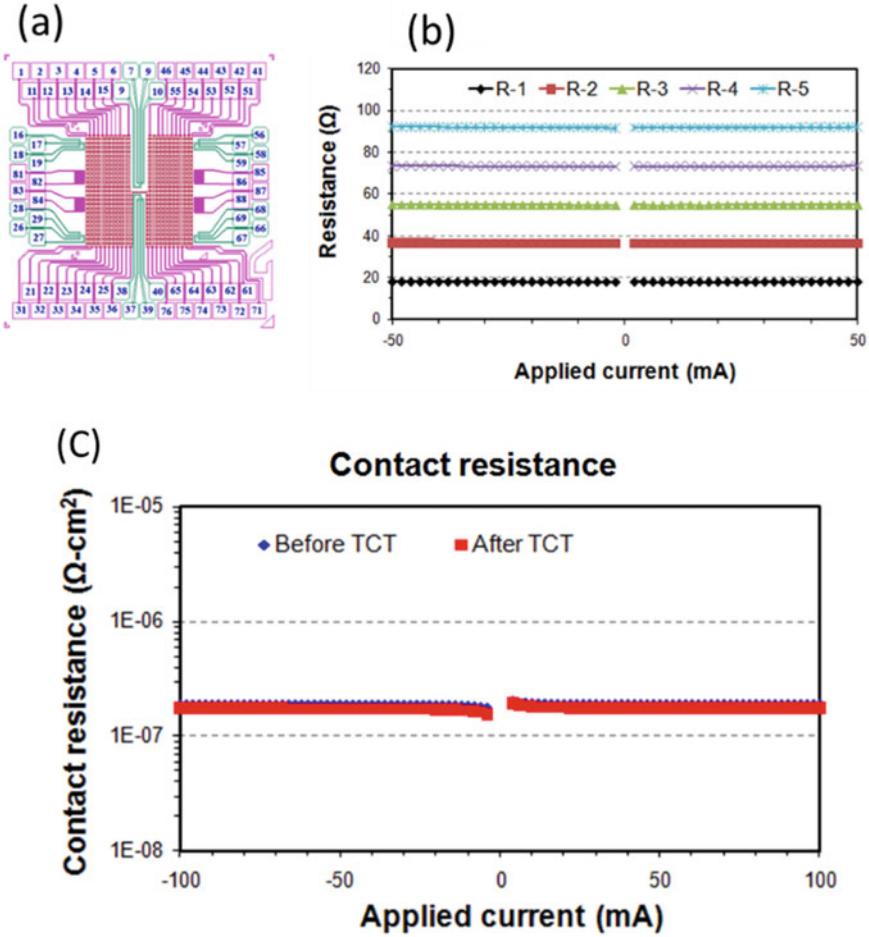


Fig. 17 **a** Test vehicle design, and **b** resistance measurements for different numbers of rows on as-bonded test vehicle, **c** resistance measurement for Kelvin structure in test vehicle before and after TCT from −55 °C to 125 °C for 1000 cycles (Adapted form Ref. [77])

hybrid bonding materials and we will review their corresponding bonding mechanism in the following.

To perform Cu/SiO_2 hybrid bonding, two-step bonding sequence is typically adopted. As shown in Fig. 18a, the SiO_2-SiO_2 bonding is conducted at room temperature and post-bonding annealing is performed at 200–400 °C for Cu-Cu bonding. Before bonding, hydrophilic surface modification of SiO_2 is typically required to enable SiO_2- SiO_2 bonding. In addition, post-bonding annealing is necessary to increase the quality of SiO_2- SiO_2 and Cu-Cu bonding. Due to thermal expansion of Cu, Cu expands at high temperatures. Consequently, the gap between Cu surfaces caused by CMP dishing can be closed during annealing. However, the main problem of this bonding technique is that Cu thermal expansion may induce high tensile stress in as-bonded SiO_2- SiO_2 interface. Thus, as-bonded SiO_2- SiO_2 interface should possess high mechanical strength to endure stress caused by Cu thermal expansion. In order to further improve the bonding quality, external compression is typically adopted on both the Cu-Cu and SiO_2- SiO_2 interfaces during bonding at high temperature, as shown in Fig. 18b. Several bonding methods are used for Cu/SiO_2 hybrid bonding, including plasma activation bonding [89, 90], direct bond interconnect [91, 92], special CMP treatment of Leti-CEA [93–95], vapor-assisted SAB, etc. [96]. Among them, direct bond interconnect (DBI) has been applied by Sony for 3D stacked back-illuminated image sensors (IMX260) used in Samsung Galaxy S7 Edge in 2016. In this technique, wafers are bonding at room temperature without external pressure after hydrophilic surface modification of SiO_2, then followed by post-bonding annealing at 125–400 °C to facilitate Cu-Cu bonding. As a result, high strength and low resistance of bonded interfaces is achieved.

For Cu/adhesive hybrid bonding, thermosetting polymer adhesive are usually used; for example, benzocyclobutene (BCB) and polyimide (PI) are two popular adhesive materials. In addition, to enable Cu/adhesive hybrid bonding, two-step bonding sequence is typically adopted with the help of pressure. For "adhesive-first" bonding technique, as shown in Fig. 19, adhesive is first bonded under thermo-compression boning and cured at a relatively lower temperature, and then the Cu-Cu

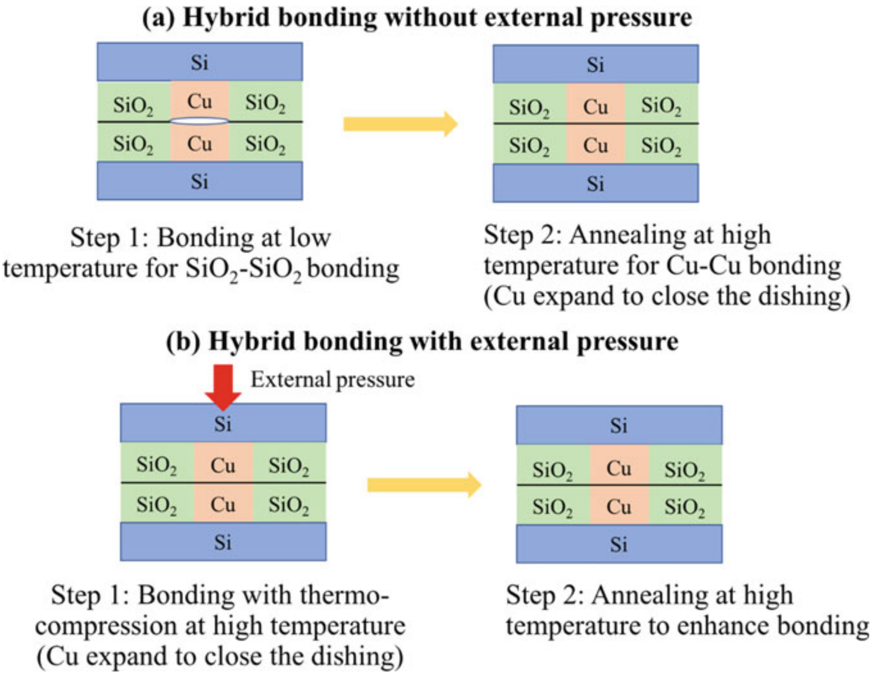

(a) Hybrid bonding without external pressure
Si
SiO2
Cu
SiO2
SiO2
Cu
SiO2
Si
Step 1: Bonding at low temperature for SiO2-SiO2 bonding
Step 2: Annealing at high temperature for Cu-Cu bonding (Cu expand to close the dishing)
(b) Hybrid bonding with external pressure
External pressure
Step 1: Bonding with thermo-compression at high temperature (Cu expand to close the dishing)
Step 2: Annealing at high temperature to enhance bonding

Fig. 18 Cu/SiO_2 hybrid bonding **a** without, and **b** with external pressure

thermo-compression bonding is performed at higher temperature [97, 98]. The main problem of "adhesive-first" bonding technique is high bonding temperature of Cu may damage the adhesive if it is not fully cured. Thus, the requirement of adhesive which needs to have high thermal stability to endure severe metal bonding conditions limits the choice of adhesive materials. In addition, high bonding temperature during Cu to Cu bonding may cause higher thermal stress and in turn results in some reliability issues. To resolve this issue, a "Cu-first" hybrid bonding technique is developed. Before longer duration of adhesive curing step, the Cu-Cu bonding is firstly performed at temperature lower than bonding/curing temperature of adhesive with shorter duration (e.g., within 10 min). Thus, the development of low-temperature of Cu-Cu bonding on the "Cu-first" hybrid bonding technique is critical. As mentioned earlier, the surface activation method or introduction of highly (111) nanotwinned Cu can effectively lower the bonding temperature; however, the study related to surface activation method on Cu/adhesive for bonding temperature lower than 250 °C is still rare. It has recently been found that using H-containing HCOOH vapor treatment could reduce the bonding temperature to 200 °C under thermo-compression bonding for 5 min to achieve strong Cu-Cu strength [99]. More research related to achieve low-temperature Cu/adhesive is required.

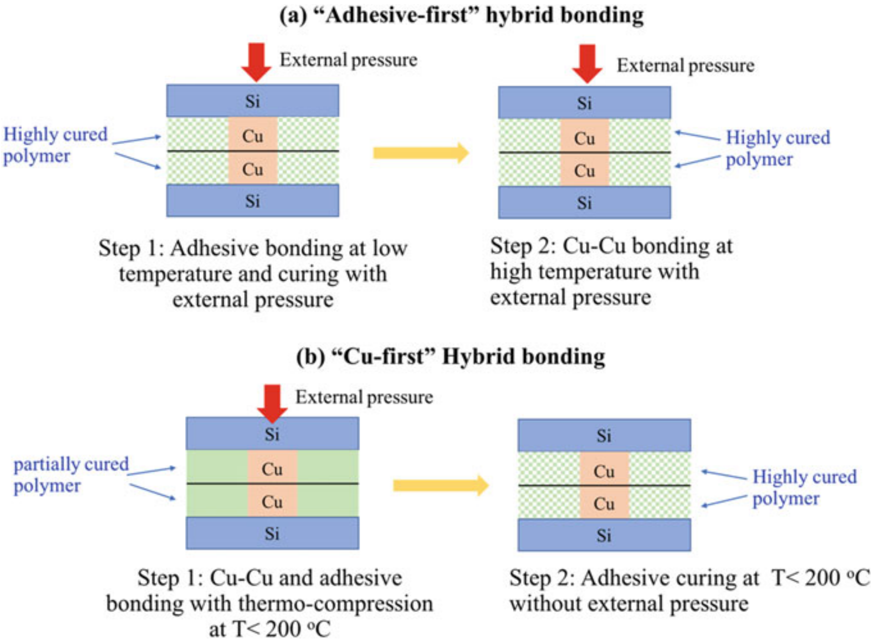

(a) "Adhesive-first" hybrid bonding
External pressure
Si
Cu
Cu
Si
Highly cured polymer
Step 1: Adhesive bonding at low temperature and curing with external pressure
External pressure
Si
Cu
Cu
Si
Highly cured polymer
Step 2: Cu-Cu bonding at high temperature with external pressure
(b) "Cu-first" Hybrid bonding
External pressure
Si
Cu
Cu
Si
partially cured polymer
Step 1: Cu-Cu and adhesive bonding with thermo-compression at T< 200 °C
Si
Cu
Cu
Si
Highly cured polymer
Step 2: Adhesive curing at T< 200 °C without external pressure

Fig. 19 **a** "Adhesive-first" process for Cu/adhesive hybrid bonding, and **b** "Cu-first" process for Cu/adhesive hybrid bonding (Adapted form Ref. [100])

Summary

In this chapter, we review the newly developed LP bonding methods and the SNB bonding method, solder joint technology, Cu-Cu direct bonding and hybrid bonding as interconnects using in the high-power devices. In the near future, more studies need be conducted quantitively to further evaluate performance under high voltage or the long-term reliability of high-power devices using newly developed TLP bonding methods or the SNB bonding method. We need to pay a special attention to the anti-electrochemical migration performance when developing a new die-attachment method for devices under the hood in the automobile application. For solder joint technology, the development of new low melting temperature solder paste (e.g., hybrid SnBi) is important. Hybrid bonding is more complicated than Cu-Cu bonding due to the requirement of simultaneously. Further research is needed to increase the bonding strength, lower bonding temperature, and control Cu dishing caused by planarization process. With the advent of various low-temperature bonding technologies, it also could facilitate the system-level heterogeneous integration of the high-power devices into a more complex power electronics system. Eventually, the viability of those emerging low-temperature bonding methods will be fully tested in the industrial production environment in terms of cost, throughput and production yield.

Acknowledgements The editors would like to thank Dr. Xin Yan of Intel Corporation for his dedicated review of this chapter.

References

1. H. Lee, V. Smet, and R. Tummala, “A Review of SiC Power Module Packaging Technologies: Challenges, Advances, and Emerging Issues”, IEEE Journal of Emerging and Selected Topics in Power Electronics, vol. 8, No. 1, pp. 239-255, 2019.
2. X. Zhong, L. Chen, B. Medgyes, Z. Zhang, S. Gao, and L. Jakab, “Electrochemical migration of Sn and Sn solder alloys: a review”, RSC Advances, vol. 7, No. 45, pp. 28186-28206, 2017.
3. T. Bakhishev, V. Subramanian, “Investigation of gold nanoparticle inks for low-temperature lead-free packaging technology”, Journal of Electronic Materials, vol. 38, No. 12, pp. 2720-2725, 2009.
4. P. Peng, A. Hu, A. P. Gerlich, G. Zou, L. Liu, Y. N. Zhou, “Joining of silver nanomaterials at low temperatures: processes, properties, and applications”, ACS Applied Materials & Interfaces, vol. 7, No. 23, pp. 12597-12618, 2015.
5. J. Yan, G. Zou, A. Hu, Y. N. Zhou, “Preparation of PVP coated Cu NPs and the application for low-temperature bonding”, Journal of Materials Chemistry, vol. 21, No. 40, pp. 15981-15986, 2011.
6. S. H. Park, H. S. Kim, “Flash light sintering of nickel nanoparticles for printed electronics”, Thin Solid Films, vol. 550, pp. 575-581, 2014.
7. K. C. Yung, C. M. Law, C. P. Lee, B. Cheung, T. M. Yue, “Size control and characterization of Sn-Ag-Cu lead-free nanosolders by a chemical reduction process”, Journal of Electronic Materials, vol. 41, No. 2, pp. 313-321, 2012.
8. L. Ding, R. L. Davidchack, J. Pan, “A molecular dynamics study of sintering between nanoparticles”, Computational Materials Science, vol. 45, No. 2, pp. 247-256, 2009.

9. K. S. Siow, “Mechanical properties of nano-silver joints as die attach materials”, Journal of Alloys and Compounds, vol. 514, pp. 6-19, 2012.
10. J. G. Bai, G. Q. Lu, “Thermomechanical reliability of low-temperature sintered silver die attached SiC power device assembly”, IEEE Transactions on Device and Materials Reliability, vol. 6, No. 3, pp. 436-441, 2006.
11. Z. Zhang, C. Chen, Y. Yang, H. Zhang, D. Kim, T. Sugahara, S. Nagao, K. Suganuma, “Low-temperature and pressureless sinter joining of Cu with micron/submicron Ag particle paste in air”, Journal of Alloys and Compounds, vol. 780, pp. 435-442, 2019.
12. C. Chen, K. Suganuma, “Microstructure and mechanical properties of sintered Ag particles with flake and spherical shape from nano to micro size”, Materials and Design, vol. 162, pp. 311-321, 2019.
13. L. Sun, M. Chen, L. Zhang, P. He, L. Xie, “Recent progress in SLID bonding in novel 3D-IC technologies”, Journal of Alloys and Compounds, vol. 818, No. 152825, pp. 1-18, 2020.
14. C. C. Lee, C. Y. Wang, G. S. Matijasevic, “A new bonding technology using gold and tin multilayer composite structures”, IEEE Transactions on Components, Hybrids, and Manufacturing Technology, vol. 14, No. 2, pp. 407-412, 1991.
15. C. C. Lee, C. Y. Wang, G .S. Matijasevic, “Au-In bonding below the eutectic temperature. IEEE Transactions on Components, Hybrids, and Manufacturing Technology”, vol. 16, No. 3, pp. 311–316, 1993
16. R. W. Chuang, C. C. Lee, “Silver-indium joints produced at low temperature for high temperature devices”, IEEE Transactions on Components and Packaging Technology, vol. 25, No. 3, pp. 453-458, 2002.
17. J. F. Li, P. A. Agyakwa, C. M. Johnson, “Kinetics of Ag3Sn growth in Ag-Sn-Ag system during transient liquid phase soldering process”, Acta Materialia, vol. 58, pp. 3429-3443, 2010.
18. H. Y. Zhao, J. H. Liu, Z. L. Li, Y. X. Zhao, H. W. Niu, X. G. Song, H. J. Dong, “Noninterfacial growth of Cu3Sn in Cu/Sn/Cu joints during ultrasonic-assisted transient liquid phase soldering process”, Material Letter, vol. 186, pp. 283-288, 2017.
19. C. C. Li, C. K. Chung, W. L. Shih, C. R. Kao, “Volume shrinkage induced by interfacial reaction in micro-Ni/Sn/Ni joints”, Metallurgical and Materials Transactions A, vol. 45, No. 5, pp. 2343-2346, 2014.
20. S. Marauska, M. Claus, T. Lisec, B. Wagner, “Low temperature transient liquid phase bonding of Au/Sn and Cu/Sn electroplated material systems for MEMS wafer-level packaging”, Microsystem Technologies, vol. 19, No. 8, pp. 1119-1130, 2013.
21. C. C. Lee, P. J. Wang, J. S. Kim, “Are intermetallics in solder joints really brittle?”, In 2007 Proceedings 57th Electronic Components and Technology Conference, pp. 648–652, 2007
22. Y. Y. Wu, C. C. Lee, “The Strength of High-Temperature Ag-In Joints Produced Between Copper by Fluxless Low-Temperature Processes”, Journal of Electronic Packaging, vol. 136, No. 1, pp. 011006–1–6, 2014
23. Y. Y. Wu, D. Nwoke, F. D. Barlowand, C. C. Lee, “The Thermal Cycling Reliability Study of Ag-In Joints between Silicon Chips and Copper Substrates Made by Fluxless Processes”, IEEE Transaction Components, Packaging, and Manufacturing Technology, vol. 4, No. 9, pp. 1420-1426, 2014.
24. Y. Huo, C. C. Lee, “The growth and stress vs. strain characterization of the silver solid solution with indium”, Journal of Alloys and Compounds, vol. 661, pp. 372-379, 2016.
25. Y. Huo, J. Wu, C. C. Lee, “Solid Solution Softening and Enhanced Ductility in Concentrated FCC Silver Solid Solution Alloys”, Material Science and Engineering A, vol. 729, pp. 208-218, 2018.
26. P. J. Rossi, N. Zotov, E. J. Mittemeijer, “Kinetics of intermetallic compound formation in thermally evaporated Ag-In bilayers”, Journal of Applied Physics, vol. 120, No. 16, pp. 165308, 2016.
27. R. Sheikhi, Y. Huo, C. H. Tsai, C. R. Kao, F. G. Shi, C. C. Lee, “Prior-to-bond annealing effects on the diamond-to-copper heterogeneous integration using silver–indium multilayer structure”, Journal of Materials Science: Materials in Electronics, vol. 31, No. 10, pp. 8059-8071, 2020.

28. Y. Y. Wu, W. P. Lin, C. C. Lee, "A study of chemical reactions of silver and indium at 180° C", Journal of Materials Science: Materials in Electronics, vol. 23, No. 12, pp. 2235-2244, 2012.
29. R. Sheikhi, Y. Huo, F. G. Shi, C. C. Lee, "Low Temperature VECSEL-to-Diamond Heterogeneous Integration with Ag-In Spinodal Nanostructured Layer", Scripta Materialia, vol. 194, pp. 113628–1–4, 2021
30. G. Q. Lu, W. Yang, Y. H. Mei, X. Li, G. Chen, X. Chen, "Mechanism of migration of sintered nanosilver at high temperatures in dry air for electronic packaging", IEEE Transactions on Device and Materials Reliability, vol. 14, No. 1, pp. 311-317, 2013.
31. C. A. Yang, J. Wu, C. C. Lee, C. R. Kao, "Analyses and design for electrochemical migration suppression by alloying indium into silver", Journal of Materials Science: Materials in Electronics, vol. 29, No. 16, pp. 13878-13888, 2018.
32. A. Vasile, I. Vasile, A. Nistor, L. Vladareanu, M. Pantazica, F. Caldararu, A. Bonea, A. Drumea, I. Plotog, "Rain sensor for automatic systems on vehicles", Advanced Topics in Optoelectronics, Microelectronics, and Nanotechnologies V. International Society for Optics and Photonics, pp. 7821–7821W, 2010
33. T. C. Cucu, I. Plotog, M. Branzei, "Mechanical Tests Regarding Low-Temperature Lead-Free Solder Pastes Application in Automotive Electronics", 2014 IEEE 20th International Symposium for Design and Technology in Electronic Packaging (SIITME). IEEE, pp. 63–68, 2014
34. Y. Liu, L. Pu, Y. Yang, Q. He, Z. Zhou, C. Tan, X. Zhao, Q. Zhang, K. N. Tu, "A high-entropy alloy as very low melting point solder for advanced electronic packaging, Materials Today Advances", Vol. 7, pp. 100101, 2020
35. V. Chidambaram, J. Hattel, J. Hald, "Hightemperature Lead-free Solder Alternatives," Microelectronic Engineering, Vol. 88, No. 6, pp. 981-989, 2011.
36. W. Liu, N. C. Lee, P. Bachorik, C. LaBarbera, "Effects of Solder Alloy Compositions on Microstructure and Reliability of Die-Attach Solder Joints for Automotive Applications", 2015 IEEE 17th Electronics Packaging and Technology Conference (EPTC). IEEE, pp. 1–7, 2015
37. H. Schoeller, S. Bansal, A. Knobloch, D. Shaddock, J. Cho, "Microstructure Evolution and the Constitutive Relations of High-Temperature Solders," Journal of Electronic Materials, Vol. 38, pp. 802-809, 2009.
38. G. Zeng, S. McDonald, K. Nogita, "Development of High-Temperature Solders: Review," Microelectronics Reliability, Vol. 52, pp. 1306–1322, 2012.
39. F. Dugal, M. Ciappa, "Study of Thermal Cycling and Temperature Aging on PbSnAg Die Attach Solder Joints for High Power Modules," Microelectronics Reliability, Vol. 54, pp. 1856-1861, 2014.
40. K. W. Moon, W. J. Boettinger, U. R. Kattner, F. S. Biancaniello, C. A. Handwerker, "Experimental and thermodynamic assessment of Sn-Ag-Cu solder alloys", Journal of electronic materials, Vol. 29, No. 10, pp. 1122-1136, 2000.
41. J. W. Yoon, B. I. Noh, B. K. Kim, C. C. Shur, S. B. Jung, "Wettability and interfacial reactions of Sn–Ag–Cu/Cu and Sn–Ag–Ni/Cu solder joints", Journal of alloys and compounds, Vol. 486, No.1-2, pp. 142-147, 2009.
42. L. Zhang, L. Sun, Y.H. Guo, "Microstructures and properties of Sn58Bi, Sn35Bi0.3Ag, Sn35Bi1.0Ag solder and solder joints", Journal of Materials Science: Materials in Electronics, Vol. 26, No. 10, pp. 7629-7634, 2015.
43. S.Y. Jang, K.W. Paik, "Comparison of electroplated eutectic Bi/Sn and Pb/Sn solder bumps on various UBM systems", IEEE transactions on electronics packaging manufacturing, Vol. 24, No. 4, pp. 269-274, 2001.
44. F. Hua, Z.Q. Mei, J. Glazer, "Eutectic Sn-Bi as an alternative to Pb-free solders", 1998 IEEE 48th Electronic Components and Technology Conference (ECTC), IEEE, pp. 277–283, 1998
45. Z. Mei, J.W. Morris, "Characterization of eutectic Sn-Bi solder joints, Journal of Electronic Materials", Vol. 21, No. 6, pp. 599–607, 1992

46. C.H. Raeder, L.E. Felton, V.A. Tanzi, D.B. Knorr, "The effect of aging on microstructure, room temperature deformation, and fracture of Sn-Bi/Cu solder joints", Journal of Electronic Materials, Vol. 23, No. 7, pp. 611-617, 1994.
47. W.R. Myung, Y. Kim, K.Y. Kim, S.B. Jung, "Drop reliability of epoxy-contained Sn-58 wt.% Bi solder joint with ENIG and ENEPIG surface finish under temperature and humidity test", Journal of Electronic Materials, Vol. 45, No. 7, pp. 3651-3658, 2016.
48. Y. Liu, K. N. Tu, "Low melting point solders based on Sn, Bi, and In elements", Materials Today Advances, Vol. 8, pp. 100115, 2020.
49. G. Humpston, D.M. Jacobson, "Indium solders", Advanced Materials & Processes, Vol. 163, pp. 45-47, 2005.
50. K. Shimizu, T. Nakanishi, K. Karasawa, K. Hashimoto, K. Niwa, "Solder joint reliability of indium-alloy interconnection", Journal of Electronic Materials, Vol. 24, No. 1, pp. 39-45, 1995.
51. J.L.F. Goldstein, J.W. Morris, "Microstructural development of eutectic Bi-Sn and eutectic In-Sn during high temperature deformation", Journal of Electronic Materials, Vol. 23, No. 5, pp. 477-486, 1994.
52. J. Seyyedi, "Thermal Fatigue Behaviour of Low Melting Point Solder Joints", Soldering & Surface mount technology, Vol. 5, pp. 26-32, 1993.
53. J. Cheong, A. Goyal, S. Tadigadapa, C. Rahn, "Reliable bonding using indium-based solders", Proceedings of Spie the International Society for Optical Engineering, Vol. 5343, pp. 114-120, 2004.
54. A.B. El-Bediwi, M.M. El-Bahay. "Influence of silver on structural, electrical, mechanical and soldering properties of tin-indium based alloys", Radiation Effects and Defects in Solids, Vol. 159, No. 2, pp. 133-140, 2004.
55. J.W. Kim, S.B. Jung, "Characterization of the shear test method with low melting point In–48Sn solder joints", Materials Science and Engineering A, Vol. 397, pp. 145-152, 2005.
56. H. Zhang, S. Lytwynec, H. Wang, J. Geng, F. Mutuku, N. C. Lee, "A Novel Bi-Free Low Temperature Solder Paste with Outstanding Drop-Shock Resistance", 2021 IEEE 71st Electronic Components and Technology Conference (ECTC), IEEE, pp. 643–653, 2021
57. Y. A. Shen, S. Zhou, J. Li, C. H. Yang, S. Huang, S. K. Lin, H. Nishikawa, "Sn-3.0 Ag-0.5 Cu/Sn-58Bi composite solder joint assembled using a low-temperature reflow process for PoP technology", Materials & Design, Vol. 183, pp. 108144, 2019.
58. S. Mokler, R. Aspandiar, K. Byrd, O. Chen, S. Walwadkar, K. K. Tang, M. Renavikar, S. Sane, "The application of Bi-based solders for low temperature reflow to reduce cost while improving SMT yields in client computing systems", Proceedings of SMTA International, pp. 318–326, 2013
59. R.S. Patti, Three-dimensional integrated circuits and the future of system-on-chip designs, Proceedings of the IEEE, 94 (2006) 1214-1224.
60. A.W. Topol, D. La Tulipe, L. Shi, D.J. Frank, K. Bernstein, S.E. Steen, A. Kumar, G.U. Singco, A.M. Young, K.W. Guarini, Three-dimensional integrated circuits, IBM Journal of Research and Development, 50 (2006) 491-506.
61. K. Tu, Reliability challenges in 3D IC packaging technology, Microelectronics Reliability, 51 (2011) 517-523.
62. Y. Liu, Y.-C. Chu, K. Tu, Scaling effect of interfacial reaction on intermetallic compound formation in Sn/Cu pillar down to 1 μm diameter, Acta Materialia, 117 (2016) 146-152.
63. R.I. Made, P. Lan, H.Y. Li, C.L. Gan, C.S. Tan, Study of the evolution of Cu-Cu bonding interface imperfection under direct current stressing for three dimensional integrated circuits, in: 2011 IEEE International Interconnect Technology Conference, IEEE, 2011, pp. 1-3
64. Y. I. Kim, K. H. Yang, W. S. Lee, "Thermal degradation of DRAM retention time: Characterization and improving techniques." IRPS, 667–668, 2004
65. T. Kim, M. Howlader, T. Itoh, T. Suga, Room temperature Cu–Cu direct bonding using surface activated bonding method, Journal of Vacuum Science & Technology A: Vacuum, Surfaces, and Films, 21 (2003) 449-453.

66. K. Chen, C. Tan, A. Fan, R. Reif, "Copper Bonded Layers Analysis and Effects of Copper Surface Conditions on Bonding Quality for Three Dimensional Integration," J. Electron. Mater., 34(12) (2005) 1464.
67. K. Chen, A. Fan, C. Tan, R. Reif, "Bonding Parameters of Blanket Copper Wafer Bonding," J. Electron. Mater., 35(2) (2006) 230.
68. B. Swinnen, et al. "3D Integration by Cu–Cu ThermoCompression Bonding of Extremely Thinned Bulk-Si Die Containing 10 lm Pitch Through-Si Vias." IEDM, 11 (2006) 1.
69. A. Huffman, J. Lannon, M. Lueck, C. Gregory, D. Temple, "Fabrication and Characterization of Metal-to-Metal Interconnect Structures for 3-D Integration." J. Instrum., 4(3) (2009) 03006.
70. E. J. Jang, S. Hyun, H. J. Lee, Y. B. Park, "Effect of Wet Pretreatment on Interfacial Adhesion Energy of Cu–Cu Thermocompression Bond for 3D IC Packages." J. Electron. Mater., 38 (2009) 2449.
71. T. H. Hung, et al. "Investigation of Wet Pretreatment to Improve Cu-Cu Bonding for Hybrid Bonding Applications" ECTC, 2021
72. J. W. Kim, S. J. Jeon, H. J. Lee, S. Hyun, Y. B. Park, "Improvement of Wafer-Level Cu-to-Cu Bonding Quality Using Wet Chemical Pretreatment." J. Nanosci.Nanotechnol., 12(4) (2012) 3577.
73. M.J. Aziz, Thermodynamics of diffusion under pressure and stress: Relation to point defect mechanisms, Applied physics letters, 70 (1997) 2810-2812.
74. C.-M. Liu, H.-W. Lin, Y.-S. Huang, Y.-C. Chu, C. Chen, D.-R. Lyu, K.-N. Chen, K.-N. Tu, Low-temperature direct copper-to-copper bonding enabled by creep on (111) surfaces of nanotwinned Cu, Scientific reports, 5 (2015) 9734.
75. Jing-Ye Juang, Chia-Ling Lu, Kuan-Ju Chen, Chao-Chang A. Chen, Po-Ning Hsu, Chih Chen & K. N. Tu, Copper-to-copper direct bonding on highly (111)-oriented nanotwinned copper in no-vacuum ambient, Scientific Reports, 8 (2018) 13910
76. C.-M. Liu, H.-w. Lin, Y.-C. Chu, C. Chen, D.-R. Lyu, K.-N. Chen, K. Tu, Low-temperature direct copper-to-copper bonding enabled by creep on highly (1 1 1)-oriented Cu surfaces, Scripta Materialia, 78 (2014) 65-68.
77. L.-P. Chang, S.-Y. Huang, T.-C. Chang, F.-Y. Ouyang, Journal of Alloys and Compounds 862 (2021) 158587.
78. L. Lu, Y. Shen, X. Chen, L. Qian, K. Lu, Science 304(5669) (2004) 422-426.
79. X. Zhang, A. Misra, H. Wang, T. Shen, M. Nastasi, T. Mitchell, J. Hirth, R. Hoagland, J. Embury, Acta Materialia 52(4) (2004) 995-1002.
80. O. Anderoglu, A. Misra, H. Wang, F. Ronning, M. Hundley, X. Zhang, Applied Physics Letters 93(8) (2008) 083108.
81. D. Bufford, H. Wang, X. Zhang, Acta Materialia 59(1) (2011) 93-101.
82. D. Bufford, H. Wang, X. Zhang, Journal of Materials Research 28(13) (2013) 1729.
83. F.-Y. Ouyang, K.-H. Yang, L.-P. Chang, Surface and Coatings Technology (2018)
84. J. Li, D. Xie, S. Xue, C. Fan, Y. Chen, H. Wang, J. Wang, X. Zhang, Acta Materialia 151 (2018) 395-405.
85. Jing-Ye Juang, Chia-Ling Lu, Yu-Jin Li, K. N. Tu and Chih Chen, Correlation between the Microstructures of Bonding Interfaces and the Shear Strength of Cu-to-Cu Joints using (111)-Oriented and Nanotwinned Cu, Materials, 11(12) (2018) 2368
86. I. Karakaya, W. Thompson, The Ag-O (silver-oxygen) system, Journal of phase equilibria, 13 (1992) 137-142.
87. D.R. Gaskell, D.E. Laughlin, Introduction to the Thermodynamics of Materials, CRC press, 2017
88. Cheng-Ta Ko, Zhi-Cheng Hsiao, Huan-Chun Fu, Kuan-Neng Chen, Wei-Chung Lo, Yu-Hua Chen, Wafer-to wafer hybrid bonding technology for 3D IC, 3rd Electronics System Integration Technology Conference ESTC, 2010
89. M.Park,S.Baek,S.Kim,S.E.Kim,ArgonplasmatreatmentonCusurfaceforCubondingin 3D integration and their characteristics. Appl. Surf. Sci. 324, 168–173 (2015). https://doi.org/10.1016/j.apsusc.2014.10.098

90. S.L.Chua,G.Y.Chong,Y.H.Lee,C.S.Tan,Directcopper-copperwaferbondingwithAr/N2 plasma activation, in 2015 IEEE International Conference on Electron Devices and Solid-State Circuits (EDSSC) (IEEE, 2015), pp. 134–137
91. P. Enquist, G. Fountain, C. Petteway, A. Hollingsworth, H. Grady, Low cost of ownership scalable copper direct bond interconnect 3D IC technology for three dimensional integrated circuit applications, in 2009 IEEE International Conference on 3D System Integration (IEEE, 2009), pp. 1–6
92. P.Enquist,Metal/siliconoxidehybridbonding,inP.Ramm,J.J.-Q.Lu,M.M.V.Taklo,Handb. Wafer Bond eds by (Wiley, Weinheim, Germany, 2012), pp. 261–278
93. L. Di Cioccio, S. Moreau, L. Sanchez, F. Baudin, P. Gueguen, S. Mermoz, Y. Beilliard, R. Taibi, Cu/SiO2 Hybrid Bonding, in P. Garrou, M. Koyanagi, P. Ramm, eds by Handb. 3D Integr (Wiley, KGaA, 2014), pp 295–312
94. L.D.Cioccio,P.Gueguen,R.Taibi,D.Landru,G.Gaudin,C.Chappaz,F.Rieutord,F.deCrecy, I. Radu, L.L. Chapelon, L. Clavelier, An overview of patterned metal/dielectric surface bonding: mechanism, alignment and characterization. J. Electrochem. Soc. 158, P81-P86 (2011). https://doi.org/10.1149/1.3577596
95. I.Radu,D.Landru,G.Gaudin,G.Riou,C.Tempesta,F.Letertre,L.DiCioccio,P.Gueguen, T. Signamarcheix, C. Euvrard, J. Dechamp, Recent Developments of Cu-Cu non-thermo compression bonding for wafer-to-wafer 3D stacking, in 2010 IEEE International 3D Systems Integration Conference (3DIC) (IEEE, Munich, 2010), pp. 1–6
96. A.Shigetou,T.Suga,Modified diffusion bonding for both Cu and SiO_2 at 150C in ambient air, in 2010 Proceedings 60th Electronic Components and Technology Conference (ECTC) (IEEE, Las Vegas, NV, USA, 2010), pp. 872–877
97. J.J.McMahon,J.Q.Lu,R.J.Gutmann,Waferbondingofdamascene-patternedmetal/adhesive redistribution layers for via-first three-dimensional (3D) interconnect, in 55th Proceedings Electronic Components and Technology, 2005. ECTC'05. (IEEE, 2005), pp. 331–336
98. Z.-C.Hsiao,C.-T.Ko,H.-H.Chang,H.-C.Fu,C.-W.Chiang,C.-K.Hsu,W.-W.Shen,W.-C. Lo, Cu/BCB Hybrid Bonding With TSV for 3D Integration by Using Fly-Cutting Technology (IEEE, Kyoto, Japan, 2015), pp. 834–837
99. W.Yang, M.Akaike, M.Fujino, T.Suga, A combined process of formicacid pretreatment for low-temperature bonding of copper electrodes. ECS J. Solid State Sci. Technol. 2, P271-P274 (2013). https://doi.org/https://doi.org/10.1149/2.010306jss
100. Tadatomo Suga, Ran He, George Vakanas, and Antonio La Manna, Chapter 8 Direct Cu to Cu Bonding and Other Alternative Bonding Techniques in 3D Packaging in book: 3D Microelectronic Packaging, Springer, 2017

Cameras in Advanced Driver-Assistance Systems and Autonomous Driving Vehicles

Zhenhua Lai

Abstract Cameras have become one of the most important sensors in advanced driver-assistance systems (ADAS) and autonomous driving (AD) vehicles. There are different ways to categorize cameras in ADAS/AD vehicles based on the camera's placement, application, and technology. Most camera systems consist of hardware components that compose the camera module and image processing components that control the hardware and perform digital operations on captured images. An overview of camera systems, hardware, image processing, and product development processes is introduced in this chapter.

1 Introduction

Cameras have become increasingly important in ADAS/AD systems. There are multiple ways to categorize cameras in ADAS/AD systems: exterior cameras and in-cabin cameras if categorized by placement; human vision cameras and computer vision cameras if categorized by application; and monocular cameras, stereo cameras, infrared (IR) cameras, and time-of-flight (ToF) cameras, etc., if categorized by technology. Human vision cameras are widely adopted as back-up cameras, surround view cameras, and rear-view mirror replacement cameras (or e-mirrors). Some example applications of computer vision cameras include lane departure detection, object (*e.g.*, pedestrians, vehicles, traffic signs) recognition, and driver monitoring.

An overview of camera systems is introduced in Sect. 2. Most camera systems consist of hardware components that compose the camera module and image processing components that control the hardware and perform digital operations on captured images. Camera hardware and image processing are described in Sect. 3 and Sect. 4, respectively. Section 5 discusses the camera product development process. A summary of this chapter is presented in Sect. 6.

Z. Lai (✉)
Level 5 Self-driving Division, Lyft, Inc, Palo Alto, CA, USA
e-mail: lai.z@northeastern.edu

Y. Li and H. Shi (eds.), *Advanced Driver Assistance Systems and Autonomous Vehicles*,
https://doi.org/10.1007/978-981-19-5053-7_7

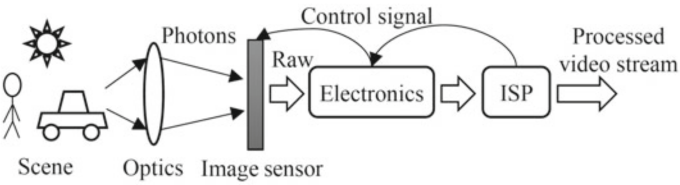


Fig. 1 Camera system overview

2 Camera System Overview

A typical camera system is shown in Fig. 1. Photons from the scene are first collected and filtered by optical components, which then form an image at the image sensor. The image sensor converts photons into raw images, which are transmitted by electronics to an image signal processor (ISP). The ISP processes raw images into a processed video stream. The ISP also controls the image sensor and provides image sensor configuration and control (*e.g.*, exposure, gain) through the electronics.

Typically, the optical components and the image sensor are packaged into a camera module with electronics. The ISP may be packaged into the same camera module or at a separate location because of various design considerations (*e.g.*, size, power, cost).

3 Camera System Hardware

3.1 Image Sensor

In digital imaging, an image sensor first converts photons into electrons, which are then digitized by a readout circuit. The two most important and common image sensor technologies are charge coupled device (CCD) and complementary metal-oxide semiconductor (CMOS).

Both CCD and CMOS sensors accumulate photogenerated charge in each pixel proportional to the local illumination intensity. In a CCD sensor, pixels are exposed simultaneously. When exposure is complete, the CCD sensor transfers each pixel's charge packet sequentially to a common output node and then converts the charge to a voltage. In a CMOS image sensor, however, the exposure of the whole frame can be either simultaneous, known as global shutter, or line-by-line, known as rolling shutter. The charge-to-voltage conversion in a CMOS sensor takes place in each pixel.

Most image sensors used in ADAS/AD systems are CMOS sensors because of their advantages in cost, power consumption, size, and the ability to integrate complex circuitry. All sensors mentioned in this chapter are CMOS sensors unless otherwise specified.

3.1.1 CMOS Sensor Architecture

CMOS sensor architecture (Fig. 2) has evolved from a front-side illuminated (FSI) technology to a back-side illuminated (BSI) technology and recently to a 3D stacked BSI technology.

The first-generation CMOS sensor architecture uses an FSI architecture (Fig. 2a). In an FSI sensor, photons are first collected by an on-chip micro-lens, filtered by a color filter, passed through multiple metal and dielectric layers, and then converted into electrons after reaching the photodiode substrate. The main challenge with FSI sensors is metal and dielectric layers reduce the throughput of photons reaching the photodiode, otherwise known as the quantum efficiency (QE). Another issue with FSI sensors is that incoming photons with high angles of incidence may map to adjacent pixels, which is known as crosstalk. The two issues above reduce the sensitivity or signal-to-noise ratio (SNR) of the sensor.

A BSI sensor solves the abovementioned issues of an FSI sensor by placing the metal and dielectric underneath the photodiode (Fig. 2b). The manufacturing of BSI sensors is much more complicated than that of FSI and is therefore more expensive.

A 3D stacked BSI sensor uses a signal processing chip to replace the traditional supporting substrate layer in a BSI sensor (Fig. 2c). Such technology has significantly reduced the size of an image sensor while simultaneously enabling the design to integrate more image processing functions and achieve faster speeds and lower power consumption.

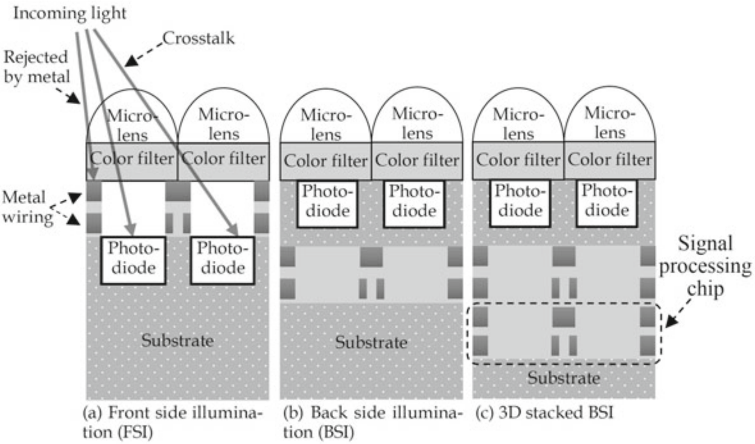


Fig. 2 CMOS sensor architecture evolvement

3.1.2 Color Filter Array (CFA)

Color filter arrays (CFA) are utilized to produce colors in an image sensor. Without a CFA, the image sensor will only be able to provide monochrome images.

The most widely adopted CFA is the Bayer pattern (Fig. 3a), which uses 2 × 2-pixel matrices with a red pixel, two green pixels, and a blue pixel to reproduce RGB colors. A main advantage of the Bayer pattern is that it is able to reproduce colors very close to human eye perception. However, the RGB filters of the Bayer pattern provide relatively poor low-light performance as they reduce the QE by filtering a majority of the light across the visible spectrum.

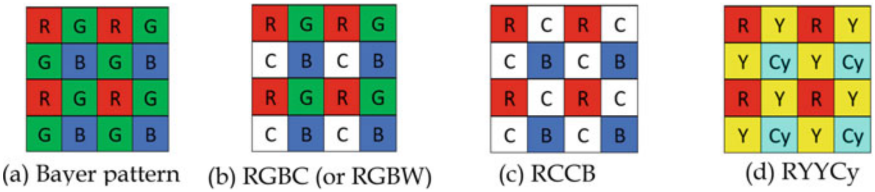
In ADAS/AD applications, most computer vision algorithms do not require high color fidelity for object detection but do have higher requirements for low-light performance. As a result, other CFAs are utilized to improve low-light performance.

One broadly accepted CFA is the RGBC (or RGBW) pattern (Fig. 3b), which is similar to the Bayer pattern but employs a clear (or white) pixel to replace one of the green pixels. The clear pixel does not have a color filter, which increases it's QE and therefore improves low-light performance but decreases its color fidelity. The RGBC pattern is broadly utilized for RGB + IR cameras, which use the clear pixel to detect infrared (IR) light.

For non-IR applications, RCCB (Fig. 3c) is more popular. Note that for RCCB and RGBC patterns, even though the clear pixel improves QE, the large QE differences between clear pixels and color pixels cause high chroma noise at low light.

Another commonly applied CFA for ADAS/AD applications is the RYYCy pattern (Fig. 3d), which uses a red pixel, two yellow pixels, and a cyan pixel in a 2 × 2-pixel matrix. There are several benefits with this CFA: (1) Red and yellow colors are elements of traffic lights; (2) yellow and cyan span broader on the visible spectrum, which means they bring better low-light performance; and (3) green and blue colors can be reproduced by using Y-R and Cy-R, respectively.

Other CFAs (*e.g.*, RCCC, RYYB, etc.) also exist in automotive applications. The choice of CFA is determined by the requirements of the applications. A comparison of popular CFAs is shown in Table 1.



(a) Bayer pattern (b) RGBC (or RGBW) (c) RCCB (d) RYYCy

Fig. 3 Color filter array

Table 1 Comparison of popular color filter arrays

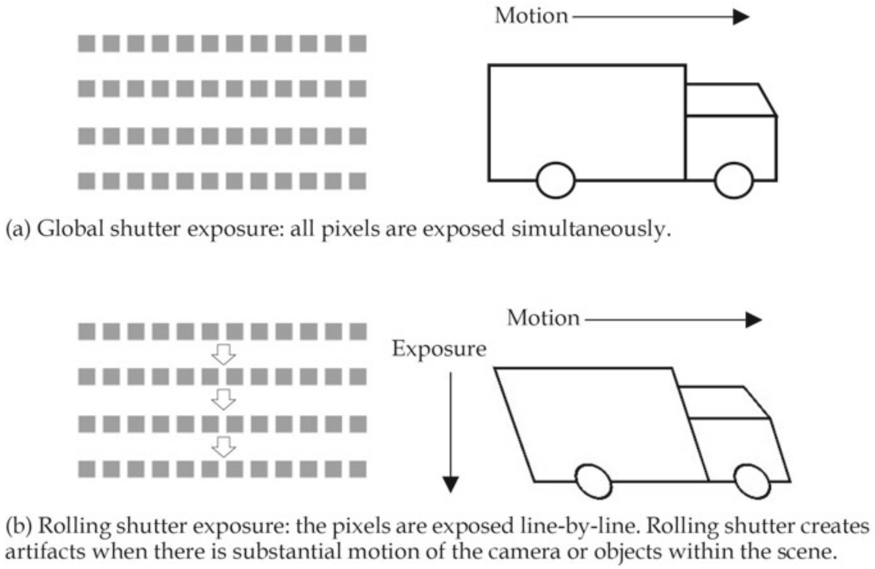
CFA	Low-light performance	Color fidelity
Bayer (RGGB)	Poor	Best
RGBC	Better	Better
RCCB	Best	Good
RYYCy	Good	Poor

3.1.3 Global Shutter Versus Rolling Shutter

There are two kinds of shutter control in CMOS image sensors: global shutter and rolling shutter. As shown in Fig. 4, a global shutter sensor exposes all pixels at the same time, while a rolling shutter sensor exposes the pixels in a line-by-line manner.

Rolling shutter sensors use a simpler electric design and usually demonstrate better sensitivity, lower cost, and smaller size. However, the disadvantage for rolling shutter sensors is when there is substantial motion of the camera or objects within the scene. As shown in Fig. 4b, since each line of the pixels starts capturing at different time frames, the moving object appears distorted in the rolling shutter sensor.

Rolling shutter sensors are capable of meeting the requirements of most of ADAS/AD applications except for some special cases where high spatial accuracy is needed; *e.g.*, some driver-monitoring systems (DMS) use global shutter cameras to capture the rapid eye movements of the driver to monitor whether the driver is focusing on the road or not.



(a) Global shutter exposure: all pixels are exposed simultaneously.

(b) Rolling shutter exposure: the pixels are exposed line-by-line. Rolling shutter creates artifacts when there is substantial motion of the camera or objects within the scene.

Fig. 4 Global shutter versus rolling shutter

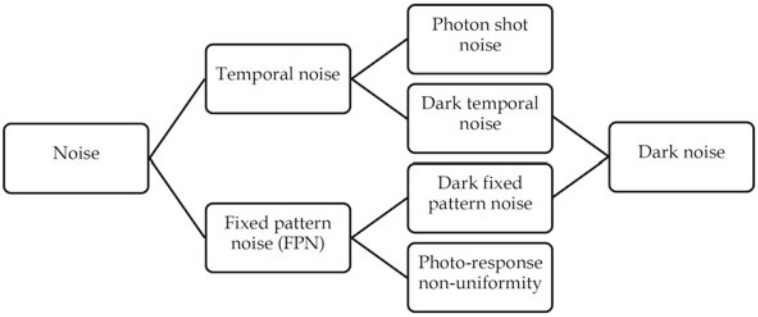


Fig. 5 Noise components

3.1.4 Concepts and Specifications of an Image Sensor

Noise

There are two major ways to categorize noise in an image sensor: by temporal behavior or by illumination (Fig. 5).

If categorized by temporal behavior, there is temporal noise and fixed pattern noise (FPN). Any noise that fluctuates over time is considered temporal noise. FPN is distributed spatially at fixed locations of an image sensor due to pixel-to-pixel variation.

Temporal noise on a dark frame, a frame without any illumination to the image sensor, is referred to as dark temporal noise, which is generally caused by variation of the dark current, the electric current that is observed when the sensor is not illuminated. Dark temporal noise is typically consistent under the same temperature and sensor gain. Photon shot noise is commonly the main source of temporal noise when the image sensor is under strong illumination. Photon shot noise originates from the uncertain and quantized nature of photons. This means that the arrival of a given photon on a specific pixel is random, and due to the discrete nature of photons, a pixel can only collect a whole photon but not a fraction of a photon. As a result, the variation in the optical intensity follows the Poisson distribution, the noise in which is equal to the square root of the signal. Therefore, the photon shot noise is equal to the square root of the signal in the electron and the photon domains. Note that in the digital number domain, photon shot noise may not be the square root of the signal as there is a conversion gain applied.

FPN on a dark frame is known as dark signal non-uniformity (DSNU). DSNU is normally a constant for a given sensor gain. FPN that varies by illumination is referred to as photo-response non-uniformity (PRNU). The photo-response of each pixel is linear to the input light. The ratio between output signal and input light may vary by pixels. Such variation among pixels is the main cause of PRNU. PRNU is linear to the signal.

Dark noise is the total noise in a dark frame. DSNU and dark temporal noise both contribute to dark noise. Dark noise is also called read noise as this kind of noise comes from the readout electronics.

In general, in a camera system, signal-to-noise ratio (SNR) increases as illumination increases. However, PRNU usually determines the maximum SNR the camera system is able to achieve.

LED Flicker Mitigation (LFM)

Nowadays, light-emitting diodes (LEDs) are widely adopted in traffic systems (*e.g.*, traffic lights, traffic signs, and vehicle headlamps) because they provide various benefits, including long lifespan, high energy efficiency, and low cost. Though with so many benefits, LEDs raise a challenging concern in the ADAS/AD industry: LED flicker.

Most LEDs are discrete in nature, which means that they are turned on and off periodically; however, the period is fast enough that they appear to be on continuously to human eyes. Different LEDs may operate under very different periods and duty cycles. If the camera captures a frame during the LED off cycle, the camera will not be able to detect the status of the LED.

In order to mitigate LED flicker, the camera needs to apply a long exposure time, which may cause motion blur, saturation, or degraded frame rate. LED flicker mitigation (LFM) is an image sensor function commonly combined with the latest high dynamic range (HDR) technologies, which will be introduced in Sect. 3.1.4.3.

Dynamic Range

Dynamic range is defined as the maximum detectable signal divided by the minimum detectable signal and is expressed in units of dB. Maximum detectable signal is limited by full well capacity, which is the largest number of electrons a pixel is able to hold. The minimum detectable signal is defined as the signal when SNR equals to one. Since minimum detectable signal is determined by read noise, in practice, we use full well capacity divided by read noise to calculate dynamic range.

Dynamic range is one of the most important specifications in ADAS/AD applications because of the large dynamic range of the real-world scene. Various technologies are employed to significantly increase the dynamic range from the traditional linear single capture mode to HDR mode. There are three common technologies to achieve HDR: multi-exposure, split pixels, and dual conversion gain (DCG).

In multi-exposure HDR, the camera takes multiple exposures of the scene, with at least one short exposure for bright signals and at least one long exposure for dark signals, and then combines all the exposures into one single HDR image. Traditionally, this approach can be implemented in the ISP. However, in ADAS/AD applications, it is preferred to implement such a feature in the image sensor in order to achieve faster capture, lower noise, and less motion artifacts. Multi-exposure HDR

Table 2 Comparison of three common HDR technologies

Technology	Dynamic range	LFM	Motion artifacts	Image quality
Multi-exposure	Best	No	Poor	Good
Split pixel	Better	Yes	Good	OK
DCG	Good	Yes	Good	Good

demonstrates various benefits in flexibility, dynamic range, and image quality. The main trade-off of multi-exposure HDR is relatively high motion artifacts and no LFM.

Split pixel HDR splits a pixel in the image with multiple sub-pixels with different sizes placed next to each other on the image sensor. Pixels with different sizes are able to perform different exposure times so that LFM can be achieved by using a long exposure time on the small pixel. Split pixel HDR minimizes the motion artifact issue in multi-exposure HDR and meanwhile achieves very good LFM capability. However, this method sacrifices color fidelity and sharpness. Split pixel HDR results in a slightly lower dynamic range compared with multi-exposure HDR.

Another widely adopted HDR technology is DCG. Conversion gain is defined as the output voltage per electron in the photodiode. A high conversion gain is desired for low-light images, while a low conversion gain is necessary for bright signals to avoid saturation. DCG sensors improve the dynamic range by changing the conversion gain based on the illumination level. DCG sensors demonstrate very good image quality with LFM capability and meanwhile minimize motion artifacts. However, the dynamic range of DCG is usually lower than that of split pixels.

A comparison of the abovementioned HDR technologies is shown in Table 2.

State-of-the-art image sensors combine multiple abovementioned HDR technologies together to achieve high dynamic range, high image quality, and LFM capability while simultaneously minimizing motion artifacts.

3.2 Optics

The main optical components in a camera module typically include a lens and a bandpass filter. Some cameras with active illumination (*e.g.,* infrared (IR) cameras, time-of-flight (ToF) cameras) also include an optical emitter.

3.2.1 Lens

A lens in a camera focuses the light rays from the scene, conventionally referred to as the object plane, to the image sensor, also known as the image plane. Though generally being referred to as a “lens,” most “lenses” in cameras are lens groups. A lens group normally includes multiple lenses packaged into a lens barrel.

Lens Parameters

Modulation Transfer Function (MTF)

Modulation transfer function (MTF) is the impulse response of the optical system represented in the spatial frequency domain. As one of the most important parameters of an optical system, MTF is used to quantify sharpness.

The value of the MTF varies from 0 ~ 1 or 0 ~ 100% and is presented with its corresponding spatial frequency, commonly with a unit of cycles/mm (or cycles/pixel, line-pair/mm, line-pair/pixel), which means the number of dark and light stripes being resolved per unit length in the image plane. MTF is usually normalized to be 1 or 100% at 0 cycles/mm as a convention.

F-number

F-number is defined as the lens focal length divided by the diameter of the lens entrance pupil, the aperture as seen from the object side, as shown in Eq. (1) below:

$$N = \frac{f}{D} \tag{1}$$

where N is the f-number, f is the focal length, and D is the entrance pupil diameter.

F-number is an important parameter of a lens as it affects sensitivity and depth of field (DoF). DoF of an optical system is the distance between the closest and furthest objects in a scene that have an acceptable sharpness in an image.

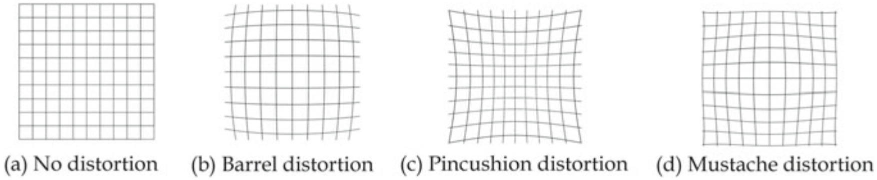
The sensitivity of the optical system is inversely proportional to the square of the f-number, which means a small f-number leads to better low-light performance. However, a smaller f-number also results in a smaller DoF. In the practice of optical design, people mostly prefer better sensitivity and large DoF. Therefore, the f-number needs to be carefully chosen to balance the needs of sensitivity and DoF.

Distortion

When light rays are projected from the scene to the image through the lens, it is very common to see that the image is distorted from the real scene. This is called distortion. Distortion happens when the magnification of the image varies with distance from the optical axis, an imaginary line that passes through the geometrical center of a lens and orthogonal to the image plane.

Distortion is categorized in three types based on how the image magnification varies from the optical axis (Fig. 6): (1) barrel distortion, when image magnification decreases with distance from the optical axis, (2) pincushion distortion, when image magnification increases with distance from the optical axis, and (3) mustache distortion, a mixture of barrel distortion and pincushion distortion.

Distortion is traditionally considered less important for human vision applications as there are mature image processing algorithms to correct distortion. However, in



(a) No distortion (b) Barrel distortion (c) Pincushion distortion (d) Mustache distortion

Fig. 6 Different types of distortion

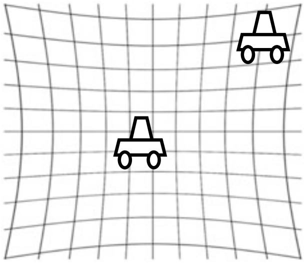


Fig. 7 Resolution change due to distortion

computer vision applications, distortion can be important as the distortion correction algorithms are not able to restore the resolution loss due to magnification increase. For example, in Fig. 7, the object occupies a 3 × 3 pixel array when projected to the image center; however, the same object only occupies a 2 × 2 pixel array when projected to the image corner due to pincushion distortion. Existing distortion correction algorithms are not able to improve object detection caused by resolution loss.

Relative Illumination (RI)

Relative illumination (RI) is a percentage of illumination at any point on the image plane divided by the maximum illumination on the image plane. When referred to in a lens or camera specification, RI is referred to as the lowest RI across the whole image plane, which means the lowest illumination on the image plane divided by the maximum illumination on the image plane.

RI is a specification used to describe shading, which is a reduction of an image's brightness from the center to the periphery. There are typically two factors contributing to shading: roll-off (or falloff) and vignetting.

Roll-off is the decrease in illumination from the image center to the corners caused by the lens. Roll-off is a physical property of the lens. The amount of roll-off varies by the lens design.

Vignetting is a reduction of illumination caused by a physical obstruction, *e.g.*, some object blocking the camera field of view (FOV), a mechanical stop inside the lens blocking light rays, etc. Vignetting can also occur when the image formed by the lens is smaller than the image area of the sensor.

In digital cameras, there is a special kind of vignetting called pixel vignetting. When incident light hits the image sensor, its power decreases as the incidence angle increases, which causes higher power loss toward the edge of the sensor. Some image sensor manufacturers shift the micro-lens on each pixel in order to compensate for the pixel vignetting.

RI is important in computer vision because even though lens shading can be corrected by ISP algorithms, such correction can only correct illumination variation but does not improve SNR. As mentioned in Sect. 3.1.4.1, in a camera system, SNR increases as signal increases. And therefore, lower RI normally leads to lower SNR at the corners.

Chromatic Aberration

Chromatic aberration, also known as "color fringing" or "purple fringing," is a phenomenon in which light rays passing through a lens focus at different locations, depending on their wavelengths. Chromatic aberration is caused by dispersion, which means the refractive index of the material varies by wavelength of the light.

3.2.2 Optical Filter

An optical filter selectively blocks light from unwanted wavelengths but passes light from desired wavelengths. In a camera system designed for the visible light range, the optical filter should block infrared (IR) light but pass visible light. Such an optical filter is called an IR cut filter. In a camera system designed for IR light, the optical filter should only pass light with the same wavelength as the optical emitter but blocks all other wavelengths. This kind of optical filter is known as a bandpass filter.

There are two kinds of optical filters: interference filters (also known as dichroic filters) and absorption filters. Interference filters use multiple layers of coatings to form a sequential series of reflective layers that resonate with the desired wavelengths but reflect and cancel the unwanted wavelengths. Absorption filters utilize the absorption spectrum of a material to absorb unwanted wavelengths. Interference filters can precisely control the wavelength ranges by adjusting the thickness and properties of the coatings but the performance degrades as the incidence angle of the incoming light increases. Absorption filters are more difficult to make to precise transmittance specifications but their performance usually does not vary with incidence angle. In practice, both kinds of filters can be used together to achieve optimum performance. For example, in a camera system designed for visible light range, blue glass is commonly used as a good absorption filter, and additional coatings can be applied to blue glass to improve its transmittance spectrum.

3.2.3 Optical Emitter

Common optical emitters in automotive cameras with active illumination are light-emitting diode (LED) and vertical-cavity surface-emitting laser (VCSEL).

Though the cost of VCSELs has decreased significantly in recent years, they are still generally more expensive than LEDs. However, VCSELs provide narrower beam divergence, smaller bandwidth, higher power, and higher efficiency, which eventually lead to better optical performance, lower power consumption, and smaller size for the whole camera system.

Whether to choose LEDs or VCSELs in the camera system will eventually depend on various factors. For example, for large FOV applications, beam divergence may become less important, and therefore, LEDs can be more suitable. In longer range applications, VCSELs provide better range because of narrower beam divergence feature.

In time-of-flight (ToF) cameras, where the emitted light needs to be modulated at high frequencies, VCSELs are able to achieve much higher modulation frequencies compared with LEDs.

For near-infrared (NIR) emitters, another factor to consider is wavelength. The most commonly used wavelength in automotive applications is 940 nm as there is a dip in the 920–960 nm region in sunlight spectrum (Fig. 8), and therefore, 940 nm offers better performance under sunlight.

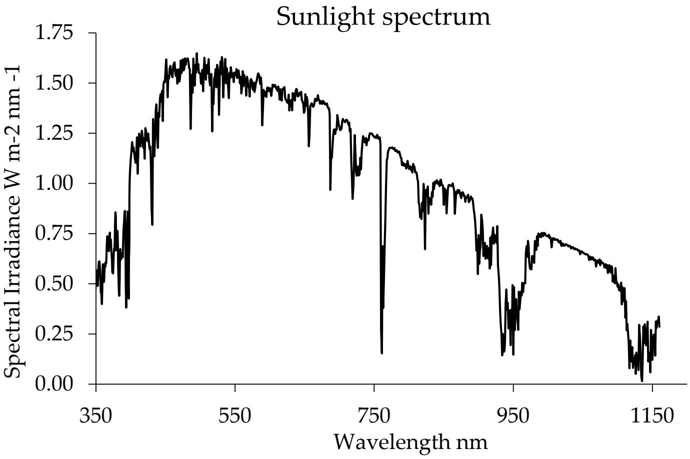


Fig. 8 Reference sunlight spectrum ASTM G-173, the global total spectral irradiance on the 37° sun-facing tilted surface [1]

3.3 Electronics

3.3.1 EEPROM and OTP

EEPROM and OTP are both types of memory that are used to store necessary camera information, *e.g.*, camera build information, calibration data, factory test data, etc.

EEPROM, Electrically Erasable Programmable Read-Only Memory, is a kind of non-volatile semiconductor memory, which means stored data will not be lost after power is turned off. OTP, One-Time Programmable, is a type of non-volatile memory that permits data to be written only once, whereas EEPROM allows data to be written and erased multiple times. OTP is generally lower cost, smaller size, and offers less memory compared with EEPROM. Some image sensors have OTP implemented.

3.3.2 SerDes

SerDes is a portmanteau of serializer and deserializer. SerDes is a functional block to transmit high speed image data. A serializer converts parallel data to serial for transmission, while a deserializer converts serial data back to parallel. For example, the raw image data from the camera sensor can be converted to serial data by a serializer for transmission. The transmitted data can be converted to parallel signal by a deserializer and then sent to the ISP.

3.4 Image Signal Processor (ISP)

An ISP is a powerful processor that provides many image processing functions and control functions. Many ISPs allow multi-channel input, which means they are able to control multiple image sensors and process their output raw image streams simultaneously.

In ADAS/AD, some cameras are designed for both human vision and computer vision applications (*e.g.*, integrated dash cameras), in which case multiple ISPs may be needed for different applications depending on the requirements of software applications.

4 Image Processing

Raw image data needs to go through a series of image processing functions, known as an image processing pipeline, in order to be converted into a video stream for applications. Calibration and tuning (also referred as ISP tuning) are two critical steps to achieve high image quality.

4.1 Image Processing Pipeline

The purpose of an image processing pipeline is to control camera operation and convert image sensor raw output into desired formats. A typical image processing pipeline is shown in Fig. 9. Please note that the specific image processing pipeline may vary. For example, the video encoding function is not included in some ISPs, and the sequence of the image processing pipeline functions may also be different. Since most automotive cameras are fixed focus, the auto-focus algorithm is not discussed in this section.

4.1.1 Black Level Subtraction

When there is no input light, the image sensor still outputs a constant level signal. The signal is called the black level or dark pedestal. Black level subtraction subtracts this black level from the raw output.

4.1.2 Decompanding

In the automotive field, most of the image sensors output 16- to 24-bit HDR images, while the SerDes function is normally only able to transmit 12-bit signal. Therefore, automotive sensors provide a function to compand higher bit linear raw images into nonlinear lower bit. The decompanding process decompands the nonlinear images into the original linear raw images.

4.1.3 Defective Pixel Correction (DPC)

Most CMOS image sensors contain some defective pixels for various reasons such as imperfections during manufacturing, transportation, and storage. Defective pixels increase during usage of time, especially in high temperature and other harsh conditions.

There are two kinds of defective pixels, static defective pixels and dynamic defective pixels. The locations of static defective pixels do not change with exposure while that of dynamic defective pixels vary with exposure.

The correct value of dynamic defective pixels can be interpolated by the values of surrounding pixels. Though static defective pixels can be corrected in the same way as dynamic defective pixels, they may also be calibrated during camera assembly. Some image sensors provide the functions of static DPC, which stores the information of

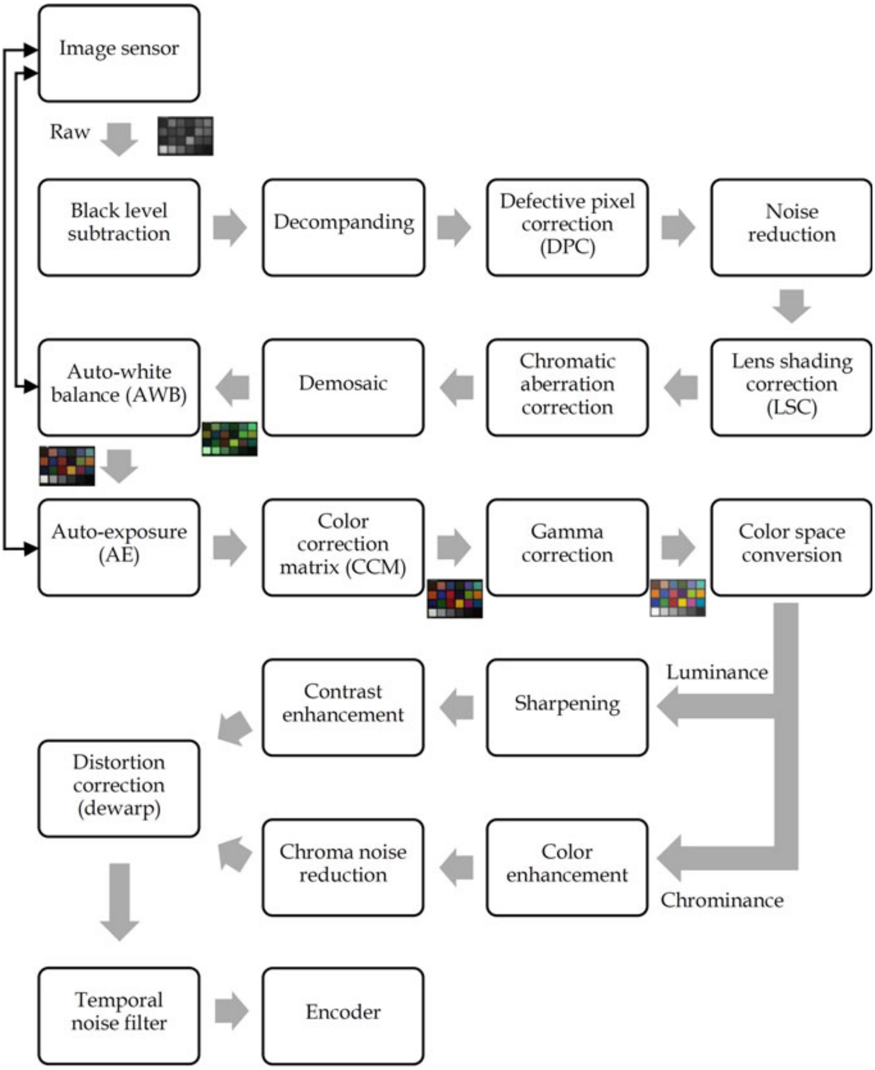


Fig. 9 A typical image processing pipeline

static defective pixels in the sensor OTP from the calibration process. Dynamic DPC function is also common in many image sensors. If the image sensor provides DPC, it may not be necessary to enable this function in the ISP.

4.1.4 Noise Reduction

Noise reduction may exist during many stages of an imaging pipeline. There are various methods to reduce different types of noise in an image, but the common ways more or less involve low-pass filtering, which results in a loss of image details. Therefore, the use of a noise reduction algorithm is usually a trade-off between noise reduction and loss of image details.

4.1.5 Lens Shading Correction (LSC)

A lens shading correction (LSC) algorithm compensates for shading of the optical system in order to generate an image with uniform illumination. There are typically two methods to perform LSC.

The first method is to calibrate the camera shading during camera assembly by capturing flat field images and recording the shading table in the camera EEPROM. Shading can be corrected by using the shading table.

The second method does not use any calibration but utilizes the low frequency portion of the image to dynamically correct shading.

Both methods provide good results. The first method requires an additional calibration process during camera assembly while the second method costs more computational resources.

4.1.6 Chromatic Aberration Correction

Since chromatic aberrations mostly affect edges of the image, most chromatic aberration correction algorithms first detect color fringes on edges and then correct the shifts in different color-channels.

4.1.7 Demosaic

Demosaic is a process to interpolate RGB images from raw images. Prior to the demosaic process, each pixel only represents a single color-channel. After the demosaic process, each pixel will have three color-channels (red, green, and blue).

A simple demosaicing approach is bilinear interpolation, in which the three color-channels are independently interpolated using symmetric bilinear interpolation from the nearest neighbors of the same color. Bilinear interpolation generates significant artifacts across edges and other high-frequency content, since it does not perform edge detection and takes into account the correlation among the RGB values. More advanced demosaicing algorithms take such correlation into account, either with better linear filters or with nonlinear filters that adapt interpolation smoothness to a measure of image activity or edginess. Later, research suggests a sequential demosaicing approach, which interpolates the luminance channel first and then reconstructs the chrominance channels based on recovered luminance information.

4.1.8 Auto-White Balance (AWB)

Auto-white balance (AWB) is a function to automatically adjust the mixture of color-channels in an image to establish a reference to white and gray colors. Most AWB algorithms follow a two-step approach: first estimate the illuminant of the scene and then adjust the color-channels (usually applies digital gains to the red and blue channels) to generate a new image where white and gray colors match the estimated illuminant. There are various AWB algorithms using different approaches.

Gray World Assumption

The gray world assumption assumes the average reflectance of a scene is achromatic, which means the mean of red, green, and blue channels in a given scene should be roughly equal. Therefore, we can balance the average values of the three color-channels to restore white and gray colors. The gray world assumption works well in most cases, except in situations when a certain color dominates the scene, such as a blue sky or a green forest.

White World Assumption

The white world assumption is based on the retinex theory of visual color constancy: Perceived white is associated with the maximum cone signals. In practice, the brightest point in an image is commonly caused by a direct reflection from a glossy surface. Such a reflection tends to reflect the color of the light source. The white world assumption also works well in practice, except when there are a few pixels with very large values in a single color-channel.

Other Algorithms

Most AWB algorithms more or less take the two abovementioned assumptions into account and cover additional scenarios. For example, illuminant voting considers the scenario when there are multiple illuminants in the scene, whereas iterative white balancing uses a local method when the scene is dominated by certain colors.

4.1.9 Autoexposure (AE)

The autoexposure (AE) function automatically determines the amount of light that falls onto the image sensor and adjusts the sensor's exposure time and gain to achieve a target brightness level at a preset region of interest (ROI). The preset ROI can be the center of the image, the whole image, or some specific regions of the image. More advanced AE algorithms use a weighted average of multiple zones as the target brightness level. Since most automotive cameras are fixed aperture cameras, the AE function mainly adjusts the integration time (or exposure time) and the gain of the image sensor.

4.1.10 Color Correction Matrix (CCM)

Even though image sensors can sense different colors, the spectral sensitivity of each color-channel normally differs from human eyes. The color correction matrix (CCM) is a 3×3 matrix used to correct the color differences caused by the spectral sensitivity difference so that the colors generated by the cameras are close to human eye response. The CCM is typically calibrated by acquiring pictures of calibrated color charts under standardized light sources.

4.1.11 Tone Mapping

In the automotive field, most image sensors can output 16- to 24-bit HDR images. However, most encoders can only output 8- to 12-bit per color-channel. Such bit depth degradation frequently results in loss of image details. Tone mapping is an image processing technique to preserve image details due to bit depth loss by changing the linear pixel response to nonlinear.

There are two types of tone mapping: global and local. Global tone mapping, also known as tone curve, is spatially uniformly applied to the whole image, while local tone mapping applies different mapping to each pixel usually based on the features extracted from adjacent pixels.

Common global tone mapping algorithms include the sigmoidal ("S-shaped") tone curve, which preserves less details near saturation and dark regions but more contrast at mid-illumination; histogram equalization, which spreads out the most frequent pixel intensity values to utilize the full digital range; and logarithmic mapping, which imitates the human response to light by reducing the contrast ratio through a logarithmic compression of luminance values.

4.1.12 Gamma Correction

Image sensors produce a linear response to the input light. However, the human perception of brightness is not linear. Therefore, the brightness of images is encoded in a nonlinear manner, normally using a power-law relationship shown in Eq. (2) below:

$$V_{\text{out}} = A \cdot V_{\text{in}}^{\gamma} \tag{2}$$

where V_{out} is the output value, A is a constant, V_{in} is the input value, and γ is the gamma value.

The gamma correction function is used to compensate the display gamma encoding so that the image response is still linear during processing. Therefore, the gamma value used in this function varies by the display system. Most display systems use a gamma value of 0.45. To compensate this value, we need to use a gamma value of 1/0.45 = 2.2.

Note that the gamma correction function is designed for human vision applications. This function is not necessary for computer vision applications.

4.1.13 Color Space Conversion

A color space is a notation to specify colors in a quantified manner. The RGB space is one of the most widely adopted color spaces. Some image processing functions require converting the images to another color space, usually to a space that uses luminance channel (represent brightness information) and chrominance channels (represent color information). Popular color spaces for this purpose are YUV, YCbCr, and CIE Lab.

4.1.14 Sharpening

The purpose of image sharpening is to enhance fine details in an image. In the spatial frequency domain, a sharpening filter emphasizes the high-frequency portion of the image; therefore, sharpening is often referred to as high-pass filtering.

Basic sharpening algorithms include first-order derivative, second-order derivative (Laplacian), and unsharp masking, which subtracts a low-passed (smoothed) image from the original image.

4.1.15 Color Enhancement

Research has shown that an image that represents the real scene may not be preferred by humans. For example, humans typically prefer colors with enhanced saturation and brightness. Color enhancement modifies the images so that they are more visually pleasing to humans. This block is traditionally designed for human vision and is normally not recommended for computer vision applications.

4.1.16 Chroma Noise Reduction

As mentioned in Sect. 4.1.4, a common side effect of most noise reduction algorithms is loss of details. One common way to solve the side effect is to separate the image into luminance channel and chrominance channels; then, we can use the luminance channel to preserve the details but apply noise reduction on the chrominance channels.

4.1.17 Distortion Correction (Dewarp)

Distortion correction, also known as geometric lens distortion correction, dewarping, or image rectification, is commonly realized by imaging a calibration target with known geometry (*e.g.*, a checkerboard chart or a chart with evenly spaced dots), extracting coordinate information from pre-defined landmarks, and then applying a lens distortion model to correlate geometric information from the image to ground truth.

Most lens distortion models categorize lens distortion into two categories: radial distortion and tangential distortion. Radial distortion causes image coordinates to shift from the ideal position inwards or outwards from the optical center, which is the imaginary point where the optical axis intersects with the image plane. Tangential distortion causes image coordinates to shift in the direction perpendicular to the line connecting them to the optical center. Note that optical center is a term used to describe the point with maximum illumination as well.

Radial distortion is the more common type of lens distortion. Therefore, the most basic models only correct radial distortion. In a radial distortion model, image coordinates are first converted from Cartesian coordinates into polar coordinates. Then, either a polynomial fit (known as polynomial radial distortion model) or a divisional polynomial fit (known as division polynomial radial distortion model) is used to convert between distorted and undistorted radii.

More advanced models consider both radial and tangential distortion. A common model of this type is the Brown-Conrady model.

4.1.18 Temporal Noise Filter

In automotive applications, the frames of a video stream are normally temporally correlated. A temporal noise filter achieves high noise attenuation by exploiting the correlations among adjacent frames. Common artifacts of a temporal noise filter include ghosting artifacts and motion blur. Therefore, motion compensation is normally applied to reduce the artifacts. More advanced algorithms combine both spatial and temporal denoising.

4.1.19 Video Codec

A video codec is software or hardware that encodes and decodes digital video streams. Codec is a portmanteau of encoder and decoder. A video encoder converts raw (uncompressed) video to a specific format and compression rate for transmission.

A video decoder decompresses the video steam. Decoders are usually in the display system for human vision or in the software application for computer vision.

There are numerous video codecs available. Broadly adopted codecs include but not limited to H.264, H.265, VC-1, VP9, etc.

4.2 Calibration

4.2.1 General Camera Calibration

There is unit-to-unit variation on camera raw images caused by variation in part (sensor, lens, PCB, etc.), quality, and the manufacturing process. The purpose of the calibration process is to generate consistent and accurate images.

Ideally, each camera can be calibrated to achieve the best image quality, and such calibration is called unit calibration. Unit calibration parameters are typically implemented at the camera manufacturing line or vehicle assembly line.

However, in production, due to constraints such as cost, space, and time, an alternative approach known as class calibration is often applied. In class calibration, a certain number of camera units are randomly selected from a large number of cameras. Assuming the randomly selected sample set represents the same statistical distribution of test parameters as all cameras, we can calibrate all the cameras in the sample set, run a regression of the test parameters on all the cameras, find a set of parameters that generate lowest errors overall for all cameras in the sample set, and apply such set of parameters to all cameras.

In production, most camera parameters go through class calibration. Unit calibration is only applied to camera parameters with high accuracy requirements.

4.2.2 Geometric Calibration

The purpose of geometric calibration may include correcting for lens distortion, converting pixel coordinates in an image to real-world units, determining the relative position of the camera in the system, or detecting the location of the camera in the scene. Though there are many aspects of the cameras to be calibrated, many people in the ADAS/AD field specifically refer to camera calibration as geometric calibration because it is one of the most important camera calibrations. Unit calibration is usually applied to geometric calibration.

A typical camera calibration procedure involves capturing images by aiming the camera at a test chart with a known pattern, *e.g.*, evenly spaced dots and checkerboard.

The test chart is placed and rotated to cover various field positions, distances, and orientations. Computer vision algorithms are applied to the test images to generate extrinsic parameters, which transfer 3D world units to the camera's 3D coordinates, and intrinsic parameters, which project the camera's 3D coordinates into the 2D pixel coordinates in an image.

Various geometric calibration algorithms exist based on different camera models. The most popular camera model is the pinhole camera model, which simplifies the lens to be a single small pinhole. Light rays from the scene pass through the pinhole and create an inverted image. The pinhole camera model cannot model extreme distortion, which typically exists in a wide FOV camera. A popular camera model that models wide FOV cameras is called fisheye camera model, which treats the camera as an omnidirectional vision system providing a 360° panoramic view of the scene.

4.3 ISP Tuning

ISP tuning, or tuning, is the process of adjusting the parameters in an image processing pipeline to achieve optimized image quality. An ISP tuning process generally includes camera characterization, class calibration, and image quality optimization.

In the camera characterization process, a series of tests are applied to the camera module to understand camera performance as well as the performance variation among cameras.

The purposes of class calibration during the ISP tuning are to reconstruct the real-world scene and minimize image quality variation among cameras. Many functions in the image pipeline require calibration to achieve the best image quality, for example, color correction and lens shading. Any required calibrations that do not undergo unit calibration will need to go through class calibration.

Image quality optimization is tuned for specific applications because of the fact that the definition of best image quality varies upon applications; for example, human vision applications prefer highly saturated colors, which normally differs from the real scene. Another purpose for the image quality optimization step is to find the optimal trade-off among image processing function blocks. For example, the cost of most noise reduction algorithms is lower sharpness, while the sharpening function may increase noise. For human vision applications, humans usually prefer high sharpness images but tend to ignore some noise. As a result, sharpness receives higher weight. On the contrary, many computer vision algorithms may prefer lower noise but are less sensitive to sharpness. In this case, noise reduction is more important.

5 Camera Product Development

It takes multiple steps to develop a camera product. As shown in Fig. 10, a camera product life cycle includes product definition, design, prototype, validation, manufacturing, implementation, and support. Note that the abovementioned process is recursive instead of strictly sequential. For example, product design is guided by product definition, which may change as a result of feasibility issues found in the design stage. Product design may need to be revised because of improvements suggested during all later stages.

5.1 Product Definition

Product definition is a critical step to ensure the success of the product from the business perspective. At this stage, certain information should be passed down from the vehicle product definition level. For example, what is the type of the vehicle (a sedan, an SUV, a pickup truck, etc.)? Who are the target consumers? What features do we plan to support (ADAS, driver monitoring, gesture control, etc.)? Which tier of the market (high-tier, mid-tier, or low tier) is the vehicle targeting at? When do

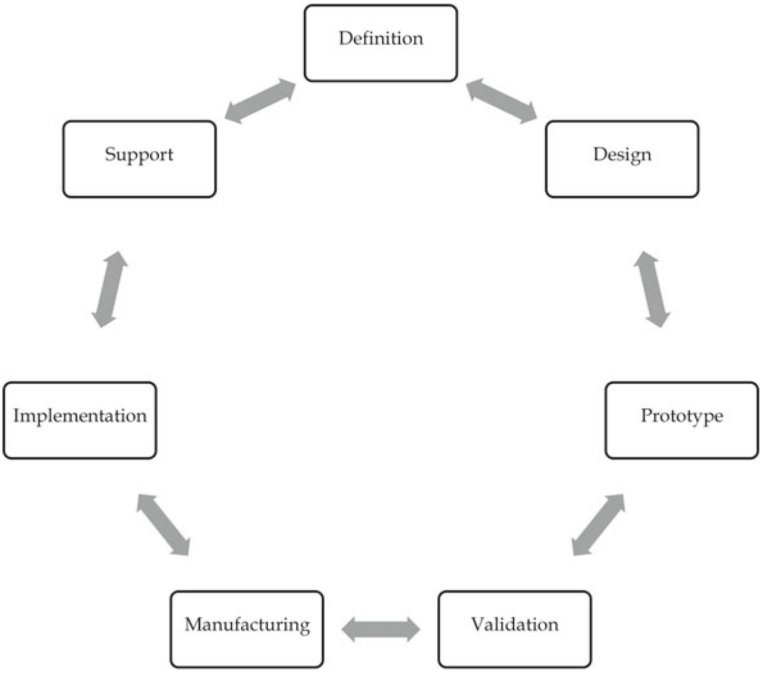


Fig. 10 Camera product life cycle

we need to deliver the product? Such information will eventually be converted to the component level as functionality, target performance, cost, schedule, and resource allocation for the planning of the camera product.

5.2 Camera Design

There are three main stages in camera design: camera system architecture, subsystem design, and component selection. Like the camera product development process, the three camera design stages are recursive as well. At each stage, we can choose whether to perform in-house design or purchase off-the-shelf systems or components. In this chapter, we describe a typical in-house design process.

5.2.1 Camera System Architecture

During camera system architecture, major camera system specifications are derived from product definition and vehicle level requirements. Such specifications should include but not limited to mechanical constraints (*e.g.*, size, weight, mounting position and angle, operation, and storage temperature), electric constraints (*e.g.*, power consumption, frame rate, voltage, interface, synchronization with other components such as Lidar, radar, and ultrasound), optical constraints (*e.g.*, lighting condition, optical resolution, FOV, operating range in distance, spectra range), reliability specifications (*e.g.*, EMI and EMC requirements, product lifetime, safety compliance, environmental conditions), and image quality requirements (*e.g.*, sharpness, SNR, color accuracy, artifacts, dynamic range).

5.2.2 Camera Subsystem Design

The purpose of camera subsystem design is to convert camera system specifications into subsystem specifications, which typically includes the number of cameras and ISPs needed in the camera system, specifications for camera module and ISP, and the interface between camera modules and ISPs. Preliminary design and simulation are preferred at this stage.

The ISP can be a dedicated hardware ISP or an electronic control unit (ECU) for a software ISP. Hardware ISPs are generally more efficient but with less flexibility compared with software ISPs. The ISP may be packaged into the same camera module or on a separate board because of various design considerations. Camera module driver and image processing functions needed should also be considered at this stage.

5.2.3 Component Selection

Camera system specifications are further converted into specifications of each component. The best practice is to characterize each component prior to the selection as the component specifications from different vendors may not align due to differences in concept definition and test methods.

Image sensors and ISPs are usually first selected because of their importance in a camera system. A lens matching the image sensor is typically the next component to be determined. The width and length of the sensor and the total track length (the span between the top of the lens surface and the image sensor, also shortened as TTL) together determine the minimum dimensions of the camera module.

Note that the interactions among components should also be considered at this stage. For example, reflection between the image sensor and the optical filter is one of the main factors contributing to flare. Therefore, the best practice is to optimize the reflectance spectrum of the optical filter to cancel light reflected from the image sensor.

5.2.4 Recursive Process

Once all components are selected, detailed design and simulation are performed for camera system review. Further modifications in design and component selection may be involved. After camera system review is accepted, the next step is to build prototypes for further design verification.

5.2.5 Design Considerations for Different Applications

Specifications of cameras are highly dependent on the applications of the cameras.

Human Vision versus Computer Vision

For human vision applications, the target is creating visually pleasing images, which normally means high sharpness, high color saturation, and no noticeable artifacts. For computer vision applications, image quality requirements depend on the requirements of software applications. Generally speaking, software applications are usually more sensitive to noise and artifacts compared with human vision applications. As the camera system advances, there is a trend to utilize the same camera system for both human vision and computer vision applications. For example, e-mirrors are traditionally designed only for human vision applications. Such camera system can also be utilized for object detection, which is a computer vision application.

Synchronization

Synchronization requirements may be needed for camera systems that involve multiple cameras (*e.g.*, surround view cameras, e-mirrors, and outward facing cameras for ADAS/AD). One common requirement is that all cameras should be synchronized to start capturing images at the same time. For a stereo vision system that involves two cameras, both cameras not only need to be synchronized to start capturing, but also need to be synchronized to apply the same autoexposure, HDR, and auto-white balance settings.

Lighting Condition

Lighting condition is another factor worth consideration. Outward facing cameras can take advantage of the headlights to illuminate the scene. In-cabin cameras may need active illumination in order to operate at low-light conditions. Therefore, many cameras designed for a driver-monitoring system (DMS), gesture control system, or an in-cabin monitoring system are illuminated with infrared emitters.

Frame Rate

30 fps (frame per second) frame rate is sufficient for most human vision application and outward facing cameras for ADAS/AD applications. However, e-mirrors normally require much higher frame rate in order to capture the high angular motion arisen from vehicles passing from the side. Some cameras designed for a driver-monitoring system (DMS) are used to capture driver's eye gaze. In this case, high frame rate (usually 60 fps) is necessary for capturing the rapid eye movement. For gesture control systems, a higher frame rate usually results in a more seamless user experience, and therefore, a higher frame rate is preferred but not necessary.

Reliability

Generally speaking, outward facing cameras require higher ingress protection level (protection provided by mechanical casings and electrical enclosures against intrusion, dust, accidental contact, and water) and stiffness compared with in-cabin cameras. Cameras designed for safety-related applications need to comply with a higher level of safety standards. A design failure mode and effects analysis (DFMEA) is critical at the design stage to address potential system failures.

Table 3 Comparison of different 3D cameras

	Passive stereo vision	Active stereo vision	Time-of-flight
Low light	Poor	Excellent	Excellent
Sunlight	Excellent	Excellent	Good
Cost	Low	High	High
Size	Large	Largest	Small
Depth precision	Low	Low	High
Spatial resolution	High	High	Low

3D Cameras

Some software applications require 3D information (*e.g.*, many gesture control cameras require 3D coordinates of the hands). Common 3D cameras in automotive applications include stereo vision cameras, which utilize two cameras to simulate human binocular vision to interpret depth information, and time-of-flight (ToF) cameras, which detect depth information by measuring the traveling time between light emission and return.

A comparison of different 3D cameras is shown in Table 3. Compared with ToF cameras, passive stereo vision cameras (without IR emission) are generally less expensive with good sunlight robustness and high spatial resolution but do not perform well under low light, and the depth precision is relatively low. Active stereo vision cameras are stereo vision cameras with IR illumination. Active stereo vision cameras work well under low light with added cost and size. ToF cameras provide the best depth precision, the smallest size, excellent low-light performance, and good sunlight robustness but with higher cost and lower spatial resolution.

Other Considerations

The mounting locations, tilting angles, FOVs, DoFs, and pixel resolutions of the cameras are critical for covering the objects of interests within the desired region. For example, for outward facing cameras for ADAS/AD applications, it is critical to design the abovementioned specifications to cover objects such as pedestrians, vehicles, and traffic signs.

5.3 Prototype

Multiple stages of prototypes are needed to turn a camera design into manufacturing. Though the number of prototype stages varies because of various reasons, generally speaking, the early prototype stages tend to verify the engineering design while the later prototype stages tend to verify the manufacturing process.

5.4 Validation

Validation is a critical step after each stage of prototypes to verify the engineering design and the manufacturing process. Learnings from the validation results should be converted to further improvements and implemented in the next stage of prototypes. Reliability tests are crucial in automotive applications as part of the validation process.

5.5 Manufacturing

Prior to camera assembly, all electrical components, image sensors, and optical emitters (if there are any) are mounted on the PCB boards in PCB fabrication houses. All camera components and customized components should arrive at the factory. Incoming quality check (IQC) should be performed to ensure quality of components.

After IQC, all components should be tested and cleaned at the manufacturing line in order to be ready for final assembly. One important step during camera assembly is to align the optical axis of the lens with the center axis of the image sensor. For optimal results, an active alignment (AA) process is recommended. During AA, there is a specially designed MTF test chart facing the image sensor. Live images are streamed to a computer to calculate MTF real-time. A robotic arm holds the lens and glues the lens to the camera once the optimal lens position is found.

After assembly is finished, the camera will be sent for calibration and outgoing quality check (OQC). Calibration data and pre-defined camera component and manufacturing data will be written to the EEPROM or OTP.

Manufacturing quality is crucial to the final image quality for several aspects: First, the accuracy of calibration data and OQC tests performed during manufacturing directly affects image quality. Second, the consistency of manufacturing process contributes to the consistency of camera quality, which further influences errors generated from class calibration. Third, artifacts may be introduced through manufacturing imperfections. For example, foreign particle intrusion may cause flare, blemish, and MTF degradation.

For optimal performance, camera manufacturing should be performed in a clean room with consistent temperature and humidity levels. A process failure mode effects analysis (PFMEA) is crucial at this stage to address potential failures that are rooted in the manufacturing process.

5.6 Implementation

The implementation step typically includes camera driver development, ISP tuning, and supporting vehicle assembly. Certain calibrations may need to be performed during vehicle assembly. For example, outward facing ADAS/AD cameras usually need to be calibrated for their relative positions to the vehicle and to other sensors (*e.g.*, Lidar, radar).

5.7 Support

Once vehicles are sold to the customers, after-sales support may be needed. Learnings on common camera issues should be considered for future camera development. Certain camera features (*e.g.*, ISP tuning, driver) may be improved through over-the-air (OTA) updates.

6 Summary

This chapter introduces cameras in ADAS/AD vehicles. Fundamental knowledge of camera system, camera hardware, and image processing are presented. A typical camera product development process is discussed. With the rapid advancement of the ADAS/AD field, more cameras will be used in ADAS/AD vehicles. New types of camera applications may occur in the future.

Acknowledgements The author is especially grateful to John Pincenti for his constructive criticisms and for contributing to Figs. 4, 6, and 7. The author would also like to express his gratitude to Rebecca Emmert and Christy Fernandez-Cull for their valuable comments and corrections. The editors would like to thank Dr. Charles Odegard from Texas Instruments for his dedicated review of the chapter.

References

1 Gueymard, C. A.., Myers, D., & Emery, K. (2002). Proposed reference irradiance spectra for solar energy systems testing, *Solar energy*, 73(6), 443–467.
2. Nakamura, Junichi (2005). *Image Sensors and Signal Processing for Digital Still Cameras*. CRC Press. ISBN 978-0849335457
3. Cressler, John D. (2017). *Silicon Earth: Introduction to Microelectronics and Nanotechnology*, Second Edition. CRC Press. ISBN 978-1498708258
4. Fossum, Eric *et al.* (2014). *A Review of the Pinned Photodiode for CCD and CMOS Image Sensors*. IEEE Journal of the Electron Devices Society. 2 (3): 33–43.
5. Durini, Daniel (2014). *High Performance Silicon Imaging: Fundamentals and Applications of CMOS and CCD sensors*. First Edition. Woodhead Publishing. ISBN 978-0857095985

6. Ohta, Jun (2007). *Smart CMOS Image Sensors and Applications (Optical Science and Engineering),* First Edition. CRC Press. ISBN 978-0849336812
7. Weikl, Korbinian (2020). *Optimization of automotive color filter arrays for traffic light color separation.* Society for Imaging Science and Technology. Volume 2020, Number 28, pp. 288–292(5)
8. Janesick, James (2007). *Photon Transfer.* SPIE--The International Society for Optical Engineering. ISBN 978-0819467225
9. Gonzalez, R. C., & Woods, R. E. (2002). *Digital image processing.*
10. Jahne, B. (2005). *Digital image processing.* Sixth Edition Springer. ISBN 978-3540240358
11. Petrou, M. P., Petrou, C. (2010). *Image Processing: The Fundamentals.* 2nd Edition. Wiley. ISBN 978-0470745861
12. Li, Xiangli (2008), *MOSFET modulated dual conversion gain CMOS image sensors*, Boise State University
13. Catrysse, Peter *et al.* (2000). *QE Reduction due to Pixel Vignetting in CMOS Image Sensors.* Proceedings of SPIE, vol. 3965
14. Li, Yuanzhen *et al.* (2005). Compressing and companding high dynamic range images with subband architectures. ACM transactions on graphics (TOG), 24(3), 836–844.
15. Tanbakuchi, Anthony A., et al. *Adaptive pixel defect correction.* Sensors and Camera Systems for Scientific, Industrial, and Digital Photography Applications IV. Vol. 5017. International Society for Optics and Photonics, 2003.
16. Yongji, Liu, and Yuan Xiaojun. "A Design of Dynamic Defective Pixel Correction for Image Sensor.“ 2020 IEEE International Conference on Artificial Intelligence and Information Systems (ICAIIS). IEEE, 2020.
17. Young, Ian T. "Shading correction: compensation for illumination and sensor inhomogeneities.“ Current Protocols in Cytometry 14.1 (2000): 2–11.
18. Cecchetto, Benjamin T. "Correction of Chromatic Aberration from a Single Image Using Keypoints.“ arXiv preprint arXiv:2002.03196 (2020).
19. Lukac, R. (Ed.). (2018). *Single-sensor imaging: methods and applications for digital cameras.* CRC Press.
20. Longere, Philippe, et al. "Perceptual assessment of demosaicing algorithm performance.“ Proceedings of the IEEE 90.1 (2002): 123–132.
21. Ramanath, Rajeev, et al. "Demosaicking methods for Bayer color arrays.“ Journal of Electronic imaging 11.3 (2002): 306–315.
22. Malvar, H. S., He, L. W., & Cutler, R. (2004, May). High-quality linear interpolation for demosaicing of Bayer-patterned color images. In 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (Vol. 3, pp. iii–485). IEEE.
23. Li, Xin, Bahadir Gunturk, and Lei Zhang. "Image demosaicing: A systematic survey.“ Visual Communications and Image Processing 2008. Vol. 6822. International Society for Optics and Photonics, 2008.
24. Land, E. H. (1964). The retinex. American Scientist, 52(2), 247–264.
25. Tkalcic, M., & Tasic, J. F. (2003). Colour spaces: perceptual, historical and applicational background (Vol. 1, pp. 304–308). IEEE.
26. Ledda, P., Chalmers, A., Troscianko, T., & Seetzen, H. (2005). Evaluation of tone mapping operators using a high dynamic range display. ACM Transactions on Graphics (TOG), 24(3), 640–648.
27. Camgöz, N., Yener, C., & Güvenç, D. (2002). Effects of hue, saturation, and brightness on preference. Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur, 27(3), 199–207.
28. Prescott, B., and G. F. McLean. "Line-based correction of radial lens distortion.“ Graphical Models and Image Processing 59.1 (1997): 39–47.
29. Lee, S. W., Maik, V., Jang, J., Shin, J., & Paik, J. (2005). Noise-adaptive spatio-temporal filter for real-time noise removal in low light level images. IEEE Transactions on Consumer Electronics, 51(2), 648–653.

30. Boo, K. J., & Bose, N. K. (1998). A motion-compensated spatio-temporal filter for image sequences with signal-dependent noise. IEEE Transactions on Circuits and Systems for Video Technology, 8(3), 287–298.
31. Zhang, Z. (2000). A flexible new technique for camera calibration. IEEE Transactions on pattern analysis and machine intelligence, 22(11), 1330–1334.
32. Heikkila, J., & Silvén, O. (1997, June). A four-step camera calibration procedure with implicit image correction. In Proceedings of IEEE computer society conference on computer vision and pattern recognition (pp. 1106–1112). IEEE.
33. Bradski, G., & Kaehler, A. (2008). Learning OpenCV: Computer vision with the OpenCV library. " O'Reilly Media, Inc.".
34. Scaramuzza, D., Martinelli, A., & Siegwart, R. (2006, October). A toolbox for easily calibrating omnidirectional cameras. In 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (pp. 5695–5701). IEEE.

Lidar Technology

Yufeng Hou and Zuoming Zhao

Abstract In the 2005 DARPA Grand Challenge, Stanford's Stanley robot car, navigated by 5 roof-mounted SICK Lidars, won the race against 4 other competitors. Since then, Lidars have gradually become crucial perception sensors for autonomous driving and ADAS, due to their ability to generate real-time point clouds with accurate 3D information of the vehicle's surroundings. Extensive research efforts have been invested into Lidar technology in both academia and the industry. As a result, a diverse variety of Lidar sensors have been created in the past decade. In this chapter, the authors aim to review the state of the art of Lidar sensors for autonomous driving or ADAS applications. The manuscript discusses the important metrics for Lidar sensor performance: detection range, field of view (FOV), angular resolution, frame rate, and eye safety. Then, different Lidar mapping methods and distance calculation mechanisms are discussed. Current status of mechanical, MEMS, FLASH, optical phased array (OPA), and frequency-modulated continuous wave (FMCW) Lidars is introduced, and their pros and cons and reliability performance are compared.

1 Introduction

Light detection and ranging (Lidar) provides the 3D depth information by sending out the laser pulse and measuring the pulse reflected back from objects. Since the first demonstration of Stanford's Stanley with 5 roof-mounted Lidars during DARPA's Grand Challenge, the Lidar has been a critical component for autonomous driving. All the following Grand Challenges proved the necessity of Lidar for autonomous navigation. All major players targeting Level 4 and above autonomous driving technology have equipped Lidars to get reliable 3D information from the surroundings. Multiple companies have successfully developed the Lidar for the autonomous driving market.

Y. Hou (✉)
DiDi Labs, Fremont, CA, USA
e-mail: yufenghou@didiglobal.com

Z. Zhao
Argo AI, LLC., San Jose, CA, USA

Y. Li and H. Shi (eds.), *Advanced Driver Assistance Systems and Autonomous Vehicles*,
https://doi.org/10.1007/978-981-19-5053-7_8

There are Lidars to cover the different distance ranges up to 250 m for autonomous systems.

Short-range Lidar covers the range up to 50 m or shorter. It normally works with a near field camera to detect and track objects nearby, like pedestrians, cyclists, and blind spots. It provides extra safety detection where cameras have difficulty to determine the distance.

Long-range Lidar covers the range from 50 m to a few hundred meters. Sometimes, the range between 50 and 100 m is also called medium range. The Lidar in this range is critical for an autonomous vehicle to be able to operate in high-speed freeway condition.

2 Overview of Current Lidar Technology for Automotive Application

Lidar has been used in autonomous driving vehicles by many companies, like Waymo, Argo, Cruise, and Aurora. It has been proved to provide additional spatial information for the surrounding driving environment. Since the first demonstration of autonomous driving with Lidar, many types of Lidar have been developed. These Lidars with different designs and principles will be discussed in the following sections of this chapter. For the automotive applications, Lidar provides 3D depth information to the autonomous driving system. It helps the autonomous driving system accurately identify the objects, track movement, and make accurate decisions for actions. In order to cover an all 360° view from a few meters to a several hundred meters distance, the Lidars for different ranges require different designs of optics, detectors, and lasers. Therefore, multiple Lidars are required for detection in short, medium, and long range.

3 Important Performance Metrics for Lidar

A Lidar needs many components to be able to generate 3D depth images. A transceiver, a beam steering device, and data processing ICs are the main building blocks. A transceiver has two key components, a receiver (Rx) and a transmitter (Tx). A transmitter generates the light pulse. It includes a laser or laser array, optics to configurate the beam, and electronics to drive the laser device. A receiver detects the photon that is reflected back from targets. It includes a photodetector, readout circuits, and optics to improve the detection efficiency. A beam steering device is used to steer the laser beam to form an image up to a 360° surrounding view. Data processing ICs process the data instantaneously to get the 3D information and transfer this information to the autonomous driving system to perform its next actions. Figure 1 shows a schematic transceiver beam getting sent to the target and reflected back from the

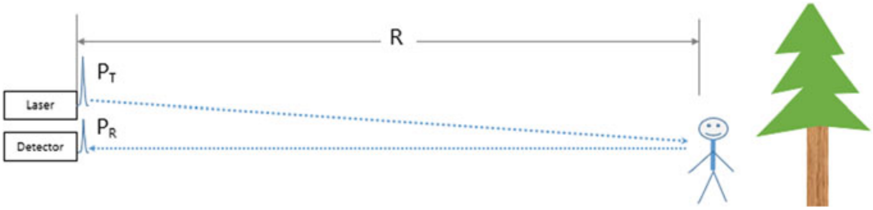


Fig. 1 Schematics of lidar detection

target. The light detected by the receiver can be expressed by Eq. 1, shown below [1].

$$P_R = P_T \frac{\sigma}{A_T} \frac{A_R}{\pi R^2} \eta_{\text{atm}}^2 \eta_{\text{sys}} \tag{1}$$

where P_R is the power received by the receiver, P_T is the transmission from the transmitter, σ is the cross section of the Lidar, A_T is the illumination area of the transmitter, A_R is the receiver area, R is the range between the Lidar and its target, η_{atm} is the transmission efficiency of the atmosphere, and η_{sys} is the transmission efficiency of the Lidar optical system. From this equation, we can conclude that the key figures of merit for the Lidar performance are range, FOV, laser power, receiver sensitivity, and angular resolution. Safety will be another factor that limits the operation of Lidar since it will limit the highest laser power that can be emitted to the target. This will also eventually limit the maximum range of a Lidar.

3.1 Range

With laser power capped by eye safety limitations, the range for a time-of-flight Lidar will be limited to the number of photons that can be detected by the receiver over the total noise. Based on Eq. 1, it can be concluded that the range will be related to the attenuation of the air, the optical efficiency of the system, and the sensitivity of the photodetector in the receiver for a given reflectance of a target. Lasers with a wavelength of from 800 to 1550 nm are used in most commercial Lidar applications. The eye safety limit varies with wavelength; the longer the wavelength, the higher the limit. Semiconductor lasers in this wavelength range are mostly III–V lasers. The photodetector is normally based on an avalanche photodetector (APD) due to its large gain. A silicon APD has a lot of advantages including its cost, integration, and process maturity. However, due to its indirect band-gap structure and 1.1 eV bandgap, it only provides reasonable sensitivity for wavelengths shorter than 1 um. For any wavelength beyond 1 um, a III–V APD will be required for good quantum efficiency.

With limited laser power and detector capability, the detector range will be strongly affected by the detector's optical design and atmosphere absorption. The solar spectrum through the atmosphere and extraterrestrial space from the American Society for Testing and Materials [2] is shown in Fig. 2. Several minimum absorption dips can be used to optimize the absorption loss. However, there will be a balance between absorption and background noise. Minimum absorption normally corresponds to a large background solar spectrum. For example, 940 nm has minimum background noise but it also has the highest absorption by water molecules. As shown in Fig. 2, there are significant advantages to using a short-wavelength IR where the solar background and absorption can be optimized. The band around 1400 nm is especially interesting because there is a wide area to optimize the absorption and background noise to achieve a large detection range. An extra advantage for the band around 1400 nm is the high cap on laser safety which will be discussed in a later section. On the other hand, the band around 940 nm has very mature III-V laser and Si-based photodetectors, which will give a big advantage for cost but less range. Currently, the commercially available Lidars which are using Si-based detectors only provide a good detector range up to 100 m with the reflectance around 10%. The long range beyond 100 m will need to be explored using the 1400 nm band.

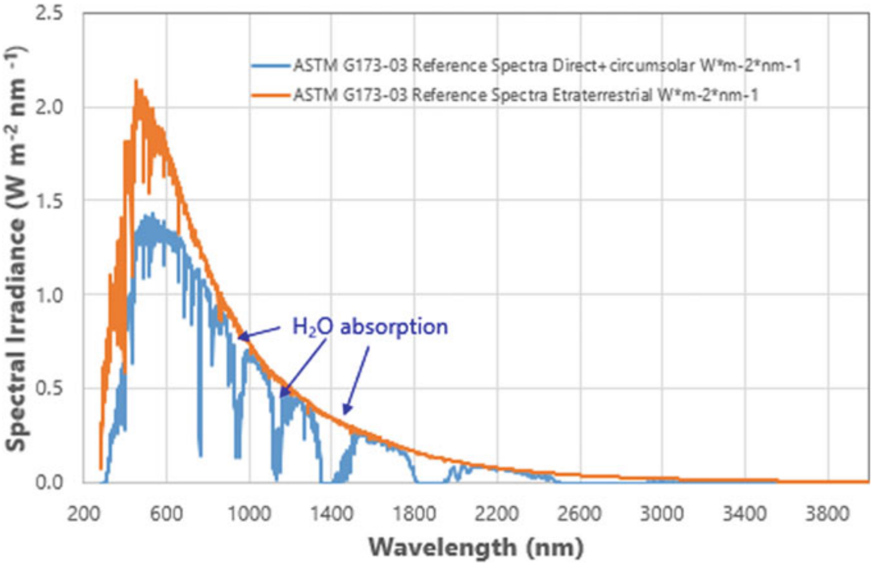


Fig. 2 Spectral irradiance from solar through atmosphere and extraterrestrial

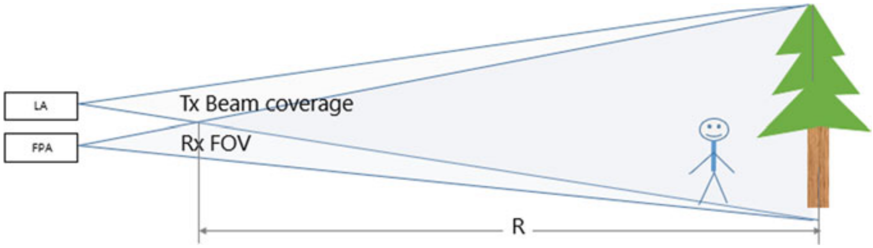


Fig. 3 Rx field of view and Tx beam coverage

3.2 Field of View

The field of view describes the maximum angle that can be viewed by a focal plane array (FPA). The larger the FOV, the wider the field the FPA can see. For an autonomous driving Lidar system, a larger FOV will give the detector larger detection areas to detect more objects. However, a larger FOV will also require larger beam lines to the field and a higher total laser power. As Fig. 3 illustrates, Lidar detects the points where a laser beam shines on. The Lidar can only detect the effective range *R* where there is sufficient overlap between the laser beam and the FPA FOV. The upper limit of range *R* will be limited by the signal-to-noise ratio of the system.

As the FOV increases, the beam coverage will also need to increase. The beam energy is quickly spread in the fast access direction. The photon reflected back to the detector will also quickly decrease. This will bring more challenges for FPA to have sufficient signal to maintain a good probability of detection on target. The large FOV will eventually eat in the margin on the maximum detection range. Therefore, it needs to balance on the range performance and FOV.

3.3 Angular Resolution/Accuracy

Angular resolution of Lidar is the smallest object that the Lidar can differentiate in angular direction where a beam is steered. It determines the capability of the Lidar to detect small objects in an angular direction. It will be reflected in the density of data points in a point cloud. Figure 4 shows the top view of the Lidar beam projection on an object. If the beam scan is between points 1 and 4, the higher the angular resolution, the more data points measured from the object. It will provide more information to the AV to detect and track an object.

The factors which determine the angular resolution are the detector sensitivity, beam steering accuracy, beam size, and FPA frame rate. The high detector sensitivity can provide advantages of small beam size and quick integration to detect the target. An accurate beam steering system will give better position accuracy in the angular direction. The small beam size will reduce the overlap between the lines and give a

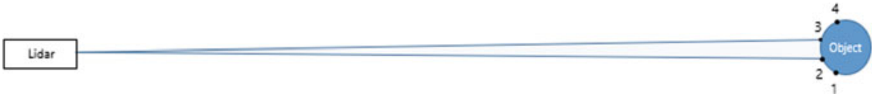


Fig. 4 Angular resolution of lidar

detector better signal-to-noise ratio. Finally, the fast frame rate is needed to process the data between angular lines. This requires the detector to have a sharp rise time and short duration after pulse to be able to differentiate each return pulse from targets.

3.4 Frame Rate

The Lidar frame rate is different from the detector frame rate. Lidar frame rate refers to the point cloud frame rate. That is determined by the beam steering frequency. These two rates will affect each other. The faster detector frame rate will give more room for Lidar frame rate which affects the point cloud quality.

The detector frame rate is the data processing speed of the FPA. It limits the Lidar's capability to detect moving objects. A high detector frame rate will allow Lidar to detect high-speed moving objects. The detector frame rate is mainly determined by the detector speed. For most APD, the ranging gate setting and readout IC time bin setting define the data process readout speed. Intrinsically, it is how fast the FPA can be armed.

The Lidar frame rate is the frame frequency of the full view of the point cloud. The beam steering system will limit the frame rate. The mechanical steering system will have a very limited frame rate around 10 Hz. Other advanced beam steering systems which will be discussed in the later sections can provide high frame rate.

3.5 Eye Safety

Eye safety defines the maximum laser exposure dose that is safe for the human eye. It uses maximum permissible exposure (MPE) to define the highest energy density of lasers that is still safe for the eye. It is usually about 10% of the dose that has a 50% chance of causing damage under worst-case conditions. The infrared light with a wavelength longer than 1400 nm is absorbed by the transparent parts of the eyes before it reaches the retina. This is the reason why there are orders of magnitude of more margin for eye safety when using long-wavelength lasers around the range of 1500 nm. Figure 5 shows the MPE at different wavelengths. From 1000 nm, there is a significant margin increase for longer wavelengths up to 1500 nm. However, the cost of the device will increase significantly due to the expensive III-V manufacturing process compared to Si process which can only work below 1000 nm.

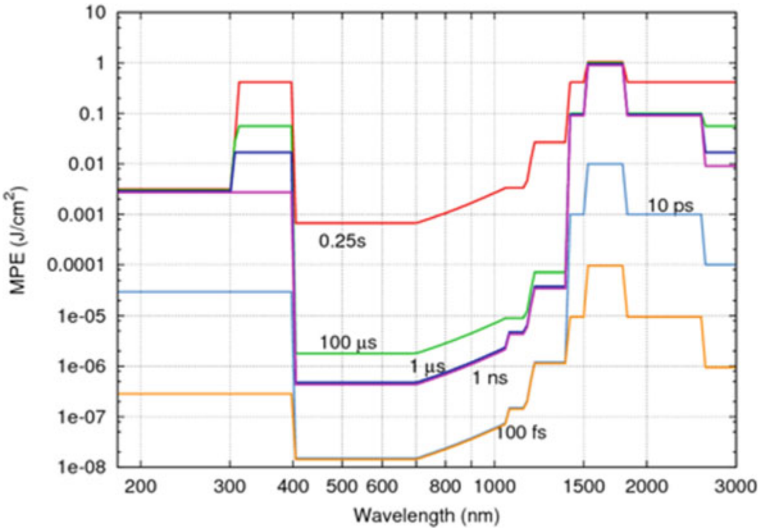


Fig. 5 MPE limit for different wavelengths [3]

4 Transmitter and Receiver

A transceiver is the module that generates the point cloud data in a Lidar. It has 3 main parts—a transmitter (Tx), a receiver (Rx), and the data process electronics. Tx is the module that emits the light pulse, and Rx is the module that detects the pulse. Lidars for autonomous driving are mostly based on semiconductors. Both photodetector arrays and laser arrays are semiconductor-based devices due to their small form factor and easy integration to data process electronics. Data process electronics are application-specific IC, FPGA, or SOC integrated with other voltage regulators or sensors on a PCB board. With different working wavelengths, the device design of semiconductors will be significantly different. Table 1 summarizes the pros and cons for different IR wavelengths. IR with Si-based Rx has a significant advantage on cost, while short-wavelength IR (SWIR) with III-V based Rx provides a large detection range which is crucial for long-range detection to enable the high-speed operation due to higher photon detection sensitivity in short-wavelength IR range.

5 Distance Calculation

The distance calculation of an AV Lidar is mostly based on the time-of-flight principle or phase shift of the light pulse. As Sect. 3 has discussed most of the figures of merit, this section will focus on distance calculation.

Table 1 The comparison for lidar applications between IR <1000 nm and SWIR >1000 nm

Metrics	IR <1000 nm	SWIR >1000 nm
Materials for Rx	Si-based Rx	III–V based Rx
Materials for Tx	GaAs-based Tx	InP-based Tx
Control/readout IC integration	Easy integration Rx with digital IC	Poor integration with digital IC
Cost of materials and process	Low cost for both Rx and Tx	High cost for both Rx and Tx
Reliability	Good reliability	Poor reliability
Manufacturing capability	Well established manufacture capability	More process development required
Development maturity	Good understanding with main failure modes	Limited research in reliability and degradation
Detector sensitivity	Low detector sensitivity for IR	High photon detection sensitivity
Detection range	Short detection range	Long detection range
Margin on eye safety	Low margin with small intensity limit	High margin with high intensity limit

5.1 Range of Time of Flight

For a time-of-flight sensor, the range distance can be expressed by Eq. 2, where R is the range distance for time of flight and t is the time for the light to reach the target and reflect back. The range is 15 cm per ns, and the accuracy of the range depends on the pulse shape.

$$R = t \times \frac{c}{2} \approx 15t \tag{2}$$

For the design to detect the edge of the pulse, the rise time of the pulse will strongly affect the uncertainty of the measurement. For the peak detection, the peak width will affect the accuracy. Current semiconductor lasers can achieve ~1 ns width pulse. The range accuracy can be within ~15 cm. However, the target reflectance will strongly modulate the peak shape. This will impact the uncertainty of detection due to the loss of signal-to-noise ratio (SNR) from transmission. Reflectance will also introduce range walk, which needs to be calibrated to avoid extra errors in range detection. The range walk is a systematic range error due to the pulse intensity change.

In order to accurately detect the target which is moving, there are two factors that need to be considered. One is the error introduced from the beam steering system. Another one is from target movement. The time of flight to a 250 m target is only 17 ns from Eq. 2. For a mechanical steering system with 10 Hz spinning rate, the error from the movement to the 250 m target can be calculated from Fig. 4 to be a quarter of millimeter. That is negligible. The movement target with speed of 120 mile/hr will only move around ~1 um. The error introduced by the movement of

Lidar is in the same order of magnitude of the target. This indicates that the effect due to beam steering, Lidar movement, and target movement will not introduce any significant impact on angular resolution.

5.2 Signal-To-Noise Ratio

As discussed previously in Sect. 3 Eq. 1, the detector needs to have enough power from the signal light to be reflected back from the target. The probability of detection will be determined by the signal-to-noise ratio (SNR). The signal is determined by the laser power, reflectance of target, and loss during transmission. The noise sources include detector noise, readout IC noise, or solar background noise. This can be expressed in Eq. 3, shown below.

$$\mathrm{SNR} = \frac{S}{N} \tag{3}$$

In Eq. 3, S is the signal power and N is the noise power. For a popular avalanche detector, the signal will be the photocurrent signal, which is directly related to the detector gain and responsivity for a given input light intensity. Noise can be from either external background or detection optics and electronics.

5.3 Factors that Affect Range Detection

In AV applications, the use conditions will affect the SNR of the Lidar. This will affect the probability of detection on the target. Weather conditions also have to be considered for Lidar development to ensure the device can remain functional in many corner cases, like rain, fog, hail, sandstorm, etc. Rain and fog will strongly scatter the light and reduce the SNR. For different systems, the scattering will be significantly different. The scattering from water drops, molecules, or other particles is inversely proportional to the fourth power of the wavelength and the range. It can be explained by Rayleigh scattering and Mie scattering in Eq. 4.

$$I = I_0 \frac{\left(1 + cos^2\theta\right)}{2R^2}\left(\frac{4\pi^2}{\lambda^4}\right)\left(\frac{n^2-1}{n^2+2}\right)^2\left(\frac{D}{2}\right)^6 \tag{4}$$

where I_0 is the light intensity before the interaction with the particle, R is the range distance between the particle and the observer, θ is the scattering angle, λ is the wavelength of light under consideration, n is the refractive index of the particle, and D is the diameter of the particle. The range detection capability will quickly decrease with the density of the raindrops. Short-wavelength Lidar will be impacted more significantly than long-wavelength ones. For very large particles, the direct reflection will take over and Lidar will lose visibility to the objects.

6 Future Direction of Lidar Developments

Although time of flight (ToF) is the current mainstream distance measurement method Lidars are using, new techniques such as frequency-modulated continuous wave (FMCW) are under development and have started to demonstrate advantages in some aspects compared to ToF [4].

6.1 Frequency-Modulated Continuous Wave (FMCW)

In a frequency-modulated continuous wave (FMCW) Lidar, the laser frequency emitted from the laser source, usually a diode to enable coherent detection [5], is modulated by a waveform generator to be varying periodically [6]. As indicated in Fig. 6, a beam splitter splits the transmitted laser into two parts. One part is projected onto the target and reflected back, while the other part goes directly to the mixer as a beat frequency and acts as a reference of the original laser in comparison with the first branch of the emitted wave. It will mix with the other branch of the emitted wave and create a beat frequency [7].

This beat frequency is proportional to the target distance [8–11]

$$f_b = \frac{4BRf_m}{c} \tag{5}$$

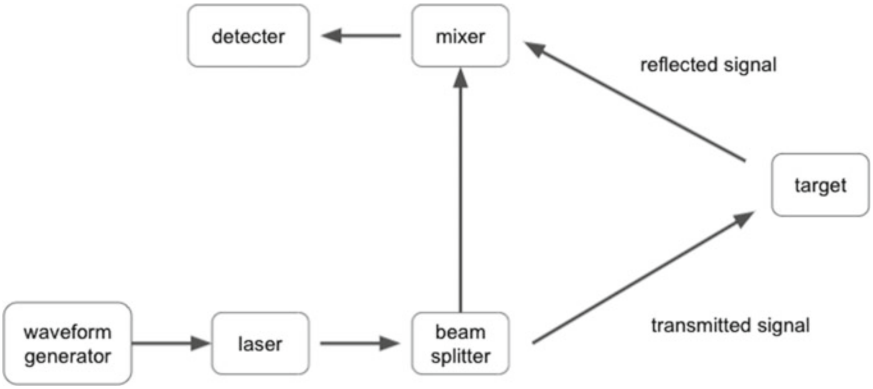


Fig. 6 Simplified FMCW lidar working principle

Here, f_b is the beat frequency, B is the bandwidth of the frequency sweep, R is the target distance, and f_m is the modulation frequency (of a triangular frequency modulation).

One advantage of the FMCW Lidar is that it has the capability of measuring not only the target's distance, but also the target longitudinal velocity and its direction. The velocity contributes an additional Doppler frequency f_d to the beat frequency f_b [12]:

$$f^+ = f_b + f_d \tag{6}$$

$$f^- = f_b - f_d \tag{7}$$

The target's distance is:

$$R = \frac{c}{8Bf_m}\left(f^+ + f^-\right) \tag{8}$$

The target's longitudinal velocity is:

$$v1 = \frac{\lambda}{4}\left(f^+ - f^-\right) \tag{9}$$

FMCW Lidar uses a continuous light source instead of a pulsed laser. This theoretically avoids any blind spots in the object detection. Its capability to measure distance and longitudinal velocity at the same time is an advantage over the ToF method. Since FMCW uses the interference of emitted/reflected laser to measure, it is less likely to be affected by ambient light, such as sunlight or other Lidars.

An incomplete list of developers of FMCW Lidars includes Aeva, Analog Photonics, Argo AI (Princeton Lightwave), Aurora (Blackmore), Baraja, Insight Lidar, OURS Technology, Psionic, SiLC, and Waymo.

7 Mapping Methods

In the last section, two different methods for single-point distance measurement used by Lidar were discussed (time-of-flight, ToF, and FMCW, amplitude modulated continuous wave, or AMCW, is not covered in this chapter). With single-point distance measurement achieved, the next step is to project this laser pointer around in the form of a 3D point cloud covering the 360° surrounding environment of the vehicle. Different methods have been applied to achieve this [4–15] including mechanical scanning, MEMS [16], optical phased array [15], et al. They each have their own characteristics and limitations and will be discussed one by one in detail in this section.

7.1 Mechanical Spinning Scanner

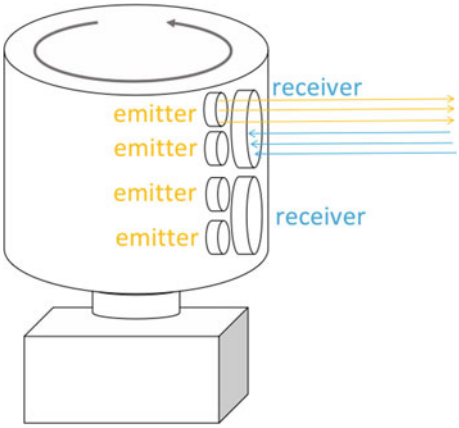
Mechanical spinning scanner Lidars are the most widely applied and most mature Lidar in the market today. It is also technically the most mature scanning mechanism. Multiple sets of laser sources and detectors are aligned in the vertical direction and mounted onto a driving motor. With the driving motor spinning 360° in the horizontal direction, every source-detector also scans a full circle and forms a line of multiple points. All these lines stack up and form a 3D point cloud, as shown in Fig. 7.

Mechanical spinning scanner Lidars usually use pulsed lasers as their light sources. In azimuth direction, their angular resolution is determined by laser pulse spacing, which is usually pretty high (0.08°) [11]. In the vertical direction, their angular resolution is determined by the number of source-detector pairs they have and is usually lower compared with that in azimuth direction. The frame rate is determined by the spinning speed of the driving motor and is usually pretty low, between 1 and 100 Hz [17].

The advantages of the mechanical spinning scanner Lidars include their technical maturity. With hundreds of autonomous testing vehicles on the road for years, they have accumulated the most experience and field mileage. They are still the choice of the majority of autonomous driving companies and their testing fleet for road use and data collection. Most of the data that have been collected and used for training and optimizing the algorithm are generated by mechanical spinning scanner Lidars. They usually have a wide 360° horizontal field of view (FOV) while other types of Lidar usually have smaller than 120° horizontal fields of view as comparison. They also have a pretty far detection range, and the uniformity of the point cloud is usually better than other types of Lidar.

Because of the high structural complexity of the multiple laser-detector pairs, a disadvantage is that mechanically spinning scanner Lidars with high line count are usually very bulky and expensive. Due to the large inertia of the rotating module, their power consumption is usually also very high (Velodyne HDL-64E consumes

Fig. 7 Simplified mechanical spinning scanner working principle



60 W) [11]. Their reliability and maintenance are also challenging. Under common driving conditions, mechanical vibrations and shocks can easily cause issues such as misalignment and fatigue.

To provide power and transmit data to/from the rotating source-detector modules, slip ring and brush are applied. Issues such as wear out and intermittent connection become common failure modes, limiting the lifetime of the mechanical spinning scanner Lidars to around two or three years. Due to the small installation base and road mileage, and the confidential R&D nature of the self-driving technology, photos showing different Lidar failure modes and FA reports on the root cause are extremely rare on the Internet. The authors will not discuss them in this chapter.

Suppliers for the mechanical spinning scanner Lidars include Velodyne, Hesai, and RoboSense.

Below, we listed some of the mechanical scanner Lidars currently on the market and their specs [18, 19].

Lidar	HDL-64E	Alpha prime	OS2 long-range lidar	Pandar 128	RS-Ruby
Company	Velodyne lidar	Velodyne lidar	Ouster	HESAI	RoboSense
Core technology	Mechanical	Mechanical	Mechanical	Mechanical	Mechanical
Max range	120	245	240	200	250
FOV (horizontal)	360°	360°	360°	360°	360°
FOV (vertical)	26.9°	40°	22.5°	16°	40°
Angular resolution (horizontal)	0.35° (20 Hz)	0.4° (20 Hz)	0.02°	0.1°	0.4° (20 Hz)
Angular resolution (vertical)	0.4°	0.11°	0.01°	0.125°	0.1°
Scan rate	~1.3 M pps (single) ~2.2 M pps (dual)	~2.4 M pps (single) ~4.8 M pps (dual)	~0.6 M pps (single) ~2.6 M pps (dual)	~3.4 M pps (single) ~6.9 M pps (dual)	~2.3 M pps (single) ~4.6 M pps (dual)

7.2 *Opto-Mechanical Scanning*

Opto-mechanical scanning refers to the usage of optical components, such as mirrors or prisms to steer the laser beam and achieve scanning. To solve the issues of early mechanical spinning scanner Lidars due to their design by integrating laser-detectors on a spinning driving motor, different methods have been used. These include decreasing the number of laser-detector pairs to reduce weight or keeping them stationary and using optical mirrors or lenses to complete the scan. Some examples are using double galvanometer scanning mirrors, gyroscopic mirrors, or Risley prisms for the scanning.

If the optical components only scan in one direction, as in a line scanner, the Lidar still needs multiple laser-detector pairs to cover the whole FOV. Some examples are a slated plain mirror, off-axis parabolic mirror, a polygon mirror, or a single galvanometer scanning mirror.

If the optical components scan simultaneously in 2 directions, only one or several laser-detector pairs are needed to cover the whole FOV.

Examples of the opto-mechanical scanning Lidar include the Scala Lidar installed on the first mass-production L3 level autonomous driving passenger vehicle—Audi A8. The Livox Lidar developed by DJI uses a Risley prism as its scanner.

7.3 MEMS Scanning

Microelectromechanical system (MEMS) mirror Lidar uses a millimeter or centimeter sized mirror to replace motor-driven, macro-scale mirrors or lens. Since there are no motors in the system and limited mechanical parts that will cause friction and wear out, MEMS Lidar greatly reduces the factors that affect the reliability and lifetime of the sensor.

From a cost perspective, every laser-detector pair in a mechanical spinning scanner Lidar costs nearly 200 dollars. A 16-line Lidar's cost is as high as 3200 dollars just for the laser-detectors. MEMS Lidar can greatly reduce the number of laser-detectors needed, thus helping to reduce the hardware cost. Meanwhile, MEMS mirrors can be designed to have wide scan angles and high scan frequencies, which generate dense point clouds and a high frame rate, improving the spatial and time resolution of the mechanical spinning scanner Lidar. The MEMS mirror is fabricated using a mature Si semiconductor foundry process and is largely immune to material fatigue, which is critical for a moving part that needs to tolerate 10^9 ~ 10^{11} duty cycles in its lifetime, as shown in Fig. 8.

Based on scan directions, the MEMS Lidar can be categorized into 1D MEMS Lidar and 2D MEMS Lidar.

The major disadvantages of the MEMS Lidar are that although MEMS mirrors are small and immune to fatigue, they are still fast-moving mechanical parts, risking its reliability and precision. This is especially notable in a shock event, which may damage the mirror. Temperature will also affect the material properties in the MEMS mirror and cause drift in scan angle and frequency, which need to be actively compensated.

An incomplete list of suppliers that are developing/manufacturing MEMS Lidars includes Aeva, AEye, Blickfeld, Cepton, Innoviz, Luminar, LeddarTech (Hybrid Flash), Livox, MicroVision, Pioneer, RoboSense, SOS Lab (SL-1), Toshiba, and XAOS.

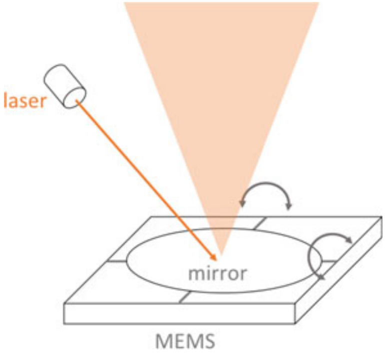


Fig. 8 Simplified MEMS scanning working principle

Among them, Luminar has Iria and Hydra published, and Ira is expected to be in mass production in 2022 and sold at 1000 dollars/unit for L3+ autonomous driving. The first generation of Innoviz Lidar, InnovizOne, has a maximum detection range of 250 m (assume 0.1 reflectivity) and is ordered by BMW. Compared with the previous generation, the latest InnovizTwo's cost has been reduced by 70%.

Below is a list of the MEMS Lidar on the market and their specs [18, 19].

Lidar	4Sight M	InnovizTwo	Dynamic view lidar	Scala2	Vision Mini	RS-LiDAR-M1
Company	AEye (continental)	Innoviz technologies	MicroVision	Valeo	Blickfeld	RoboSense
Core technology	MEMS/ToF	MEMS/ToF	MEMS/ToF	MEMS/ToF	MEMS/ToF	MEMS/ToF
Max range	300	300	250	200	150	200
FOV (horizontal)	60º	125º	100º	133º	120º	120º
FOV (vertical)	30º	40º	25º	10º	50º	25º
Angular resolution (horizontal)	0.1º	0.07º	0.03º	0.125º	0.2º	0.2º
Angular resolution (vertical)	0.1º	0.05º	0.03º	0.6º	0.6º	0.2º
Scan rate	4 M pps	–	10 M pps	0.25 M pps	–	1.5 M pps

7.4 Flash

The name Flash refers to the way the Lidar works similar to the flashlight of a camera when taking a photo. The full field of view is illuminated by a flood laser source,

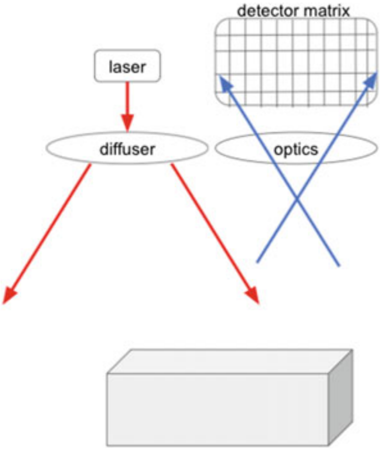


Fig. 9 Simplified flash lidar working principle

usually in a pulse form. An array of photodetectors at the image plane captures the time-of-flight (ToF) signal from every pixel and calculates distance information, which generates a point cloud (Fig. 9).

The advantages of Flash Lidar include the following: (1) A true solid state with no moving parts greatly improves its resistance to vibration and shock. (2) The flood laser source illuminates the entire FOV, which is a more reliable way for object detection compared to scanning certain points in space. (3) Flash Lidar uses an optical lens, which has already matured for years in cameras.

The disadvantage of the Flash Lidar includes the following: (1) Using flood illumination means every pixel in the image is only a small fraction of the returning laser power, which leads to low signal-to-noise ratio (SNR). Low SNR greatly limits the detection range of the Flash Lidar. (2) To compensate for the low SNR, high power laser source is one option; however, the heat management and power will become a problem for a compact Flash Lidar. (3) Low SNR means Flash Lidar is more prone to be affected and cheated by other Lidars from another vehicle. (4) Flash Lidar is relatively new and lacks maturity and experience.

A incomplete list of companies developing Flash Lidars includes Argo AI (Princeton Lightwave), Benewake, Fastree3D, LeddarTech (Pixell), Newsight Imaging, Phantom Intelligence (purchased by LeddarTech in 2020), RoboSense, Sense Photonics, SOS Lab (ML-1), TetraVue, Valeo, and Vergence Automation.

7.5 Optical Phased Array (OPA)

In an optical phased array (OPA) Lidar, the laser power is split into an array of transmitters. The phase of each transmitter can be controlled individually. By tuning the relative phase shift among transmitters, a laser beam can be formed and steered (Fig. 10).

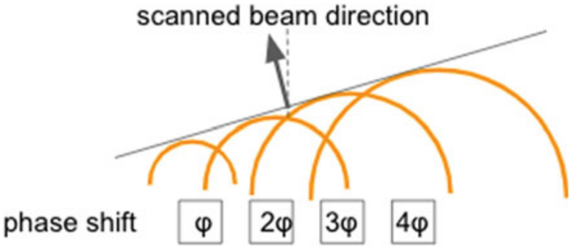


Fig. 10 Simplified OPA Lidar working principle

The advantages of OPA Lidar include the following: (1) There are no moving parts in the Lidar, which ensures good reliability and prevents any extra noise during the operation. (2) OPA Lidars use optical lens, which have already matured for years in cameras.

The disadvantages of this scanning method include the following: (1) In the transmitting and receiving stages, the dissipated light of the laser beam on the side lobes will reduce the efficiency and detection range of the Lidar. For example, the Quanergy S series of OPA Lidar only has an effective detection range of 11 m and currently is applied in low-speed short-range scenarios, such as parking assistance. (2) This concept is rather new and lacks maturity and experience.

An incomplete list of companies developing OPA Lidars includes RoboSense, Quanergy, Analog Photonics, NepTec, Voyant Photonics, and XAOS.

8 Discussion

The pros and cons for different scanning mechanisms of Lidar are summarized in the table below.

Scanning method	Pros	Cons
Mechanical	• Good precision in single-point distance measurement • High resistance to interference • Tolerance to work under high power laser	• Difficulty to meet automotive industry standards • Vertical scan angle fixed, difficulty in assembly and mass production
MEMS	• Highly integrated, compact in volume • Low wear out • Mature Si wafer processing easy for mass production	• Difficulty to control high precision high frequency scan • High wafer fabrication requirements • Limited FOV, cannot achieve 360 degree coverage, needs a combination of multiple units

(continued)

(continued)

Scanning method	Pros	Cons
Flash	• No scan needed • Fast imaging speed • Highly integrated, compact in volume • Mature optical CMOS technique, easy for mass production	• Laser power limitation • Relatively short detection range • More prone to interference and crosstalk • Low angular resolution
OPA	• Fast scanning speed • High scanning precision • High controllability	• Side lobes affect detection range and angular resolution • Complicated, hard to manufacture

Acknowledgements Both of the authors would like to thank their family and friends for their love and support in the pandemic time when this article is completed. Special thanks is to Angella Zhao for the grammar check.

The editors and authors would like to thank Dr. Bryan Tracy from Tesla for his review and advice.

References

1. Paul F. McMannamon. Lidar Technologies and Systems. SPIE, 2019
2. ASTM G173-03 (2003) Standard tables for reference solar spectral irradiances: direct normal and hemispherical on 37° tilted surface. https://www.astm.org/g0173-03.html
3. https://en.wikipedia.org/wiki/Laser_safety
4. Royo S, Ballesta-Garcia M. An Overview of Lidar Imaging Systems for Autonomous Vehicles. Applied Sciences. 2019; 9(19):4093. https://doi.org/10.3390/app9194093
5. Wang D, Watkins C, Xie H. MEMS Mirrors for Lidar: A Review. Micromachines. 2020; 11(5):456. https://doi.org/10.3390/mi11050456
6. C. -P. Hsu et al., "A Review and Perspective on Optical Phased Array for Automotive Lidar," in IEEE Journal of Selected Topics in Quantum Electronics, vol. 27, no. 1, pp. 1–16, Jan.-Feb. 2021, Art no. 8300416, doi: https://doi.org/10.1109/JSTQE.2020.3022948
7. Petermann, K. Advances in Optoelectronics; Springer: Berlin, Germany, 1988.
8. Amann, M.C.; Bosch, T.M.; Lescure, M.; Myllylae, R.A.; Rioux, M. Laser ranging: A critical review of unusual techniques for distance measurement. Opt. Eng. 2001, 40, 10–20.
9. Agishev, R.; Gross, B.; Moshary, F.; Gilerson, A.; Ahmed, S. Range-resolved pulsed and CWFM Lidars: potential capabilities comparison. Appl. Phys. B 2006, 85, 149–162.
10. Uttam, D.; Culshaw, B. Precision time domain reflectometry in optical fiber systems using a frequency modulated continuous wave ranging technique. J. Lightw. Technol. 1985, 3, 971–977.
11. Aulia, S.; Suksmono, A.B.; Munir, A. Stationary and moving targets detection on FMCW radar using GNU radio-based software defined radio. In Proceedings of the IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), Nusa Dua, Indonesia, 9–12 November 2015; pp. 468–473.
12. Wojtkiewicz, A.; Misiurewicz, J.; Nalecz, M.; Jedrzejewski, K.; Kulpa, K. Two-dimensional signal processing in FMCW radars. In Proceeding of the XXth National Conference on Circuit Theory and Electronic Networks; University of Mining and Metallurgy: Kolobrzeg, Poland, 1997; pp. 475–480.
13. You L., Javier I. Lidar for Autonomous Driving: The principles, challenges, and trends for automotive Lidar and perception systems. IEEE Signal Processing Magazine 2020 1053

14. Raj T, Hashim FH, Huddin AB, Ibrahim MF, Hussain A. A Survey on Lidar Scanning Mechanisms. Electronics. 2020; 9(5):741. https://doi.org/10.3390/electronics9050741
15. B. Behroozpour, P. A. M. Sandborn, M. C. Wu and B. E. Boser, Lidar System Architectures and Circuits, in IEEE Communications Magazine, vol. 55, no. 10, pp. 135–142, Oct. 2017, doi: https://doi.org/10.1109/MCOM.2017.1700030
16. C. Rablau, LIDAR – A new (self-driving) vehicle for introducing optics to broader engineering and non-engineering audiences, in Fifteenth Conference on Education and Training in Optics and Photonics: ETOP 2019, ETOP 2019 Papers (Optical Society of America, 2019), paper 11143_138.
17. https://velodyneLidar.com/products/hdl-64e/
18. https://autonomoustuff.com/lidar-chart
19. https://www.reddit.com/r/MVIS/comments/qa4m7h/lidar_comparison_chart_as_of_1017 2021/
20. Sun X, Zhang L, Zhang Q, Zhang W. Si Photonics for Practical Lidar Solutions. Applied Sciences. 2019; 9(20):4225. https://doi.org/10.3390/app9204225

Radar Technology

Mohammad Emadi

Abstract Over the last decade, new applications have emerged for radar technology in the automotive industry, such as adaptive cruise control, blind spot detection and automatic emergency braking. It is expected that the continued development of radar will unlock new capabilities for autonomous vehicles and safety systems. Indeed, advancements in integrated circuit design have enabled the development of very high-frequency radars with sophisticated signal processing and machine learning techniques. As a consequence, it is now possible with low power, small form factor and low-cost sensors to have rich point cloud data in dense environments. However, many challenges remain that can reduce precision, recall and accuracy of automotive radar, intra- and inter-platform interference or jamming, multi-radar fusion, multipath effects, false targets and detection ambiguity (range, velocity and angle). Many approaches have been developed to address these challenges; massive hybrid MIMO (Multiple-Input Multiple-Output), compressed sensing methods, sparse arrays, PMCW (Phase-Modulated Continuous Wave) Radar, Artificial Intelligence (AI) algorithms and others. Some of these approaches have already been applied to automotive radars, and some likely will be used in the coming years. This chapter will review why radar is unique, the challenges and the state-of-the-art solutions. We will use intuitive and simplified examples to provide the background needed to understand the utility of radar and these techniques.

1 Introduction

Increasing the driver's awareness of a vehicle's surroundings and preventing potentially hazardous operations are the main objective of Advanced Driver-Assistance System (ADAS). ADAS alerts drivers of potential risks in advance and, in some cases, plays a preventive role in mitigating collisions. Accordingly, the vehicle systems need a way to perceive the surrounding environment using sensors. Primarily, these sensors include camera, lidar and radar technologies. Many systems will use multiple types of

M. Emadi (✉)
Zada Labs, San Jose, CA, USA
e-mail: mo@zadarlabs.com

Y. Li and H. Shi (eds.), *Advanced Driver Assistance Systems and Autonomous Vehicles*,
https://doi.org/10.1007/978-981-19-5053-7_9

Table 1 Camera, lidar and radar capability comparison

	Camera	Lidar	Radar
Data type	2D Images	3D point cloud (x, y, z)	4D point cloud (x, y, z, doppler)
Color (RGB)	Yes	No	No
Computational burden	High	Medium	Low
Environmental condition	Daytime Clear	Daytime Nighttime Clear	Daytime Nighttime All-weather

technology together to improve performance. This is due to the fact that each of these types of sensors has its own benefits and drawbacks (Table 1). For instance, cameras are quite useful for identifying or classifying targets in a vehicle's surroundings. However, they are inefficient in providing valuable data in adverse environmental conditions like rain, snow, fog, intense light, dust and nighttime. Also, cameras typically require significantly more computing capability for data processing when compared to other sensors.

Lidar has a lower computational burden than a camera due to providing a point cloud rather than an image for data processing. The inherent benefit of a point cloud is the addition of the depth or distance of the detected points. Lidar also has better detection capability in some environments where a camera may be limited, such as nighttime. However, lidar remains hindered in poor weather conditions (i.e., rain, snow or fog). Furthermore, Lidar has difficulty detecting color (a traffic signal) or interpreting text (road sign).

Radar, similar to lidar, provides a point cloud of detections. In addition, radar directly measures the velocity information of the detections.[1]

Radar also has the capability of detecting glossy, dark and transparent objects and materials that impose a challenge to camera and lidar. Furthermore, advanced radars are capable of classifying objects at long distances. This specific capability is due to the radar system's ability to perform direct distance and velocity measurements and non-line-of-sight detection.

To demonstrate the difference in performance between lidar and radar, let us consider a scenario in which a human is approximately 200 m from the sensor and is between the sensor and a vehicle (Fig. 1). The goal is to detect both the human and the vehicle.

Using the sensor resolution and approximate range, we can calculate the expected number of detections for each sensor. Even though the lidar resolution is better, 0.2° in azimuth and elevation, the fact that lidar is a line-of-sight sensor limits the number of detections. In the radar case, although the resolution is worse, 0.5°, the sensor gets several detections from the human as well as the vehicle behind them. This is a crucial benefit to radar and its non-line-of-sight capability.

[1] FMCW Lidars also have this capability.

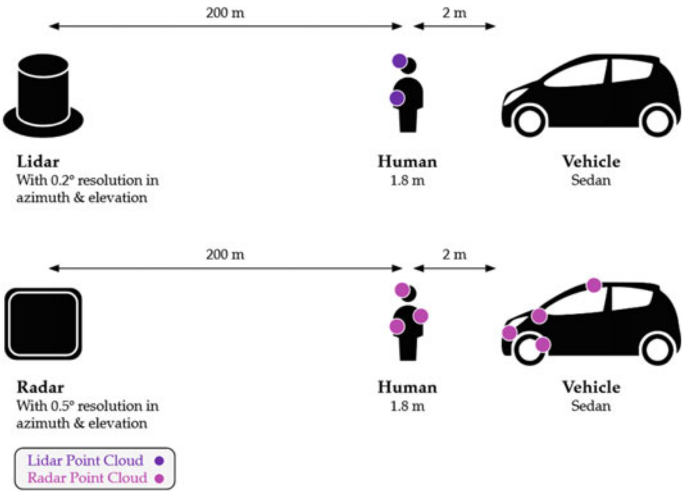


Fig. 1 Radar point cloud compared to lidar point cloud

The many benefits of radar have made it an essential ADAS sensor. Current ADAS applications fall into three general categories of radar. Short-range radar (SRR) is designed and tuned to detect targets up to 30 m and typically has a wide field of view (FOV). They are mainly used for parking assistance and pre-crash use cases. Mid-range radar (MRR) is typically for up to 80 m or 130 m and can also have a wide FOV and is usually used for blind spot detection, lane change alert and lane keep assist. Long-range radar (LRR) is for applications requiring more than 180 m in range with minimum FOV coverage (usually no more than ± 20°).

Advancements in integrated circuit design, antenna design and signal processing algorithms have enabled the development of radar that supports all the workloads of an LRR, MRR and SRR in a single design. In some cases, a single sensor can be used in place of multiple LRR, MRR and/or SRR (Fig. 2).

Altogether, advancements in camera, lidar and radar are resulting in more performant sensors. However, each modality is unique and has inherent limitations to the technology. It is important to note that in the application of sensors, more than performance is considered when selecting the best solution. Economic, aesthetic and reliability implications of each technology also play a role in the cost/benefit analysis. In their current form, lidar sensors are electro-mechanical and sensitive to vibration and impact. Some advancements in solid-state lidar may offer a solution in the future. Lidar also requires an unobstructed view of the environment, and it cannot be placed behind a fascia or easily hidden from view. Lastly, lidar continues to be relatively expensive compared to radar both for purchase and repair/replacement.

Most of the current radars are a fully solid-state technology which keeps costs low and offers better reliability. It is also able to be integrated behind a fascia or radome as needed, which offers more flexibility for designers. Radar has been implemented in mass for ADAS applications across the automotive industry, proving its viability

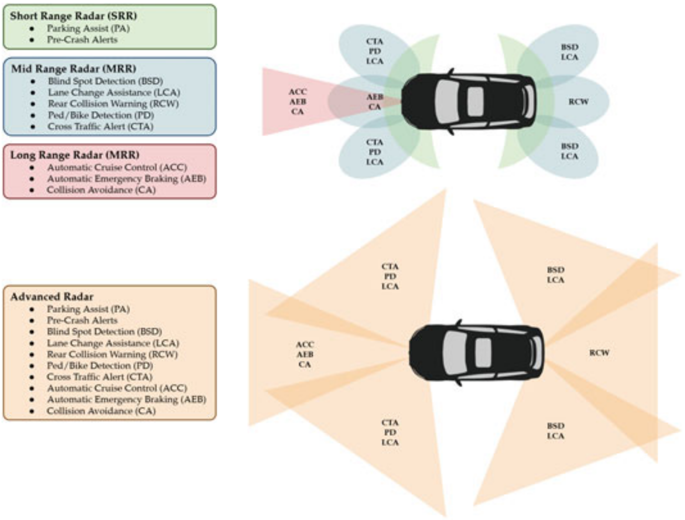

Short Range Radar (SRR)
Parking Assist (PA)
Pre-Crash Alerts
Mid Range Radar (MRR)
Blind Spot Detection (BSD)
Lane Change Assistance (LCA)
Rear Collision Warning (RCW)
Ped/Bike Detection (PD)
Cross Traffic Alert (CTA)
Long Range Radar (MRR)
Automatic Cruise Control (ACC)
Automatic Emergency Braking (AEB)
Collision Avoidance (CA)
ACC
AEB
CA
CTA
PD
LCA
AEB
CA
BSD
LCA
RCW
Advanced Radar
Parking Assist (PA)
Pre-Crash Alerts
Blind Spot Detection (BSD)
Lane Change Assistance (LCA)
Rear Collision Warning (RCW)
Ped/Bike Detection (PD)
Cross Traffic Alert (CTA)
Automatic Cruise Control (ACC)
Automatic Emergency Braking (AEB)
Collision Avoidance (CA)

Fig. 2 Current radars and future radars for ADAS applications

as a useful sensor modality. Understanding the technology and its advancements is essential for understanding ADAS.

2 Radar Physical Design

Radar operates by transmitting electromagnetic waves into space and receiving the reflection of those waves to estimate a target's distance from the radar, its velocities and the angle of arrival (AoA) both in azimuth and in some cases elevation. Because a radar uses these waves to estimate the position of the target in 3-dimensional space as well as the velocity, it is often referred to as a 4-dimensional sensor. In general, all radars share a similar architecture; however, there are various approaches to implementation.

2.1 Radar Architecture

Generally, radar architecture is composed of five sections (Fig. 3):

1. RF transceiver
2. RF antenna
3. Signal processing
4. Data processing
5. System management.

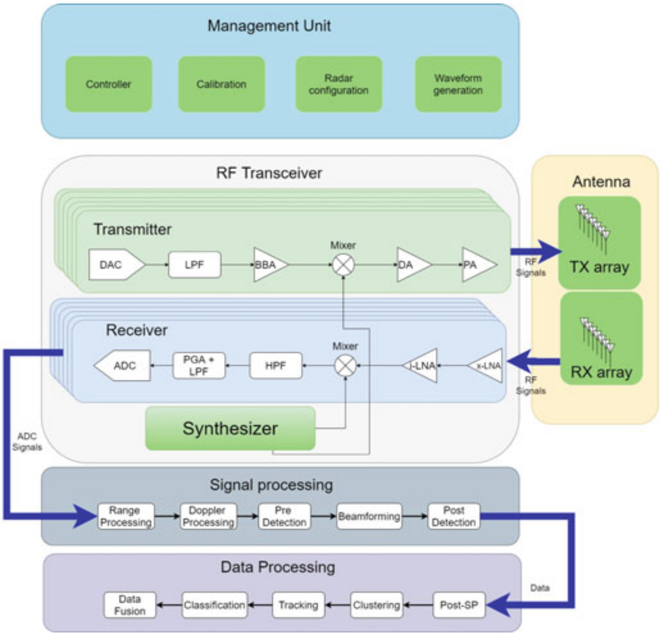


Fig. 3 General radar architecture

2.1.1 RF Transceiver

The RF (Radio Frequency) transceiver both generates the signal to the antenna and receives the reflections from the antenna. The received signals are converted from analog to digital samples in the receiver for downstream signal processing. The RF transceiver is composed of three sub-sections: transmitter, receiver and synthesizer.

The **transmitter** typically has a **digital-to-analog converter (DAC)** for converting the digital waveform to analog. A **low pass filter (LPF)** is then implemented to filter out any DAC spurs from the quantization distortion errors. The **baseband amplifier (BBA)** will amplify the signal before it is up-converted from the baseband to the RF by the **mixer**. The **driver amplifier (DA)** and **power amplifier (PA)** increase the power level and match the impedance to be transmitted by the antenna. Some radars may use an external PA to increase the power. In some designs, a **phase shifter** is implemented to change the phase of the transmitted signals on each antenna elements.

The **receiver** likely has an **external low-noise amplifier (x-LNA)** or an **internal low-noise amplifier (i-LNA)** or both to amplify the received signal and increase the signal-to-noise ratio (SNR). A **mixer** is then used to down-convert the RF to the baseband before the **high pass filter** (HPF) eliminates unwanted close distance high-power signals for FMCW radars. The **programmable gain amplifier** (PGA) is implemented to amplify and control the signal to an appropriate level for the system.

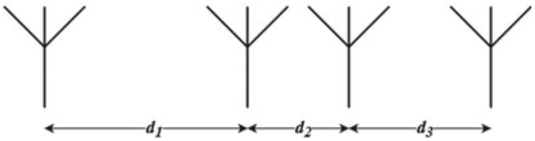


Fig. 4 Simple 1D linear Rx array

Finally, an **analog-to-digital converter** (ADC) converts the received signals back to digital for signal processing.

The **synthesizer** provides the RF for the mixers when up-converting or down-converting the signal. Currently, mass market ADAS radars are operating at 24 GHz, usually for SRR applications, and 76–77 GHz, usually for LRR applications. However, newly developed radars have expanded to 76–81 GHz which can provide better resolution and detection performance compared to lower frequencies.

Research on higher frequency bands, above 81 GHz, is ongoing but faces challenges specifically related to semiconductor technology and performance at the sub-THz frequency ranges [1]. In addition to technological challenges, the Federal Communications Commission (FCC) regulates frequency band usage in the United States and has requirements for automotive radar applications such as radiated power limits and emissions. In 2017, the FCC released the 76–81 GHz band for automotive radar applications. It is expected that other regions in the world will do the same [2].

2.1.2 Antenna

The antenna is responsible for converting the RF signal from the transceiver to electromagnetic (EM) waves for transmission and converting the received EM reflections to RF signals prior to being processed by the transceiver.

Some of the most common antenna element types utilized in ADAS applications include **patch antenna** (vertical polarization), **balanced feed vertical polarized antenna**, **slotted arrays** and **dual polarized patch antennas**. Each antenna type will have different implications on radar performance and system design. Selection or design of the antenna element is usually driven by the application's requirements.

Single-Input Multiple-Output (SIMO)

Using an array of antenna elements has advantages in improving detection and resolution performance. Different array designs can achieve dramatically different results. Moreover, by changing the phase of the signal sent to each antenna, it is possible to steer the transmitted or received power to a specific direction.

Single-Input Multiple-Output (SIMO) refers to an array with a single transmit (TX) and multiple receive (RX) antenna elements. The angular resolution of the SIMO radar depends on the number of RX antenna elements.

Let us assume we have a simple 1D linear array of receiver (Rx) antennas (Fig. 4).

The total length of the array is equal to:

$$D = d_1 + d_2 + d_3 \tag{1}$$

When we consider diffraction through a circular aperture, the angular resolution or beamwidth of the antenna array is equal to:

$$\Delta\theta = 1.2\frac{\lambda}{D} \tag{2}$$

where

λ the wavelength of light
$\Delta\theta$ angular resolution in radians

Angular resolution refers to the minimum angular distance at which two targets can be separated, or resolved, from each other. In other words, two targets less than $\Delta\theta$ apart, will be indistinguishable from a single target. By increasing *D,* it is possible to improve (reduce) angular resolution. However, if the distance, d_n, between antenna elements increases beyond $\lambda/2$ we will see the presence of grating lobes (very strong side lobes). For example, if the radar has a main lobe at beam angle of 30°, another lobe (the grating lobe) may appear at 65°, depending on the relationship between d_n and λ. This relationship resembles the Nyquist Theorem in the space domain.

The Nyquist Theorem states that to detect a signal whose frequency is f_c, the sample frequency, F_s, must be greater than $2 \times f_c$. Otherwise, there will be ambiguity in the frequency domain. Similarly, an array of antenna elements is sampling space, and to avoid the presence of grating lobes, the distance between antenna elements needs to $d_n \leq \lambda/2$.

In summary, to achieve approximately 3° angular resolution, the array would require 40 antenna elements with $d = \lambda/2$.

In practice, the use of Fast Fourier Transform (FFT) for frequency estimation (in the frequency domain) or angle of arrival (AoA) in the space domain requires uniformity of *d* across all antenna elements. Commonly called uniform arrays, it is related to uniform sampling theory. Additionally, it may not be practical to implement such a large array of antenna elements as it would increase **cost**, **complexity** and **power consumption**.

Multiple-Input Multiple-Output (MIMO)

One approach to increase the perceived size of the array without the addition of physical antenna elements is to employ **Multiple-Input Multiple-Output (MIMO)** techniques which will increase the array size by adding virtual antenna elements. When additional transmitting (TX) antenna elements are placed at some distance, from the perspective of the target it is as if receiving (RX) antenna elements have been

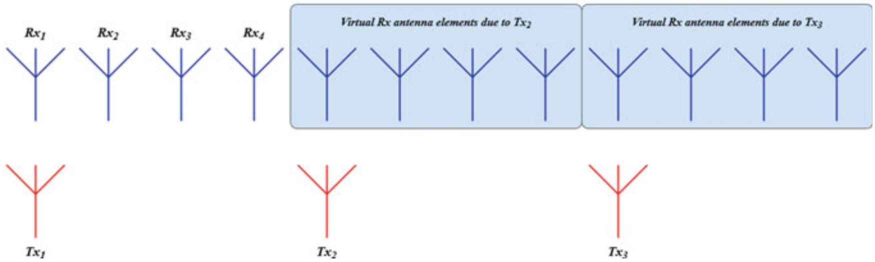

Rx_1
Rx_2
Rx_3
Rx_4
Virtual Rx antenna elements due to Tx_2
Virtual Rx antenna elements due to Tx_3
Tx_1
Tx_2
Tx_3

Fig. 5 TX-MIMO, uniform linear array (ULA), 3 Tx and 4 Rx to have 12 elements

moved, and thus from a system perspective, it is as if we added receiving antenna elements (Fig. 5). This approach is known as **TX-MIMO**.

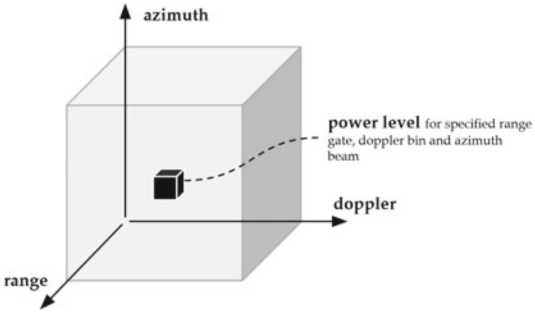
Other techniques such as non-uniform array (based on non-uniform sampling theory) and RX-MIMO (sparse array) [3] have been used in some applications as well.

2.1.3 Signal Processing

Signal processing techniques are applied to the RF receiver data to estimate the **range** (distance), **doppler** (range-rate) and the **azimuth and elevation angles** of the target. This is often viewed as a four-dimensional search in the range-doppler, range-azimuth and elevation domains. For simplicity, Fig. 6 shows a 3D cubic without the elevation angle domain.

Typically, signal processing begins with **range processing** to analyze all available data in the range domain. This is done by processing the collected data in each range gate, or sometimes referred to as a range slice. A **range gate** is an interval of range (or time delay from transmission) within which the reflected signals are measured. Different types of radars will perform range processing differently. For example, Frequency-Modulated Continuous Wave (FMCW) radars will likely use

Fig. 6 Cubic data at the output of radar signal processing


azimuth
power level for specified range gate, doppler bin and azimuth beam
doppler
range

an FFT approach. In Pulse-Modulated Continuous Wave (PMCW) radars, a pulse compression technique may be used.

After separating the range slices, several reflected chirps (pulses) can be used to perform **doppler processing** and estimate the doppler of each detection or target. Often, the range and doppler processing results are used to generate a 2D plan referred to as a range-doppler heat map. In some cases, the next processing step is **pre-detection** which can reduce the downstream computational burden by reducing the number of points processed by using techniques to remove unneeded data for beamforming. **Beamforming** is the process of shaping the received (or transmitted) signals to estimate the **angle of arrival** (AoA) of the received signals or the angular (azimuth) location of the range-doppler data (2D beamforming). In the case of 3D beamforming, the elevation angle will also be estimated to provide a 4D **point cloud** for data processing.

In recent years, many of the outward advancements in radar have come from the development of new signal processing algorithms and the implementation of improved digital electronics.

2.1.4 Data Processing

Depending on the richness of the provided point cloud, **data processing** methods can range from simple filtering to advanced machine learning and artificial intelligence techniques. The primary objective of data processing is to develop the data in a way that is meaningful for the application. In many cases, this includes clustering, tracking and classification of targets. If there are other sensors, radar or otherwise, involved in the application, data fusion may be needed as well. All these techniques come with a computational burden. The design and implementation of these processing algorithms must consider both performance and efficiency.

2.1.5 System Management

Management of the system is critical in maintaining expected operation as well as modifying the radar performance or tuning to suit the application. **Knowledge-based radars** actively modify the radar tuning based on key detection feedback and processing outcomes. Sometimes known as adaptive radars, these sensors will optimize operation based on targets identified as threats or environmental conditions. By actively adapting RF transceiver performance, signal processing and/or data processing methods based on the knowledge of the environment or detection scenario, the radar can reduce the likelihood of inappropriate action taken by the platform. For example, appropriately integrating feedback from the tracking engine could improve the probability of detection while maintaining or reducing the probability of false alarm [4].

2.2 Radar Categories

Most available ADAS radars on the market today are **monostatic,** meaning the transmitter and receiver are collocated within the physical design. These radars are commonly a **TX-MIMO** design and support **Rx digital beamforming** and in some cases **Tx digital beamforming**.

Bistatic or **multistatic**, **RX-MIMO (sparse array)** and **networked** radars also have potential for ADAS applications but are currently in development or have limited adoption.

2.2.1 Monostatic Radars

The monostatic radar's collocation of the transmitter and receiver in the physical design reduces the system complexity and improves the synchronization of antenna elements due to their proximity. However, the closeness of the TX and RX antenna elements can result in **leakage** or the unwanted interaction between transmitted and received signals. It is possible to use a single antenna for both TX and RX if a non-concurrent transmission and reception technique is used. Some **pulse radars** have been known to have a non-concurrent approach. In contrast, most **continuous wave (CW) radar** applications require different TX and RX elements and support concurrent transmission and reception. Leakage can still occur even with discrete TX and RX antenna elements if the system design does not consider the total signal path to minimize the behavior. In addition, the design must also minimize the effects of the antenna housing, or **radome**, which can reflect the transmitted signal. Mitigation strategies include the orientation and placement of the TX and RX antennas, grounding techniques and many other methods.

2.2.2 Bistatic or Multistatic Radars

In **bistatic** radars, the transmitter and receiver are separated by a distance typically comparable to the expected distance to the desired target. The radar cross section is considerably larger than a monostatic design. With a larger cross section, the reflectivity of targets will also be increased which improves detectability. More transmitters and/or receivers can be used to increase the size of the array and create a **multistatic radar**. The primary challenges to this approach are the increased system complexity and difficulty to synchronize the signal. These challenges have not yet been resolved for automotive applications.

2.2.3 RX-MIMO (Sparse Array)

A sparse array [3] radar increases the number of antenna elements virtually, similar to other MIMO designs. However, in this case the distance between RX antenna elements, d, is greater than $\lambda/2$. Implementation of a sparse array has significant technical challenges with respect to signal correlation and processing.

2.2.4 Radar Networks

An automotive radar network combines the data from different monostatic radars. This could be a network of radars on the same vehicle (intra-vehicle), or between vehicles (inter-vehicle), or between a vehicle and infrastructure. All these applications have data synchronization and device connectivity challenges that must be solved before being useful in an ADAS application.

2.2.5 Digital Beamforming

Beamforming is the signal processing technique in which the RX and/or TX signals are shaped to focus on a desired azimuth and/or elevation. This can be done in the **analog** or **digital** domain. Almost, all the available automotive radars in the market today are using one form of digital beamforming to acheive an enhanced performance.

TX Beamforming

Using TX beamforming, the radar directs the TX antenna beam in a specific direction in a way that constructively combines the power of several signals to increase the detectability in the desired direction. The total increment in power level will be equal to:

$$20 \times \log(N_{\mathrm{Tx}}) \tag{3}$$

where

N_{TX} the number of TX antenna elements

The beam can also be rotated by using phase shifters, in the analog domain, or generate the needed phase prior to analog conversion in the digital domain. TX beamforming usually lowers the field of view (FOV) which will increase the radar's scan time. Some ADAS radars rely on TX beamforming in specific applications where the FOV requirements are small. For example, most LRR required only ± 20° FOV in azimuth. One can use TX beamforming only on the elevation side.

RX Beamforming

ADAS radars most commonly use **RX digital beamforming** techniques. After the RX signals from different channels are received (physical and/or virtual in the case of MIMO), it is possible to simultaneously have beams in different directions over the required FOV. This is accomplished through various advanced beamforming methods applied during signal processing. Some examples include the following:

- Spectral-Based Methods (MUSIC [5])
- Min-Norm [5]
- Capon [6]
- Iterative Techniques [7] (IAA [8]
- Iterative Capon [9]
- OMP [10])
- Parametric-Based Methods (ESPRIT [11])
- Maximum Likelihood [11]
- Beam-Spaced Methods [12].

3 Waveform Design

A radar illuminates a signal into space and, based on the received reflected signals, estimates the target's range, range-rate (doppler), azimuth angle and elevation. To accomplish this, the transmitted wave needs to have specific characteristics. For the scope of this discussion, we will focus on **pulse**, **pulse coded** and **FMCW radars** as well as different waveforms for MIMO applications.

3.1 Pulse Radar

Pulse radars transmit a **pulse train**, or a series of pulses, into the environment (Fig. 7).

Several pulses with active time equal to τ are sent into space. These pulses will be repeated at the **pulse repetition interval (PRI)**.

There are two main reasons for sending several pulses into the environment.

1. By integrating several pulses on the receiver, the signal-to-noise ratio can be increased, and radar will have better probability of detection and lower probability of false alarm.
2. The doppler estimation will be improved with more samples.

Pulse Repetition Frequency (PRF), 1/PRI, is sometimes used in place of PRI. The reflected signal will have a delay due to the distance of the target to the radar. This delay is calculated as Δt.

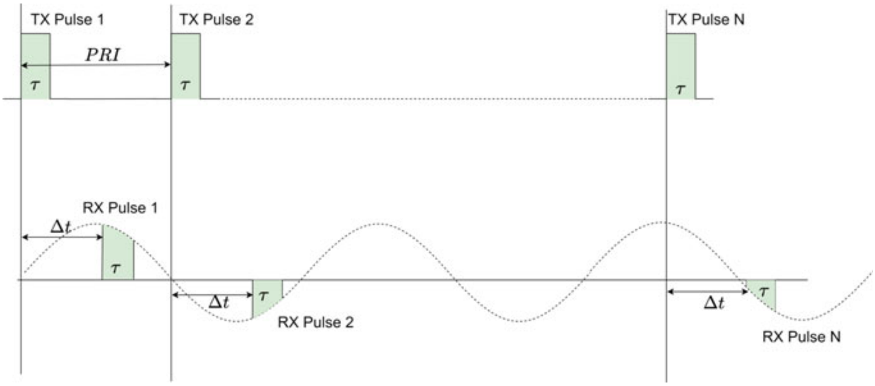


Fig. 7 Transmitted and received pulses of a pulse radar

$$\Delta t = \frac{2 \times R}{c} \tag{4}$$

where

c the speed of light
R target range

Maximum detection range is equal to:

$$R_{\max} = \frac{c}{2 \times \mathrm{PRF}} \tag{5}$$

The pulse radar estimates **doppler** by evaluating the phase change of the received signals over the estimated range. Furthermore, the reflected signals will be modulated by a doppler frequency. The relationship between doppler frequency ($\mathrm{f_d}$) and radial velocity ($\mathrm{v_d}$) of the target is used to estimate the doppler.

$$f_d = \frac{2 \times v_d}{\lambda} \tag{6}$$

where

λ wavelength of light

By sampling the received signal at the PRF, and considering the Nyquist Theorem, the maximum measurable doppler ($f_{d_{\max}}$) and the maximum radial velocity ($\mathrm{v}_{\mathrm{d}_{\max}}$) are equal to:

$$f_{d_{\max}} = \pm \mathrm{PRF}/2 \tag{7}$$

$$v_{d_{\max}} = \pm\frac{\lambda \times \mathrm{PRF}}{4} \tag{8}$$

Range resolution, the capability of the radar to separate two targets at different distances (ranges), is calculated as ΔR.

$$\Delta R = \frac{c \times \tau}{2} \tag{9}$$

where

τ is active pulse time

It is more straightforward to use radar bandwidth, $\mathrm{BW} = 1/\tau$, which results in:

$$\Delta R = \frac{c}{2 \times \mathrm{BW}} \tag{10}$$

Doppler resolution, the capability of the radar to separate two targets at the same range with different range-rates (dopplers), is calculated as Δf_d.

$$\Delta f_d = \frac{\mathrm{PRF}}{N} \tag{11}$$

where

N number of pulses sent by the radar

Or when substituting in the relationship between doppler and radial velocity (Eq. 4):

$$\Delta v_d = \frac{\lambda \times \mathrm{PRF}}{2 \times N} \tag{12}$$

An Example in Pulse Radar

Let us assume we want to design a pulse radar with 1.5-m range resolution and maximum range of 150 m.

Based on Eqs. (9) and (10), the pulse width and BW of the radar are equal to 5 ns and 200 MHz, respectively. Also, based on maximum range (5) the radar PRI and PRF will be 1 us and 1 MHz, respectively.

Key Considerations:

1. We assumed there was no coupling present between the doppler and range. It means the variation of pulse amplitude over the active time due to the doppler is very low. This is true when the pulse width is very small compared to doppler velocity.

   In this example, if the relative radial velocity (range-rate) between the radar platform (Ego) and the target is 150 mph (67 m/s), then the doppler shift for

77 GHz radar ($\lambda = 3.9$ mm) according to (6) is equal to 34.4 kHz, which results in no change over $\tau = 5$ ns.

2. We assumed the target will stay in the same range gate over the integration time. This means the target's movement will not be higher than half of the range gate size (range resolution) during the integration time.
3. Sampling the pulse is limited by the receiver BW (both for the baseband analog amplifiers and ADC). This results in an imperfect pulse shape (rectangular) and may only allow for one or two samples of the pulse. Consequently, this causes a reduction in the received power level of the signal. Often known as a **straddling effect**, it can occur in the range, doppler and angle estimation domains.
4. In practice, advanced algorithms are regularly implemented to improve range and frequency estimation to achieve higher resolution and accuracy.

A critical issue for pulse radar design is the ratio of active time, τ compared to PRI. From the example:

$$\frac{\tau}{\mathrm{PRI}} = \frac{5\ \mathrm{ns}}{1\ \mu\mathrm{s}} = 0.5\%$$

In other words, the radar is effectively active for only 0.5% of the operation time. The implication is a significant reduction in received signal power and detectability of the radar. Because of this characteristic, pulse radars are not widely found, if at all, in ADAS applications. Pulse coded and frequency-modulated radars attempt to solve this issue.

3.2 Pulse Coded Radar

By adding coding to the transmitted waveforms, it is possible to substantially increase the active time of the radar while maintaining resolution performance. This technique is generally known as **fast coded pulse modulation** (Fig. 8).

As depicted (Fig. 8), the entirety of the active time is increased by sending several short pulses with different phases during each PRI. All the formulas discussed in relation to pulse radars still hold for pulse coded approaches. However, to estimate a

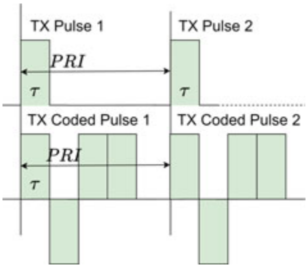


Fig. 8 Stylized example of fast coded pulse modulation technique

more accurate delay of the received reflected signals, a pulse compression technique is needed. Furthermore, it is necessary to apply a coding technique that limits correlation with the signal delay. Many coding techniques are available for use [13]: Baker, Walsh-Hadamard, Zadoff-Chu and so on.

In addition to solving the issue of low received signal power, pulse coding provides an opportunity to reduce interference. By using heterogeneous coding between transmitters, the intra-radar interference can be reduced. Similarly, by using diverse coding schemes between radars, the inter-radar interference can be reduced.

However, pulse coded radars have some implementation issues that need to be considered carefully:

1. An **ADC with high sampling frequency**. Pulse compression techniques are applied in the digital domain and require high sampling frequency capabilities for analog-to-digital conversion.

   For example, simultaneous localization and mapping (SLAM) methods typically require 10 cm or better range resolution. This results in a 1.5 GHz BW and an ADC sampling frequency of 3 GS/s (with an effective number of bits >9 bits) or better. Designing ADC with this specification is really cumbersome and will be very power hungry unless we reduce the effective number of bits. It is important to note that having a lower effective number of bits will result in a low dynamic range performance of the radar. Low dynamic range is sometimes referred to as a **near-far problem**, meaning it is difficult for the radar to detect targets with low reflectivity at long range when targets with high reflectivity are nearby.
2. **Linear baseband amplification**. Linearity of the baseband amplifier over the bandwidth is needed for both the transmitter and receiver. Without it, pulse coded radars tend to generate ghost target detections and have higher side lobes around targets in the range domain.
3. **Regulatory certification** of pulse coded radars can be challenging. It is difficult to eliminate or mask the side lobes produced by the series of pulses to pass the various band leakage requirements for certification.

Examples of challenging scenarios for a pulse coded radar:

1. Detecting a human (low reflectivity object) at long range (i.e., 150 m) when a high reflectivity target (a bus) is nearby (near-far problem).
2. A low reflectivity and high reflectivity targets are in close proximity and not moving (nonlinearity effect).
3. The high range resolution (i.e., < 10 cm) is regularly required for SLAM and/or classification algorithms while maintaining the high resolution requirements in doppler, azimuth and elevation domains.
4. It is not possible to compensate for saturation of the analog baseband (on the receiver) due to jamming. This is why in some cases even the performance of pulse coded radars is worse than FMCW radars with respect to interferences.

3.3 FMCW Radar

Frequency-Modulated Continuous Wave (FMCW) radars transmit a signal called a **chirp** in which a continuous wave's frequency increases linearly over time (frequency modulation) (Fig. 9). The transmitted and the received signals are combined at the mixer to form a frequency proportional to the delay called the **intermediate frequency (IF)**.

Using the intermediate frequency (f_{IF}), the range can be calculated:

$$R = \frac{c \times f_{\mathrm{If}}}{2 \times s} \tag{13}$$

where s, the **slope** of the linear frequency-modulated (LFM) signal and is equal to:

$$s = \frac{\mathrm{BW}}{\mathrm{PRI}} \tag{14}$$

where

BW bandwidth
PRI pulse (chirp) repetition interval

Based on the ADC sampling frequency (F_s), the **maximum detectable range** ($\mathrm{R_{max}}$) will be equal to:

$$R_{\max} = \frac{c \times Fs}{2 \times s} \tag{15}$$

The formulas for range resolution, doppler resolution and doppler ambiguity will be the same as pulse radar.

An Example in FMCW Radar

Let us assume we want to design a pulse radar with 1.5-m range resolution and maximum range of 150 m. With an ADC sampling frequency ($\mathrm{F_s}$) of 50 MHz, the required slope (15) will be 50 MHz/us and the required BW (10) will be 200 MHz

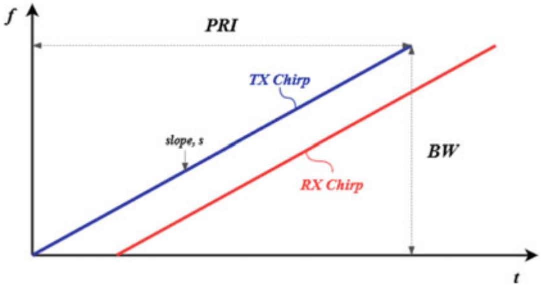


Fig. 9 FMCW TX chirp and RX reflection

with a PRI (14) equal to 4 usec. It is important to note we have achieved the same range performance requirements with a lower ADC sampling frequency compared to a pulse coded design. The ability to use lower frequency sampling is a key reason why most available ADAS radars on the market today are FMCW.

In addition, coded FMCW techniques are used to achieve better robustness against intra- and inter-radar interference and reduce the probability of being jammed during operation.

3.3.1 Fast Coded FMCW Radars

Fast coding techniques multiply the chirp (during the PRI) by a code, similar to pulse coded radars. The code needs to be orthogonal and not sensitive to the doppler effect. Adding the code inherently adds complexity and will reduce the dynamic range performance. At the moment very few, if any, available radar chips support this feature.

3.3.2 Slow Coded FMCW Radars

Slow coding techniques utilize various codes over different chirps (between radars or between TXs on the same radar). The codes used need to keep their orthogonality after range processing. **Constant amplitude zero autocorrelation (CAZAC)** waveforms are good for these types of use cases. The FFT of a CAZAC sequence is also CAZAC and not sensitive to the doppler effect [14]. There is the potential for leakage between TXs or radars when using coding, which results in lower dynamic range in the doppler domain. Many available transceivers in the market today support slow coding techniques.

4 Link Budget Analysis for FMCW Radar

Understanding the link budget helps in achieving a better estimation of the maximum range information which can be extracted from a detection. This is done by analyzing the power level of a received reflected signal using the **radar equation** and considering phenomenon such as **noise** and **nonlinearity**.

4.1 Radar Equation

The power of a received signal from a target can be modeled as:

$$P_r = \frac{P_t G_t G_r \lambda^2 \sigma}{(4\pi)^3 R^4} LF \tag{16}$$

where

P_r	received power
P_t	transmitted power
G_t	transmit antenna gain
G_r	received antenna gain
σ	reflectivity of the target
R	distance of the target
L	losses in the radar
F	multipath effect.

In addition to the received power, there will be the presence of noise. In general, **noise** comes from many different sources of the system and the environment, for example, receiver **thermal noise** or the **noise figure (NF)**, **impairments**, **clutter** and **interference**. The final **signal-to-noise impairments-clutter-interference (SNICIR)** determines the level of signal to background noise prior to signal processing. Several techniques can be applied to increase the ratio and extract the signal. The increase in the SNICIR is considered a **processing gain *(PG)***. The final SNR (logarithmic scale) will be equal to:

$$\text{SNR}_{\text{final}} = P_r - P_{N_{\text{th}}} - P_{\text{Impairments}} - P_{\text{Clutter}} - P_{\text{Interference}} + \text{PG} \tag{17}$$

4.1.1 Antenna and Array Gains

An **isotropic antenna** has no directivity, and it will radiate energy equally in all directions and is said to have “no gain.” Typically, an automotive radar antenna has some directivity and unequal radiation in some direction and will transmit energy accordingly having between 5 and 15 dBi gain based on the field of view requirements. Beamforming techniques on the transmitter or receiver will increase the gain in a particular direction. A practical formula for the gain could be $10 \times \log(N)$ where N is the number of antenna elements. If the system is using digital beamforming, the receiver antenna gain can be calculated as part of the processing gain.

4.1.2 Noise Figure

The RF transceiver is composed of the transmitter and receiver (Fig. 3). Transmitted power tends to be between 10 and 15 dBm **(dB per milliwatt)** depending on the integrated circuits (ICs) used. Some designs use external power amplifiers (PA) and low-noise amplifiers (LNAs) outside the transceiver to increase the transmitted power and reduce the thermal noise power (Fig. 10) and improve final SNR. However, there

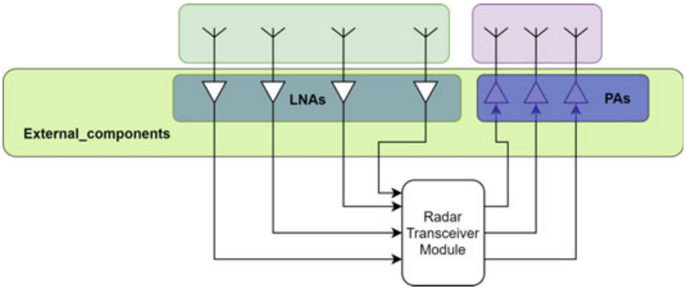


Fig. 10 Example of external PA and LNA design

are integrity and quality challenges specific to ADAS that have made this approach unsuitable for automotive radars.

Thermal noise on the receiver is one of the significant contributors to reducing the final signal-to-noise ratio (17). It is an important limiting factor for detection, specifically at long range. Usually, thermal noise is represented by the noise figure.

A practical formula for calculating thermal noise ($P_{N_{\text{th}}}$) in logarithmic scale (dBm) is:

$$P_{N_{\text{th}}} = -174 + \text{NF} - 10 \times \log(\text{BW}) \tag{18}$$

This formula shows that if we increase the BW of the radar (for better range resolution), the thermal noise power will increase and will result in lower detectability.

The typical NF of a 77 GHz automotive radar is between 10 and 15 dB at maximum receiver gain (G_r). If the gain of the receiver is reduced, the NF will increase accordingly.

4.1.3 Receiver Impairments

In operation, receiver impairments are a significant factor in limiting the SNR. Two principal impairments are **nonlinearity** and **phase noise**.

Receiver Nonlinearity

When a high-power signal is received at the input of the receiver, its gain will not behave in a linear way. The **nonlinearity** will produce harmonics that mainly appear in the range domain as false targets. Commonly, an **input third-order intercept point (IIP3)** figure of merit is used to indicate the nonlinear effect (Fig. 11).

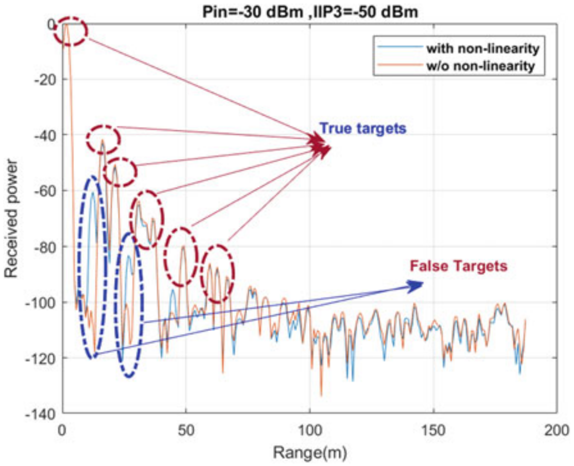


Fig. 11 Nonlinearity effect on the detection

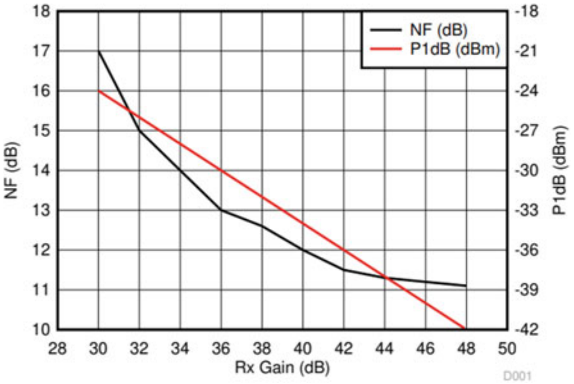


Fig. 12 TI AWR2243 linearity and thermal noise [15]

From the integrated circuit perspective, there is an inverse relationship between thermal noise and nonlinearity. For example, if the NF is good (low thermal noise), the linearity will be poor (high IIP3) (Fig. 12).

In Fig. 12, P1dB represents the **one dB compression point** or the output power level at which the gain decreases by 1 dB from its constant value and, in other words, the point at which the device becomes nonlinear and will produce harmonics, distortions and intermodulation. The relationship with IIP3 in logarithmic is:

$$\mathrm{IIP3} = \mathrm{P1dB} \ + 10 \tag{19}$$

Because of this relationship, it is necessary for the system to change the receiver gain based on the environment, power of received signals and/or the ADC's dynamic range. There are some **advanced automatic gain control (AGC)** and fixed gain techniques that accomplish this. For example, different gains could be required for

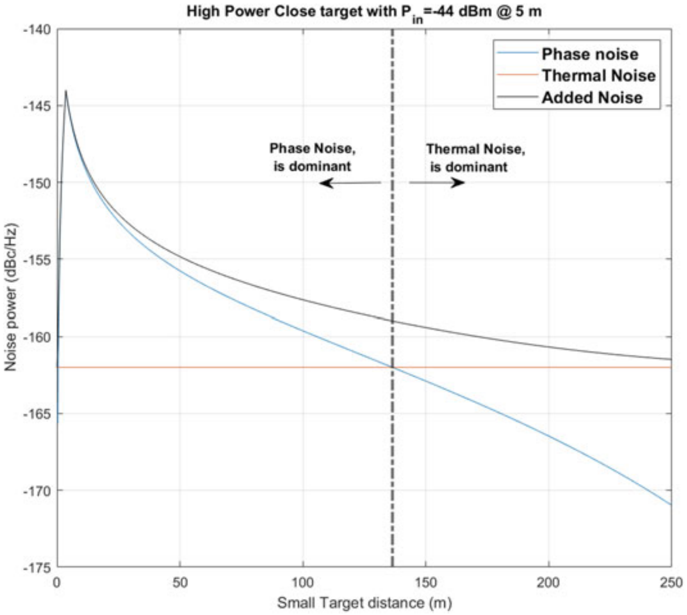


Fig. 13 Phase noise contribution compared to thermal noise

better performance indoors, or in an urban environment or on an interstate. To improve the effectiveness of the tradeoff between NF and nonlinearity, a high pass filter can be used in the receiver to manage high-power reflections at close range.

Phase Noise

Experimental data have shown phase noise to be a significant factor in limiting the maximum detection range and other parameters of a radar. It has become a critical factor in the design of automotive radars as performance requirements have increased with the emergence of new ADAS applications. **Phase noise** is a measure of the frequency stability within the oscillator and is different from background noise of the electrical system.

In FMCW radar, the phase noise will persist even after mixing the transmitted and reflected signals due to the difference in reflected signal power over the detection range (Fig. 13).

Figure 13 shows a high-power signal with input power (P_{in}) of approximately − 44 dBm at 5 m distance to the radar. And we swept the small target (Bike) that had swept through the range. The total noise figure is dominated by phase noise until approximately 140 m range at which point thermal noise becomes dominant. This behavior of phase noise over the range, and dependent on close target distance and reflectivity, makes it challenging to cancel out in practice.

Other Impairments

Other notable impairments include **transmitter nonlinearity**, **synthesize frequency spurs**, **DC offset**, **ADC quantization noise** and **inphase quadrature imbalance**. All of which will increase noise power [16].

4.1.4 Losses

Received signal power is subject to various sources of losses which appear as L in the radar equation. Table 2 shows a general summary of approximate loss values.

From Table 2, the antenna is the source of about a half of the losses. It is challenging to have good directivity and gain [G_t and G_r, (16)] while also having wide beamwidth (large FOV). This most commonly results in **attenuation**, reduction in signal strength, near the edges of the FOV. There is additional attenuation as the array directivity is rotated to the FOV edges during beamforming and will be proportional to the cosine of the angle from broadside (center). Similarly, directivity over the entire bandwidth (i.e., 76 GHz–81 GHz) often results in losses or reduction of the maximum range performance near the edge of the frequency band.

The radome adds losses due to the interaction of the electromagnetic waves and the radome materials. **Dielectric permittivity** is the measure used to consider which materials may be best suited for a radome. Most commonly, polytetrafluoroethylene (PTFE) and polycarbonates are used with attenuation on the order of 1–2 dB.

The number and routing of antenna elements to the RFICs will have losses on the order of 2–3 dB. PCB design of a radar plays a critical role in minimizing **trace losses**.

Windowing, or weighting, is a signal processing technique used to reduce the spreading of signals into adjacent frequencies or ranges bins (side lobes) around the center frequency (main lobe). While the benefit is better detectability performance, windowing does incur a reduction in signal power.

Table 2 Typical losses in automotive radars

Loss	Approx. value (dB)
Antenna Gain Drop at the edge of FOV	**6**
Antenna Gain Drop at the edge of frequency	**4–6**
Array gain loss	**2–3**
Radome loss	**1–2**
Trace loss	**2–3**
Windowing loss	**6**
Straddling loss	**3**
Attenuation of rain @ 200 m	**2**

Limitations to the signal sampling capabilities, number of range gates, number of doppler cells and azimuth/elevation beams are responsible for **straddling losses** in the system.

Weather also reduces the signal power, although usually relatively small. In the case of rain, the electromagnetic wave will have some propagation loss on the order of 1–2 dB at longer ranges.

4.1.5 Multipath Effect

In radar applications, the **multipath effect** is the phenomenon in which a reflection is received by the antenna along multiple paths. It is a natural phenomenon of radar technology and is represented as *F* in the radar Eq. (16). In some scenarios this behavior is beneficial for non-line-of-sight (non-LOS) detection. For example, a line of vehicles in front of the ego vehicle (Fig. 14a), or a target approaching an obstructed intersection (Fig. 14b). However, multipath behavior is also responsible for some adverse effects such as **intermittent detection** and **ghost targets**.

Intermittent Detection

Reflected signals have the potential to combine **constructively** (same phase) and will increase signal power or **destructively** (180° out of phase) and will decrease signal power. When the signals are combined destructively, the decreased power will result in a momentary loss (blind spot) of detection, known as **intermittent detection** (Fig. 15).

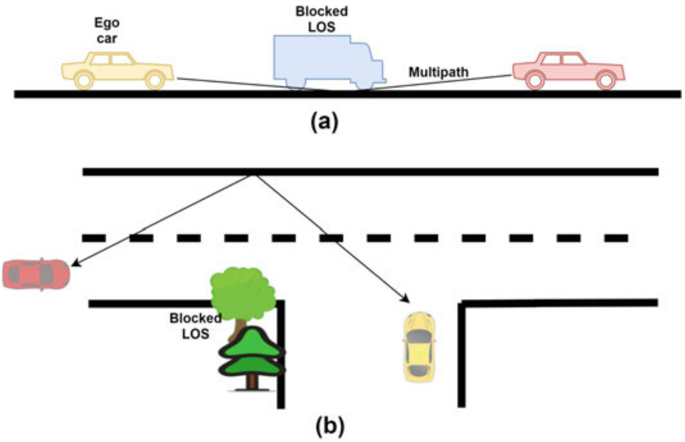


Fig. 14 Non-line-of-sight detection due to multipath effect

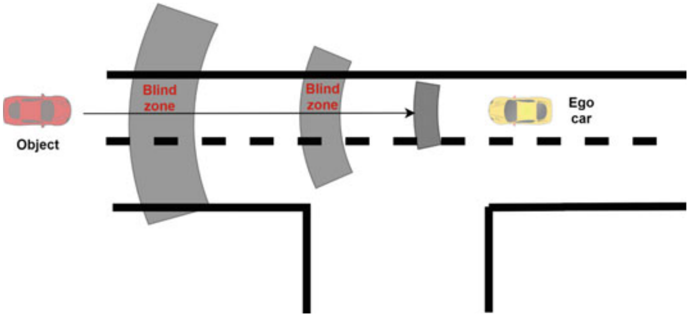


Fig. 15 Intermittent detection due to multipath effect

The severity of the multipath effect depends on many factors such as target location, radar placement and position, target's surface, radar polarization and operating frequency. The most important of which is the radar's placement and position on the vehicle, specifically regarding the relationship to ground. The highly reflective nature of the road will result in higher intermittent detection as the radar relative height is increased [17]. At higher positions on the vehicle, the detection range of the radar will improve.

Ghost Targets

Multipath reflections from other targets or objects in the FOV can create the appearance of a target in the wrong location, a **ghost target**. This phenomenon can appear as a **mirror target,** in which the real target and the ghost target move as if they are mirroring each other, or a **shadow target,** in which the real target is "shadowed" by one or many ghost targets (Fig. 16).

4.2 Target Reflectivity

A target's reflectivity is known as its **radar cross section (RCS)**, is represented by σ in the radar Eq. (16) and is measured in dBsm (decibel square meter). The RCS parameter accounts for how much of the incident wave will be reflected by the target. Different targets will have different reflectiveness dependent on the incident angle (Fig. 17). In general, for a given frequency and angle the RCS of a target can be measured to support the calculation of the link budget.

In the case of a typical sedan [18], the RCS can change from 10 to 30 dBsm depending on the angle of incidence of the reflection. Similarly for a human [19], the RCS can change from -8 to -3 dBsm. For calibration of a radar, it is standard practice to use a predefined trihedral reflector of known RCS. Table 3 shows typical RCS results of various targets.

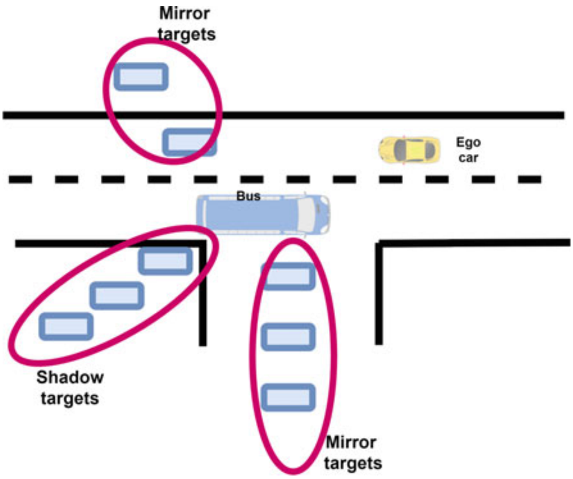


Fig. 16 Shadow and mirror ghost targets due to the multipath effect

4.3 Processing Gain

Processing gain (PG) is the cumulative effect of the range processing, doppler processing and beamforming gains. By using signal processing algorithms, the PG can reduce the effective BW of the received signal, thus reducing the power of the noise and improving the SNR. In some cases, **pre-** and **post-detection** techniques are used to improve the SNR further by integrating coherent and noncoherent signals.

Once these effects have been calculated, the radar link budget analysis can estimate the detectability of the radar at a specific range. Figure 18 shows the typical detectability output of a sedan using a relatively small radar.

It is important to note that Fig. 18 assumes a highly reflective ground which is the source of the several blind spots at some distances. It is also assumed the radar requires at least 20 dB SNR to detect targets.

5 Challenges and Solutions for Automotive Radars

In addition to the multipath effect, there are other significant challenges to automotive radars:

1. Interference
2. Over- and underclustering
3. Classification capability
4. Lack of resolution.

Fig. 17 RCS of different targets **a** typical sedan [18] and **b** human [19]

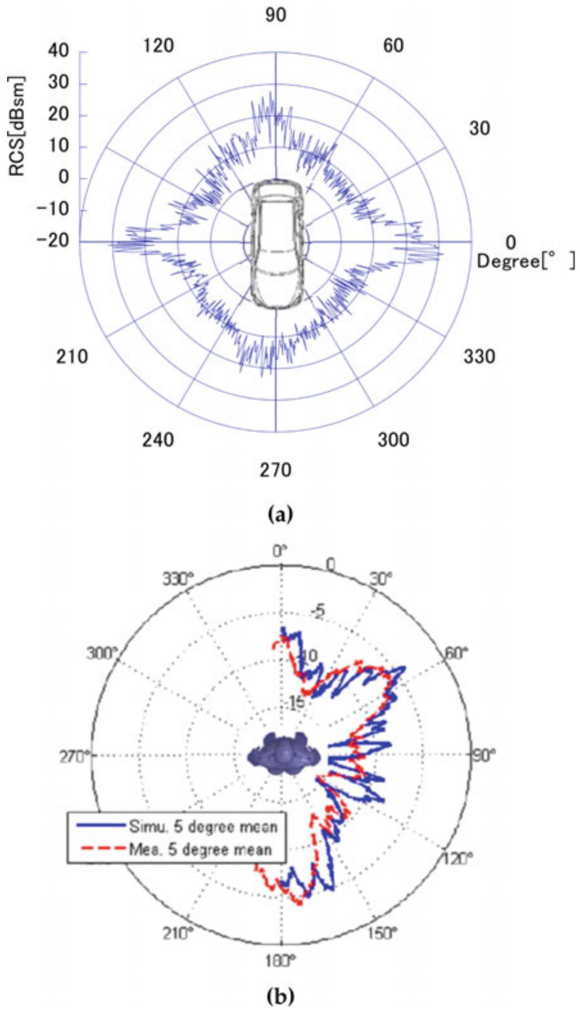


Table 3 Expected RCS of various targets

Target type	RCS, front (dBsm)	RCS, side (dBsm)
Human	−8	−3
Bike	−4	0
Motorbike	0	8
Bike + Human	1	9
Car	15	25

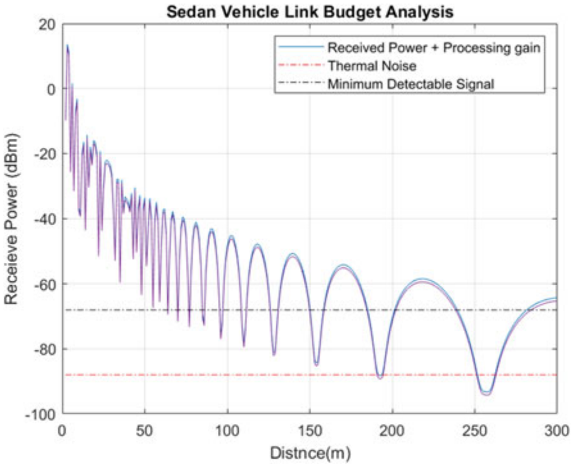


Fig. 18 Detectability output of a typical sedan over range

5.1 Interference

Intra-radar interference occurs when several transmitters of the same radar are transmitting at the same time. Many MIMO designs are challenged by intra-radar interference, slow phase coded division (SCD-MIMO), fast coded division (FCD-MIMO), doppler division multiplexing (DDM-MIMO), range division multiplexing (RDM-MIMO), frequency division multiplexing (FDM-MIMO) and polarization division multiplexing (PDM-MIMO). Any radar in which transmitting antennas are transmitting at the same time will result in some leakage between TX-MIMO channels in the range-doppler plane. To avoid this effect, a TDM-MIMO approach will lower the SNR; however, it will also lower doppler ambiguity [20].

Intra-vehicle interference is brought on by unwanted interaction between different radars on the same platform or vehicle when more than one radar is transmitting nearly simultaneously at the same frequency. The resulting effect will be receiver saturation of randomized detections along the signal's angle of arrival, effectively creating a detection **blind spot** (Fig. 19).

Intra-vehicle interference detections typically have random doppler measurements; therefore, the effect can be mitigated using constant false alarm rate (CFAR) algorithms. Some radar design and operation techniques will also mitigate the intra-vehicle interference such as dithering frequency (PRI, phase, starting frequency) or interleaved radars (time, polarization, frequency).

Inter-vehicle interference is the unwanted signal interaction between radars on different vehicles and is often referred to as **cross-interference**. This type of interference is momentary on the order of a few transmitting chirps, usually less. Similar to intra-vehicle interference, the result is a blind spot along all ranges over an angle of detection (Fig. 20).

Cross-interference detections have randomized doppler measurements, and therefore, CFAR algorithms can be used to mitigate the effect. To avoid inter-vehicle

Fig. 19 Intra-vehicle interference effect

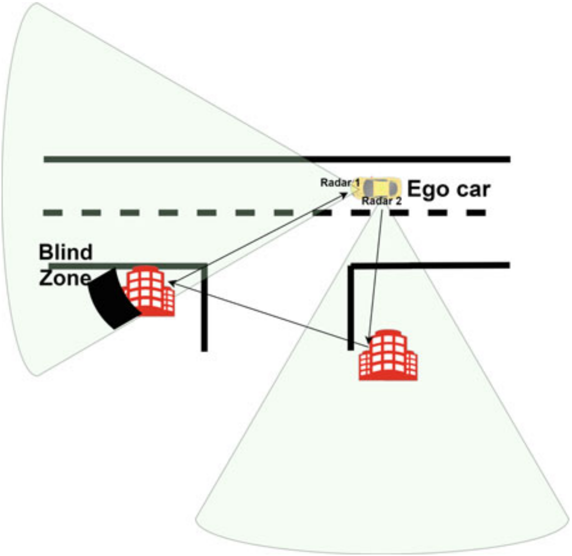
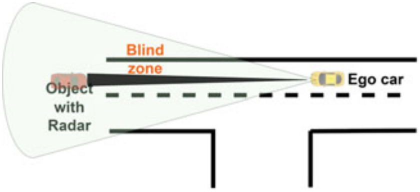


Fig. 20 Cross-interference effect due to inter-vehicle interferences



interference, frequency hopping, slow or fast coding, transmit randomization and other techniques can be adopted [21].

Many newer techniques such as adaptive nulling, multi-polarized radars, ADC smoothing (reduce peak to average power ratio) and reconstruction have been developed.

5.2 Under- and Overclustering

When two targets are clustered as one, it is known as **underclustering**. Depending on the clustering algorithms used, the source of underclustering is likely due to lack of resolution performance in angles or the **same doppler effect** when the relative motion of the radar and two targets is the same. This scenario is a consistent challenge in radar processing. Figure 21 shows a single clustered object between the two vehicles. Ideally, each vehicle would have its own cluster.

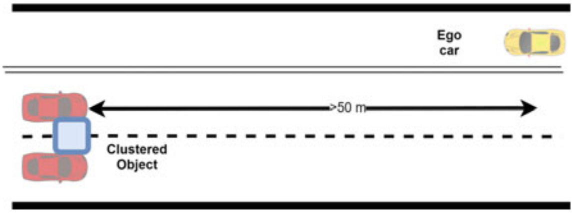


Fig. 21 Example of underclustering

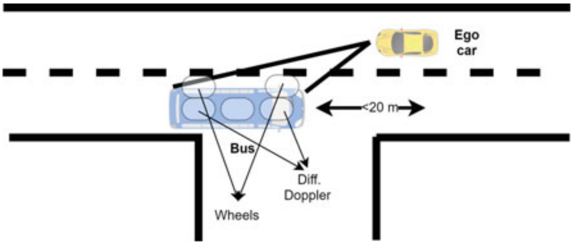


Fig. 22 Example of overclustering

Improvement in resolution of any detection domain (range, azimuth, elevation, doppler) will reduce the likelihood of underclustering as well as potentially more advanced clustering algorithms [22]. As this chapter has discussed, the improvement of resolution, in any domain, has inherent challenges and tradeoffs in computation burden, design and complexity.

Overclustering issues occur when a single target results in multiple clusters (Fig. 22). The variation in the point detections of a target is the main source of overclustering. That is to say there is variation in the doppler measurement of the point detections of the same target. For example, the wheels of a vehicle will have a distinct doppler compared to the body of the vehicle. This makes it difficult to clearly discern the extents of the target **(bounding box**). Strong clustering algorithms can be combined with machine learning techniques to help mitigate overclustering.

5.3 Classification

Target recognition is a nascent field in radar science. The interest in being able to use radar point clouds for classification is a fairly recent requirement. As ADAS systems advance to higher levels of automation (SAE L1-L5 footnote), the need to discern more information about the target has become crucial. In tandem, the advancements in higher definition radar have enabled the development of new study and methods of classification.

Classification can be applied on a per-scan basis or over multiple scans as part of the data processing path. It is common to leverage feedback from other data processing methods, such as tracking to improve classification performance.

The **range profile** is a 1D view of the detections associated with a specific target. The higher the resolution performance, the clearer the range profile. However, obstruction of a target by other objects or by the target itself (**shadowing**) as well as multipath effects limits the usefulness of the range profile for classification.

A **cross-range profile** is an extension of the range profile over azimuth and/or elevation. Similarly, better resolution radar will enable better classification performance.

The **range-doppler profile** adds the doppler domain information to enable analysis of **structural motion** (i.e., wheel spinning) and **bulk motion** (vehicle turning) of a target. This can be particularly useful for artificial intelligence and machine learning approaches to classification. The range-doppler profile of a target over time is also referred to as the **micro-doppler profile**.

Some widely used classification techniques for automotive radar include decision treeing, **support vector machine** learning models (SVM) and **convolutional neural networks** (CNN).

5.4 Lack of Resolution

Having low resolution in azimuth and/or elevation angle will reduce the capability of using radar for target classification and can be the source of other issue such as underclustering. Although targets may not be able to be separated in azimuth or elevation based on resolution alone, it is possible to separate them by utilizing other domains such as range and doppler. In general, resolution can be broken into two types: **static resolution** and **dynamic resolution**.

5.4.1 Static Resolution

Static resolution is defined as the capability of a radar to separate two targets with the same reflectivity in a single domain while all other domains are exactly the same. This is not a particularly useful measurement as it is not representative of the vast majority of real-world scenarios. Especially if the ego system is moving, the static objects in the environment will have different doppler (radial velocity) due to the effects of relative angular motion. Static resolution can however be calculated using the formulas (2), (10) and (12).

To achieve better static angular resolution, one can use techniques like TX-MIMO and RX-MIMO explained in the previous sections.

5.4.2 Dynamic Resolution

There are two applications of dynamic resolution: **cross-domain resolution** and **adaptive resolution**.

Cross-domain Resolution

Cross-domain is defined as the capability of a radar to separate two targets in a single domain while there is additional separation in another domain by at least 1 bin.

For example, in measuring resolution in azimuth angle there are two static 10 dBsm reflectors in the same elevation angle and separated by (at least) 1 range bin.

In this specific case, the resolution of radar will be its accuracy. **Accuracy** of a sensor refers to how close the estimated value of the detection for a target is compared to its true value (azimuth, elevation, range and/or velocity). It is important to note that accuracy is not the same as resolution. It is possible to achieve better accuracy than the resolution would suggest using various signal processing techniques. Accuracy is significantly dependent on SNR and is defined as follows:

$$R, v, \theta_{\text{accuracy}} = \kappa \frac{\Delta R, v, \theta}{\sqrt{\text{SNR}}} \tag{20}$$

R range
v velocity
θ detection angle

where κ is a constant and SNR represents the signal-to-noise ratio of the target detection. This relationship shows that high SNR achieves an accuracy that is better than resolution. When using a reasonable probability of detection and low probability of false alarm, a practical estimation of accuracy is on the order of one tenth of the resolution. The application of clustering and tracking algorithms can improve this relationship.

Adaptive Resolution

Adaptive resolution is the application of knowledge-based signal processing techniques that consider the environment and update the radar to achieve improved target separability.

For example, an autonomous perception system has different resolution requirements for radar when operating in a freeway environment as opposed to when in an urban environment. The radar signal processing techniques can take into account certain operational characteristics of the system to update the approach to resolution estimation.

The application of adaptive resolution is a part of a more general concept known as **software-defined imaging radar** (**SDIR**). SDIR delivers an imaging radar platform in which the functionality and performance are defined and managed by the software layer. This provides the ability to dynamically configure the sensor in real time and eliminates the need for application-specific radar hardware. The outcome is a highly adaptable, flexible and scalable imaging radar. A single software-defined radar platform has the potential to support many of today's automotive **short-range radar** (SRR), **mid-range radar** (MRR) and **long-range radar** (LRR) applications. Additionally, the SDIR software provides deeper data along the processing chain for feature development at the system level.

5.5 Data Fusion

Data fusion is the act of merging data from multiple sources as part of the general processing workloads. The expected outcome is more robust detection. Fusion can take place at various processing steps as well as between sensors of the same or different modalities.

5.5.1 Radar-Radar Fusion

As an example, an autonomous truck may be outfitted with a significant number of radars of different types throughout the system (Fig. 23).

In order to fuse the data from the radar sensor on the platform, they need to be calibrated and synchronized. Typically, radar calibration is performed in a controlled environment and is focused on the sensor hardware itself, **intrinsic calibration**. However, when utilizing multiple sensors, their orientation to each other as well as the local center of the system is also required, **extrinsic calibration**.

Over time as the system is in operation, it is possible for a sensor's calibration to drift. Traditionally to remedy this problem, the sensor would need to be recalibrated in a controlled environment (intrinsically) and in some cases manually (extrinsically). This is often referred to as **offline calibration**. However, **online calibration** would be the ability to recalibrate the sensor while it is in use for the application. This is extremely challenging as it usually requires highly accurate maps with pre-identified static options that can be used for reference during calibration. Should the map or the known position of the vehicle be unreliable, the online calibration will result in an error.

There are some new techniques that have been proposed for online calibration [23]. They are primarily based on the usage of doppler information or azimuth smearing effects on targets.

Data synchronization between different radars can be equally challenging. Using Kalman filters and its variants, track-to-track fusion-based methods, time interpolation and even some polynomial model-based estimation (even recurrent neural networks) can be used to ensure proper synchronization of data.

5.5.2 Radar-Other Sensor Fusion

In many ADAS features today, camera and radar data are processed together to support features such as lane keep assist, lane departure warning and automatic emergency braking.

There is a continuing discussion of data fusion between radar, camera and/or lidar for the next generation of ADAS features. Radar on its own provides significant benefits to any system through its ability to directly measure radial velocity and position in nearly all environments and at potentially longer ranges when compared

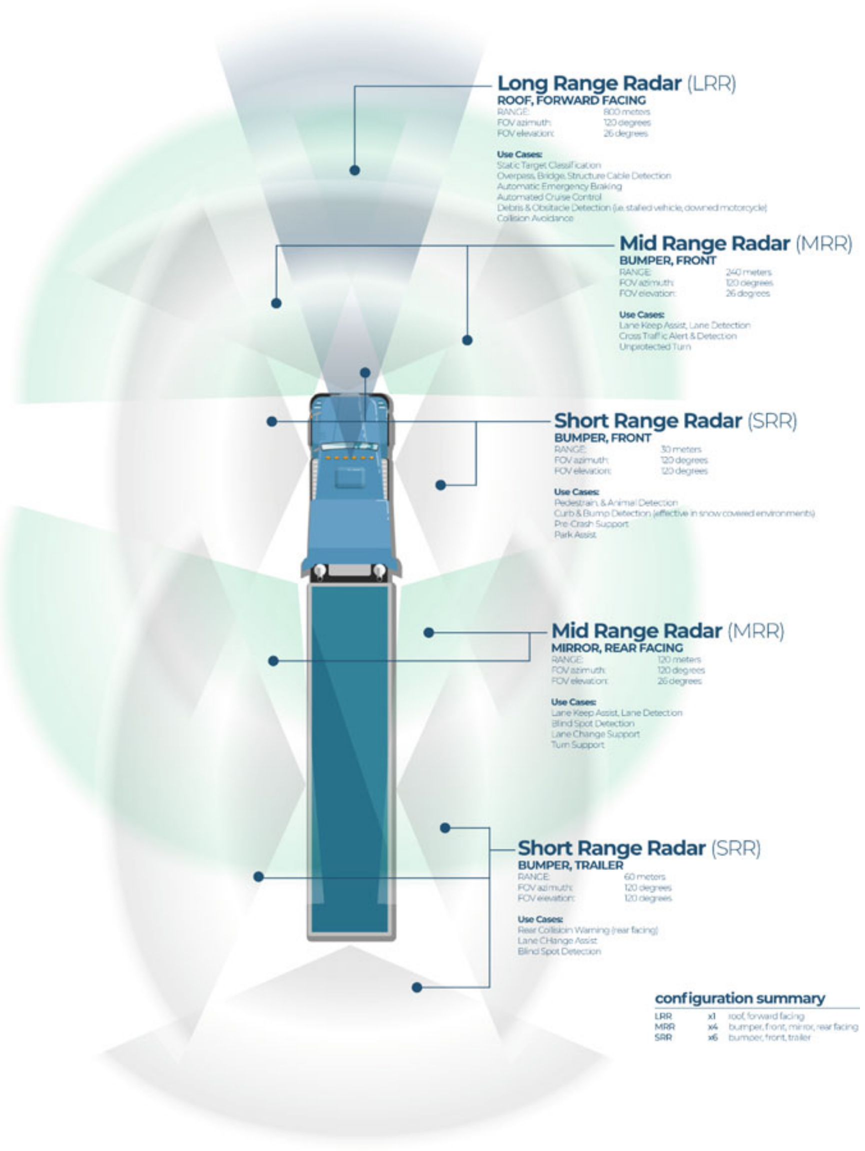
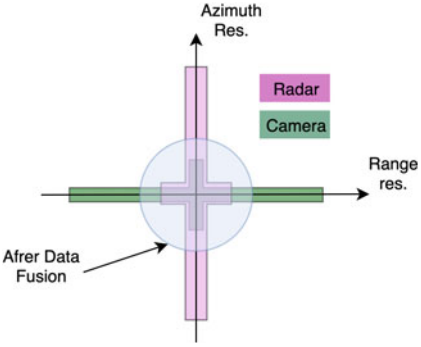


Fig. 23 Example of autonomous truck radar array

to camera. In contrast, cameras usually provide superior resolution in azimuth and elevation compared to radar. When taken together, or fused, the advantages of radar and camera can drastically improve the overall detection performance (Fig. 24).

Fig. 24 Camera versus radar



There are different approaches to achieve radar-camera fusion. **Early fusion** is the approach of considering the data deeper in the processing chain prior to determining an object's properties based on any single sensor modality. Conversely, **late fusion** is the consideration of the detected object properties after they are identified independently by each modality. Additionally, hybrid **fusion**, or a combination of early and late fusion, is being explored as well. Regardless of the approach, it is generally accepted that proper data fusion will provide the higher levels of detection performance required by the next-generation ADAS features.

5.6 Radar Integration

In this section, we will briefly review some challenges in integration, including mechanical challenging and packaging, software/firmware integration and radar placements.

5.6.1 Mechanical Challenges

Typical radar consists of the electronics (typically 1–2 PCBAs), the electronics housing, the radome and any required EMI shielding for the electronics. The radome is attached to the electronics housing to provide a sealed environment for the electronics. The antenna array, electronics and electrical connectors are mounted within the housing, which provides vibration protection and thermal management. The housing is typically made out of die-cast aluminum for thermal performance, but injection-molded plastic housings can be used if the radar is especially low power. The radome is the radio-transparent covering over the antenna array that keeps the radar sealed but allows the antenna array to transmit and receive. Radomes are typically made from injection-molded high-performance plastics for radio transparency. Any internal EMI shields are usually made from radio-absorbing plastic or radio-blocking metal.

Mechanical design of the radome is critical in order to maximize overall radar performance. The thickness, antenna gap and material selection all impact the radome's RF transparency. Because the transmitted signal passes through the radome twice, once on transmit and once on return, RF transparency of the radome should be maximized in order to maximize detection range and accuracy. The selection of the radome's thickness and spacing depends on the radome's material and the radar's frequency of operation. The exact required thickness and spacing are a system design question and will not be discussed here. In general, the lower the dielectric constant and dissipation factor of the radome, the better. Some common volume production radome materials are PBT-GF30 and PEI-GF20. While these materials do not have the best RF transparency, they offer a good balance of RF transparency, manufacturability and environmental compatibility. Other potential options include polycarbonate, Teflon and PPE (polyphenyl ether).

The optimum radome thickness and spacing will have tight tolerances in order to maximize radar performance. Thus, the injection-molded or CNC machined radome must be fabricated to a consistent thickness and designed to install with a consistent gap from the antenna array.

An optimally packaged radar requires tight collaboration across mechanical, electrical and radar system design domains. The radar's main mechanical design challenge is packaging the required antenna array and processing electronics in the required enclosure size. As the popularity of imaging radars increases, the market constantly demands smaller, higher resolution radars. Improved resolution and antenna array size are inherently at odds—the more angular resolution required, the larger the array must be if all other factors (e.g., signal processing and transmit wavelength) are kept equal. The mechanical designer must work with the radar system designer to find the appropriate balance between resolution and size. From an environmental point of view, automotive radars are required to operate over a wide range of harsh conditions. Every application will have its own set of environmental lifetime standards. These environmental standards include operational temperature and humidity range, thermal lifetime cycling, thermal shock, water and dust ingress protection, mechanical shock survival and vibration lifetime. The two largest environmental design challenges are water ingress and operational temperature range.

Typical ingress protection ratings for all-weather radars are IP68 or higher, which means that the typical radar is protected against full water submersion. A radar can use o-rings, co-molded elastomer seals, ultrasonic welding or adhesive to seal the radome to the enclosure. The electrical connector can either be directly co-molded into the housing or sealed like the radome. Sealing methods that use o-rings or co-molded elastomers tend to "pump" moisture into the enclosure from the outside environment due to pressure differentials arising from thermal cycles. Over time, moisture can build up inside the otherwise sealed housing with no way to escape, causing eventual electrical failure. Radars sealed this way require an ePTFE environmental vent to allow any accumulated moisture to escape while preventing liquid water ingress. Radars hermetically sealed with ultrasonic welds or adhesives do not require vents as they are not susceptible to this phenomenon.

Automotive radars are required to operate over a wide temperature range. A typical temperature rating could be from −40 to + 85 °C ambient. At low temperatures, plastics and sealing compounds should be designed and selected to avoid brittle failure, and at high temperatures, high glass transition materials should be selected to avoid creepage. Plastics are usually glass-reinforced, and common options are PBT-GF30 and PEI-GF20. The electronics housing must sufficiently cool the internal electronics at the maximum operating temperature. Given power consumption and maximum ambient temperature, the housing must provide a thermal path with sufficiently low thermal resistance from IC junction to ambient to keep the most sensitive electronics inside their operating temperature range. If the ambient temperature is 85 °C, the IC junction temperature can be 125 °C or more. The most sensitive ICs are typically the RF frontend and signal processing ICs. The RF performance of the frontend ICs degrades with increasing temperature, so good thermal performance is critical to maximize overall radar performance.

5.6.2 Radar Mounting

Placement and orientation of the radar on the vehicle can have significant implications. Any obstruction material (i.e., radome, body panel, fascia) will likely alter radar performance. There are high-frequency simulation tools available to assist in analyzing and identifying the best location for radar placement. Once installed, there are various testing techniques that can be used to adjust the integration and tune the radar parameters as needed. Factors to consider when integrating a radar include the following:

- Antenna/radar orientation [24]
- Radar distance to the edge of the platform
- Radar distance to the roof
- Radar tilt, yaw and roll on the platform
- Radar external radome materials (e.g., bumper).

5.6.3 Software/Firmware Integration

When considering safety in a radar system, the software stack should be considered both holistically and in a piece-by-piece approach. Software does not operate in isolation; it is dependent on the hardware safety features provided by the radar and the compute platform. In fact, levels of redundancy and monitoring in hardware are a prerequisite to sufficient safety from a software point of view.

A typical radar system will have two major software components: the **sensor software** (referred to as the embedded or firmware layer) and the **computer software** (which runs AI/ML applications and more complex and resource intensive signal processing algorithms). Each component can be broken up further depending on the system architecture.

The **embedded software** could run a light-weight operating system, such as a Real-Time Operating System (RTOS), which maintains the time and performance guarantees required for operation of the radar. The embedded software is also in charge of monitoring, fault detection, and determining if the radar is in a working state, as well as doing low-level data ingestion and signal processing. In general, the embedded system should produce radar data and guarantee that the data are valid. For **SDIR-enabled** radar, the embedded software is also responsible for dynamic configuration of the radar based on environment, application and system requirements. For the embedded software, the system hardware should provide several layers of fault detection and redundancy. For memory protection, Error Correction Code (ECC) and Built-In Self-Test (BIST) can be used to detect memory fault or failures in sections of memory. In order to protect from unknown failure cases (software bugs, hardware faults or any unpredictable miscellaneous failures), a hardware **watchdog** should be used to reset the system to a clean state. The watchdog must be capable of detecting when the embedded software stops running and then resetting the hardware. Other safety-detection techniques such as voltage monitoring, temperature monitoring, radar block detection, radar bad data detection and other failure detection methods should be used and validated via **Fault Insertion Testing** (FIT) and **Design Failure Mode and Effect Analysis** (DFMEA).

After receiving radar data from the sensor, a **compute stack** is used to perform a perception algorithm on top of the data (whether it is a point cloud, cluster, tracks or some low-level data, e.g., ADC outputs). A compute stack is going to be responsible for ingestion, responding to system faults received from the sensor or the compute hardware and running the signal processing or AI/ML algorithms. The compute should provide at least a software watchdog to determine if the signal processing algorithms fail and should continuously monitor any failure notifications or reset that come from the embedded software. Importantly, the compute is responsible for determining if the radar reached some bad state or is sending bad data, validating the ingested data for errors (e.g., using CRC or TCP/IP techniques) and ensuring that the signal processing section also continues to run and provide valid data.

Zooming out, the software system is providing several key features for safety when using radar systems: monitoring the system for hardware, software or unknown failures, reporting an error when a failure is detected and resetting the radar to a clean state if a failure occurs. The embedded software and compute software must work together to determine if the radar is working correctly and to ensure that detections received at the output of the radar software are valid for use for higher level perception algorithms.

6 Summary

Radar is a critical part of ADAS applications today with many unique features and capabilities. However, radar continues to evolve and develop in large part due to advancements in integrated circuit design. As the needs and requirements of new

ADAS features continue to become more stringent, radar is presented with new opportunities and challenges, most notably:

- Multipath effect
- Higher resolution
- Interference
- Underclustering and overclustering
- Advanced classification.

Acknowledgements Ali Mostajeran, Kurt Fredlund and David Wu of our Zadar team in helping to clarify and establish my thoughts for this chapter. The editors would like to thank Dr. Xuefeng Wang from Cruise for his technical review.

References

1. A. Mostajeran et al., "A High-Resolution 220-GHz Ultra-Wideband Fully Integrated ISAR Imaging System," in IEEE Transactions on Microwave Theory and Techniques, vol. 67, no. 1, pp. 429–442, Jan. 2019, doi: https://doi.org/10.1109/TMTT.2019.2874666.
2. Federal Communications Commission Office of Engineering and Technology Laboratory Division, "EQUIPMENT AUTHORIZATION GUIDANCE FOR 76–81 GHz RADAR DEVICES", Apr. 2019, 653005 D01 76–81 GHz Radars v01r01.
3. P. P. Vaidyanathan and P. Pal, "Sparse sensing with coprime arrays," 2010 Conference Record of the Forty Fourth Asilomar Conference on Signals, Systems and Computers, 2010, pp. 1405–1409, doi: https://doi.org/10.1109/ACSSC.2010.5757766.
4. M. Emadi, J. Izadian, Ali Mostajeran, R. ZHANG, "Systems and methods for adaptive gating in initialization of radar tracking", 2021. [Online]. Available: https://patents.google.com/patent/US20210199792A1.
5. Xiao-Liang Yu and K. M. Buckley, "Bias and variance of direction-of-arrival estimates from MUSIC, MIN-NORM, and FINE," in IEEE Transactions on Signal Processing, vol. 42, no. 7, pp. 1812–1816, July 1994, doi: https://doi.org/10.1109/79.298289.
6. L. Yang, M. R. McKay and R. Couillet, "High-Dimensional MVDR Beamforming: Optimized Solutions Based on Spiked Random Matrix Models," in IEEE Transactions on Signal Processing, vol. 66, no. 7, pp. 1933–1947, 1 April1, 2018, doi: https://doi.org/10.1109/TSP.2019.2799183.
7. M. Emadi and K. H. Sadeghi, "DOA Estimation of Multi-Reflected Known Signals in Compact Arrays," in IEEE Transactions on Aerospace and Electronic Systems, vol. 49, no. 3, pp. 1920–1931, July 2013, doi: https://doi.org/10.1109/TAES.2013.6558029.
8. Y. Sun, P. Hu, J. Pan and Q. Bao, "An IAA-based DOA Estimation Method for PBR in Coherent Environment," 2019 Photonics & Electromagnetics Research Symposium - Fall (PIERS - Fall), 2019, pp. 3050–3055, doi: https://doi.org/10.1109/PIERS-Fall48861.2019.9021594.
9. Mohammad Emadi, K. H. Sadeghi, Amir Jafargholi, and F. Marvasti, "Co Channel Interference Cancellation by the Use of Iterative Digital Beam Forming Method," Progress In Electromagnetics Research, Vol. 87, 89–103, 2009. DOI:https://doi.org/10.2528/PIER08100403.
10. Mohammad Emadi, Ehsan Miandji, Jonas Unger," OMP-based DOA estimation performance analysis", Digital Signal Processing, Volume 79, 2018, Pages 57–65, https://doi.org/https://doi.org/10.1016/j.dsp.2019.04.006.

11. M. Viberg and H. Krim, “Two decades of statistical array processing,” Conference Record of the Thirty-First Asilomar Conference on Signals, Systems and Computers (Cat. No.97CB36136), 1997, pp. 775–777 vol.1, doi: https://doi.org/10.1109/ACSSC.1997.680549.
12. Sina Naderi Shahi, Mohammad Emadi, and K. H. Sadeghi, “High Resolution DOA Estimation in Fully Coherent Environments,” Progress In Electromagnetics Research C, Vol. 5, 135–148, 2009.
13. Skolnik, M.I, “Radar Handbook, Third Edition”, McGraw-Hill Education, 2008.
14. S. Cao and N. Madsen, “Slow-time waveform design for MIMO GMTI radar using CAZAC sequences,” 2018 IEEE Radar Conference (RadarConf18), 2018, pp. 1456–1460, doi: https://doi.org/10.1109/RADAR.2019.8378779.
15. Texas Instruments. “AWR2243 Single Chip 76 to 81 GHz FMCW Transceiver”, 2020.
16. Lydi Smaini, “RF Analog Impairments Modeling for Communication Systems Simulation: Application to OFDM-based Transceivers”, 1st Edition, Wiley, 2012.
17. J. Izadian, M. Emadi, “Mitigating multipath effect on radars for effective target detection”, 2021. [Online]. Available: https://patents.google.com/patent/US20210199760A1.
18. T. Motomura, K. Uchiyama and A. Kajiwara, “Measurement results of vehicular RCS characteristics for 79GHz millimeter band,” 2018 IEEE Topical Conference on Wireless Sensors and Sensor Networks (WiSNet), 2018, pp. 103–106, doi: https://doi.org/10.1109/WISNET.2019.8311576.
19. M. Chen, M. Kuloglu and C. Chen, “Numerical study of pedestrian RCS at 76–77 GHz,” 2013 IEEE Antennas and Propagation Society International Symposium (APSURSI), 2013, pp. 1982–1983, doi: https://doi.org/10.1109/APS.2013.6711649.
20. S. Sun, A. P. Petropulu and H. V. Poor, “MIMO Radar for Advanced Driver-Assistance Systems and Autonomous Driving: Advantages and Challenges,” in IEEE Signal Processing Magazine, vol. 37, no. 4, pp. 98–117, July 2020, doi: https://doi.org/10.1109/MSP.2020.2978507.
21. J. Izadian, M. Emadi, “Reducing Radar Signal Interference based on Semi-random and Random Configuration”, 2021. [Online]. Available: https://patents.google.com/patent/US20210190901A1.
22. M. Emadi, J. Izadian, Ali Mostajeran, R. ZHANG, “Sequential clustering”, 2021. [Online]. Available: https://patents.google.com/patent/US20210200209A1.
23. M. Emadi, J. Izadian, Ali Mostajeran, R. ZHANG, “Systems and methods for blind online calibration of radar systems on a vehicle”, 2021. [Online]. Available: https://patents.google.com/patent/US20210199759A1.
24. J. Izadian, M. Emadi, “Adaptive tilting radars for effective vehicle controls”, 2021. [Online]. Available: https://patents.google.com/patent/US20210199758A1.

Electrochemical Power Systems for Advanced Driver-Assistant Vehicles

Wen Li

Abstract Battery packs have been widely used as the main power source for advanced driver-assistant vehicles. The status and challenges related to electrochemical batteries, including material choices, energy density, performance, battery design, safety, reliability, cost, and development trend, are reviewed and addressed in this chapter. Meanwhile, other types of the electric power sources, such as fuel cells and capacitors, and challenges related to power management systems, are briefly introduced and discussed.

Keywords Battery · Battery management system · Fuel cell · Capacitor

1 Introduction

The increasing worldwide energy consumption (Fig. 1) has created enormous interest in alternative energy sources such as solar, wind, nuclear, biofuels, and hydrogen. In the recently released study from International Energy Outlook (IEO) 2019, the world energy consumption will grow by nearly 40–80% between 2018 and 2050 based on a survey from different countries.

Among all types of energy consumed every year in the world, that for transportation accounts for 20–35% of the total energy used every year based on the study from IEO2019. The internal combustion engine has been the main power source for the automotive vehicle, which uses one of the fossil fuels products (gasoline, diesel) to provide power for the internal combustion engine. There are some negative impacts associated with internal combustion engines, such as environmental pollution and CO_2 emission which causes the world average temperature to increase over the last 100 years.

The ideal alternative energy source should be easily accessible, environmentally friendly, and cost competitive compared to the existing energy supplies. Electric energy, which is considered as green and sustainable energy, has drawn more and more attention in recent years for transportation application and has become a

W. Li (✉)
Google, Saratoga, CA, USA

Y. Li and H. Shi (eds.), *Advanced Driver Assistance Systems and Autonomous Vehicles*,
https://doi.org/10.1007/978-981-19-5053-7_10

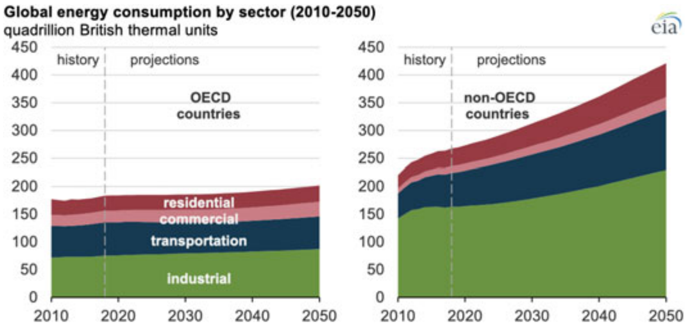


Fig. 1 Global energy consumption by section (2010–2050). (*Source* US Energy Information Administration, International Energy Outlook 2019). OECD: The Organisation for Economic Co-operation and Development (https://www.oecd.org)

promising alternative in the transportation industry. With the successful commercialization of lithium-ion rechargeable batteries in the 1990s, the gaps of the energy density and power density between gasoline-engine and battery are further minimized, which makes it possible that battery-powered electric cars provide comparable performance and operation range compared to combustion engine cars with reasonable cost.

In the last few decades, governments, especially the USA, China, and countries in Europe, have been providing strong support to develop electric-powered vehicles. Auto companies across the world have devoted great efforts to push for wider commercialization of electric-powered automobiles. Some German automakers (e.g., Volkswagen) even declared plans to go all electric by 2026 [1].

There are many different types of electrochemical energy storage/conversion devices. Among them, batteries, capacitors (electrochemical storage devices), and fuel cells (electrochemical conversion devices) are the three major types of the electrochemical devices that are widely used as power supply for transportation applications. Figure 2 compares the energy density and power density of those three different types of electrochemical devices vs. internal combustion engine [2].

In this chapter below, the status, performance, cost, and challenges related to those three types electrochemical devices will be discussed with more focus on batteries.

2 Batteries

2.1 Introduction

After Italian physicist Alessandro Volta built and described the first electrochemical battery in 1800, battery technology has penetrated to almost all areas related to people’s daily life. Battery-powered vehicles have been used since 1834 when

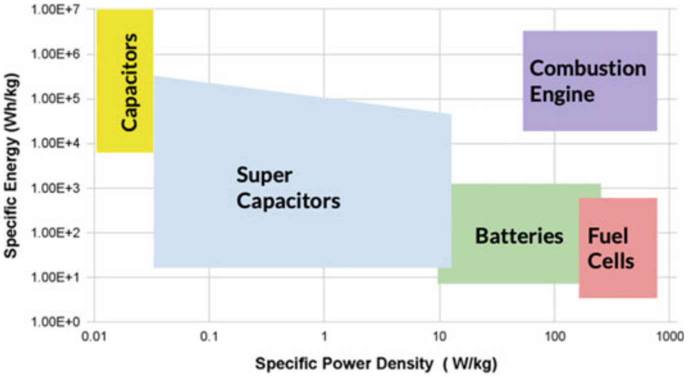


Fig. 2 Comparison of energy density and power density

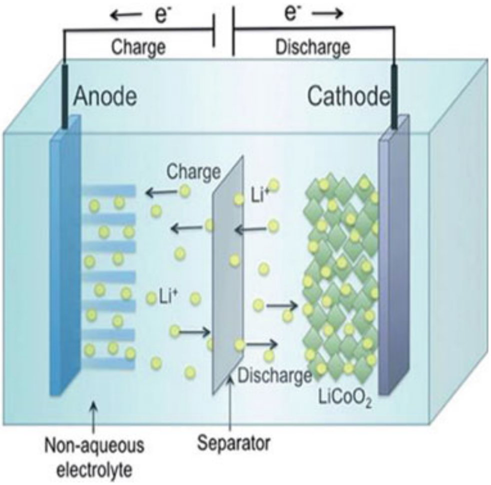
Robert Anderson first invented a crude battery electric-powered carriage. However, at that time, the battery had to be replaced after each use because of the absence of rechargeable batteries. In 1859, the first rechargeable battery, lead–acid battery, was invented in France by Gaston Plante. Vehicles powered by rechargeable batteries were also developed along with the development of new battery technologies. As shown in Table 1 listed, the rechargeable batteries showed relatively lower energy density compared to gasoline (>2000 Wh/kg) and higher cost. Therefore, batteries have not been widely used by the public for a long time.

Batteries are energy storage devices, which storage electric energy to chemical energy during charge and release chemical energy to electric energy during discharge. Figure 3 [4] shows structure of Li-ion battery and working mechanism during a cycle.

Table 1 Properties and application summary of different chemistries used in rechargeable batteries [3]

Specification	Lead-acid	Ni–Cd	Ni-MH	Metal alloy		
				$LiCoO_2$	LiMn2O4	$LiFePO_4$
In use since	Late 1800s	1950	1990	1991	1996	1999
Specific energy density (Wh/kg)	30–50	45–80	60–120	120–270	100–250	90–220
Nominal discharge voltage	2 V	1.2 V	1.2 V	3.6 V	3.8 V	3.3 V
Cycle life (cycles to reach 80% capacity)	200–300	1000	500–800	500–1000	500–1000	1000–2000
Typical charge time	6–12 h	1–2 h	1–2 h	1–2 h	1–2 h	1–2 h
Transportation application example	Golf cart	N/A	Prius hybrid	N/A	Nissan leaf	Electric bus

Fig. 3 Schematic of charge and discharge process of Lithium-ion batteries



During charge process, a charging circuit applies a voltage higher than the equilibrium voltage between cathode and anode, forcing a current to flow from cathode to anode, the Lithium ion with positive charge migrates from cathode to anode side through the non-aqueous electrolyte (containing lithium salt) and pores within the separator, then embedded in the inter-layers within the material particles in anode electrodes. During discharge, Lithium ion moves in reversed direction (from anode to cathode side), and the electrons move from anode to cathode to balance the charge.

Like any battery packs, the battery pack for electric-powered vehicles (EV) consists of two main parts: battery cells (Cells) and battery management system (BMS).

- Battery cell is the device used to store the energy during charge and provide energy during the discharge process. The main components in battery cells are two electrodes (cathode and anode), and separator (in-between the two electrodes), and electrolyte (lithium salt dissolved in organic solvent). As discussed earlier, the typical discharge voltage of a Li-ion battery is between 3.2 and 3.7 V depending on the chemistry used. However, the typical voltage of battery packages for electric vehicles is 300–400 V, so individual cells have to be connected in series to reach the target working voltage, then connected in parallel to provide the capacity as needed by pack design.
- BMS is the system comprising both hardware and software to control the charge and discharge process of the battery pack, to maintain the battery operating at proper temperatures, and to provide the protection needed to ensure the safety and the reliability of the battery pack under normal and abuse conditions. We will discuss this topic in more detail in the later part of this chapter (Fig. 4) [5].

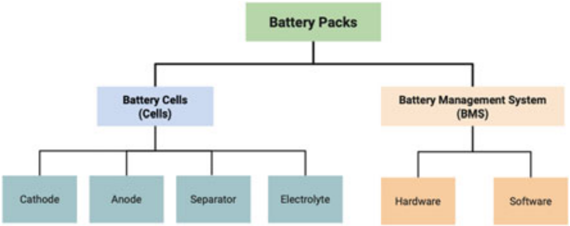


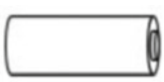
Fig. 4 Components within battery packs for EV applications

2.2 *Types of Battery Cells*

As indicated in Table 2, based on the external structure of the battery cells, there are mainly three types of battery cells that are widely used in electric vehicles: cylindrical, prismatic, and pouch. Some key information comparison.

- Cylindrical cells are named by the shape of the cell. Cylindrical cells were widely used in laptop computers in 1990s with standardized size (e.g., 18,650, 18 mm-diameter and 65 mm-height) and have been manufactured in very large quantities for a long time. It is well known that the first few generations of Tesla EV were using 18,650 batteries with individual cell capacity of about 3.5 Ah to make battery packs and later switched to 21,700 cells (21 mm-diameter and 70 mm-length) because the new size could provide higher individual capacity (3–5.1Ah) compared to 18,650. Most of the cylindrical cells have good mechanical strength because metals can be used as casing and have built-in safety features such as venting feature and short-circuit protection.
- Prismatic cells are also named by the shape of the cell. Prismatic Li-ion cells were first developed for consumer electronics (e.g., phone, camera) application using rectangular metal can as casing and were widely used in 1990s–2000s

Table 2 Summary of the three types of battery cells based on external structure

Cell type	Cylindrical cell	Prismatic cell	Pouch cell
			
Casing	Hard	Hard	Soft
Shape	Cylindrical	Rectangle	Rectangle/irregular
Size	Fixed size	Flexible	Flexible
Capacity	Low individual	Low to high Individual	Low to high individual
Cost	Low	Medium	Low to medium
Makers	Panasonic/LGC	LGC/SDI	LGC/CATL
Notes	Built-in safety feature	Higher manufacture cost	Dimension change in-use

as replaceable batteries for cell phones. However, after Apple introduced the concept of the built-in non-replaceable battery, the demand for prismatic cells dropped year-over-year dramatically. However, because of the mature process and good individual mechanical strength, more and more efforts have been made to optimize this cell format for EV applications. Prismatic cells can be manufactured to different sizes to achieve much higher individual cell capacity compared to cylindrical cells.

- Pouch cells are also called polymer cells. The name refers to the polymer casing material used to manufacture the cells. Since the polymer material can be used to form a pouch bag with almost any shape as the battery casing, the pouch cells have more flexibility to manufacture, even irregular-shaped cells can be made with polymer pouch casing. The L-shaped battery in the iPhone is a good example to demonstrate the potential of the cell format. However, because of the soft casing, the mechanical protection design is more critical at the pack level design when using pouch cells to make battery packs for EV applications. In addition, since the overall volume of the electrodes inside the battery increases during charge and decreases during discharge, and the overall volume keeps increasing during cycles, the dimension change of the cells during usage will have more impact on the overall pack design and performance during usage.

2.3 Battery Cell Internal Structure

We have discussed the external shapes and structures of the battery cells and their impact on the performance and safety. Another one important point related to the performance of the battery cells is the internal structures. There are mainly two different internal structures of the battery cells: the jellyroll structure and the stacking structure. Figure 5 provides the illustration of both structures. Table 3 also summarizes the impact of the internal structures on the performance of the battery cells, which is as important as external structures discussed in previous paragraphs.

- The jellyroll structure is formed by the winding process where the machine winds the two electrodes and separator together. The winding process is designed as a continuous process to improve the production efficiency and minimize the manufacturing cost. However, since the jellyroll cannot have intimate contact with the four corners of a rectangular casing, the corner space is wasted, and the package efficiency of the overall cell is relatively low.
- Stacking structure is formed by the stacking process where the machine stacks the electrode and separator layer by layer. The stacking speed is relatively slower compared to the winding process. The overall manufacturing cost of the stacking process is higher compared to the winding process. However, benefits coming from this structure are also significant: 1. The stacking structure has a higher packing efficiency at the corners, especially for thicker cells; 2. Each layer can be made separately, therefore an irregularly shaped electrode can be made, allowing

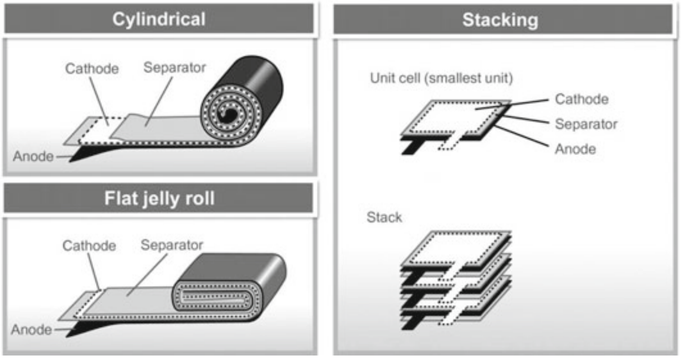


Fig. 5 Illustration of internal cell structures

a much flexible cell format; 3. Since there is electrode terminal on each layer of the electrode, the electronic path length within each electrode is much shorter in the stacking structure compared to the jellyroll structure, which provides much better high-power performance for the cells.

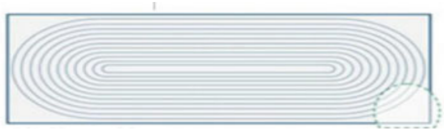
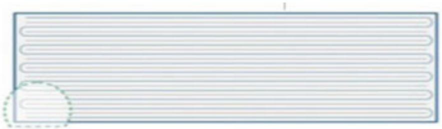
2.4 Battery Cell Manufacturing Process

As discussed earlier, battery packs for electric vehicle applications are assembled from hundreds or thousands of the individual battery cells. The reliability and pack usage life directly depend on the consistency of individual cells. The cell manufacturing process is the key part to achieve consistent capacity, internal impedance, and performance behavior of the cells.

As shown in Fig. 6, the Li-ion pouch cell manufacturing process can be roughly divided into three steps [6]:

- Electrode-making process: The active material in cathode or anode is mixed with the binder, conductive agent, and dispersion agent to form a slurry. The slurry is then coated on the metal foil substrate (aluminum for cathode and copper for anode) when dispersion agent is evaporated at the same time. The coated electrodes are further pressed by roller press to achieve the targeted thickness and further slitted to smaller width per the designed value. In the electrode-making process, the terminal tabs and insulation tapes were attached to the electrode based on the design. The finished electrodes are further dried to minimize the remaining moisture content in the electrodes before the cell assembly process. The electrode-making process is the key to make cells with uniform capacity, as the consistency of the slurry and coating weight in unit area will directly impact the final capacity of each cell. Also, the electrode with defects could cause internal short or capacity fast fading during usage.

Table 3 Comparison of two internal cell structures (winding and stacking)

Internal structure	Winding (jellyroll)	Stacking
		
Energy density	Good for normal thickness	Good for low/high thickness
Max power	Good	High
Cell shape	Rectangular/cylindrical	Flexible
Manufacture cost	Good	High

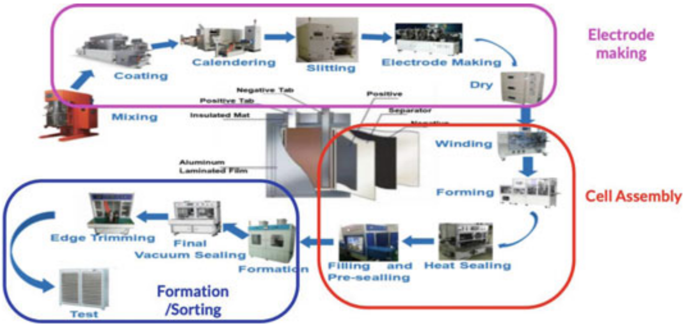


Fig. 6 Process flowchart of typical pouch cell manufacturing process

- Cell Assembly Process: Cathode, anode, and separator are used during this process to form the jellyroll or stacking structure first. Then jellyroll is inserted into the pouch bag which is formed during the forming process. The pouch bag with jellyroll inside is then heat sealed with one side opened. After another drying process to further remove the moisture in the pouch bag and jellyroll, the electrolyte is injected followed by a pre-sealing process to fully close the pouch bag. The cell assemble process controls the electrode alignment in the jellyroll which is key to ensure the reliability of the cell, since the misalignments of the jellyroll could cause shorting between electrodes.
- Formation and Sorting Process: The pouch cell with electrolyte and jellyroll sealed inside then goes through the formation process, which is to charge and discharge the cell using the fixture to apply the pressure and temperature. During the formation process, the stable solid–electrolyte interface (SEI) layer is formed in the cell to ensure the stable cell performance during usage. As by-product gas is usually generated during the formation process, vacuum sealing is needed to remove the gas inside the pouch bag, and extra pouch material in the sealing area is further trimmed. The finished cells are then tested to check for capacity, internal resistance, and some key dimensions. All the battery cells have to be characterized in the sorting process; the sorting data will later be used to match the cells during the module assembly process.

The battery cell manufacturing process has been improved dramatically during the last few decades. More and more automation processes have been introduced in almost every step, and real-time feedback with machine adjustment functions has been also applied in more and more processes. As a result, the quality and consistency of the cells have been continuously improved with higher production rate and improved yield rate, which have further driven down the cost of the manufacturing process.

2.5 Chemistry Choices for Li-Ion Battery for EV Applications

In Li-ion batteries, cathode and anode are the most important chemical materials. Various combinations of the material chemistries can be used to make battery cells. Each electrode material combination has its disadvantages and advantages in terms of performance, cost, safety, and other parameters.

2.5.1 Cathode Chemistry for Li-Ion Battery

When Li-ion was first commercialized in 1990s, lithium cobalt oxide ($LiCoO_2$ or LCO), which was invented by JB Goodenough [7, 8], was the cathode material, and graphite was the anode material. Dramatic improvements of the energy density, safety, and performance of Li-ion batteries have been made since then, and the improved LCO material is still widely used in Li-ion batteries for various applications, especially as the power supply for portable consumer electronic devices (smart phone, laptop, etc.). However, the structural instability of LCO in fully charged state could cause potential safety risk for EV applications. More importantly, the price of cobalt source has kept increasing due to the increasing worldwide demand and the limited overall available resources on earth, which has created a significant barrier for this material to be widely used in EV applications.

Compared to LCO, other cathode materials have become more popular for batteries for EV applications, not only because of their relatively more stable crystal structures at delithiated state, but also because of their relatively lower costs. Table 4 summarizes some key factors of these new materials: energy density, power density, safety, cycle life, cost, and industrial readiness. Detail discussion of each cathode material will follow [9].

Lithium Iron Phosphate ($LiFePO_4$, LFP)

LFP cathode was developed by Goodenough team and the first report was published in 1996 [10]. Compared with LCO, LFP has many very attractive advantages as a cathode material for EV applications:

Table 4 Comparison of different cathode materials for Li-ion batteries

Materials	LCO	LFP	LMO	NCA	NMC
Energy density	High	Low	Good	High	High
Power	High	Good	Good	High	Good
Lifetime	Good	High	Low	High	Good
Safety	Good	High	Good	Good	High
Cost	High	Low	Good	Low	Good
Average discharge potential	~ 3.8 V	~ 3.2 V	~ 3.9 V	~ 3.7 V	~ 3.7 V

- Excellent chemical and thermal stability, mainly due to its different crystal structure compared to LCO and other lithium transition-metal oxide cathode materials.
- Due to its stable crystal structure, cells made with LFP cathode have demonstrated excellent cycle life. Commercially available cells can achieve more than 2000 cycles with remaining capacity still higher than 80% of initial capacity.
- Very good safety performance and the ability to tolerate abuse conditions. This property is also due to the stable crystal structure.
- Very low raw material cost. The elements used to make LPF are widely available and abundant.

Unfortunately, LFP suffered from low electronic conductivity when it was developed initially, which impacted the power energy when high discharge rate was needed. Different approaches have been used to improve the overall electronic conductivity, such as surface coating with more conductive material (carbon, metal), making the particles to nano-size to shorten the electronic and ionic transfer length within the particles. With those improvements, LFP is able to achieve acceptable charge/discharge rate. A123 and BYD are two main players which manufacture material and cells with the LFP chemistry. Another disadvantage of LFP is the lower nominal voltage (~3.2 V per cell) during discharge, which reduces the specific energy density compared to other cathode materials.

LFP is widely used in many electric cars, such as BYD e6. But its properties make it more suitable for electric-powered buses and trucks for the following reasons: 1) the size and total energy of the LFP battery pack for those applications are much bigger compared to those of regular cars, and energy density is not as important as for smaller vehicles, and 2) LFP has excellent safety and cycle performance.

Spinel Lithium Manganese Oxide ($LiMn_2O_4$ or LMO)

Spinel LMO was first reported by Thackeray and co-workers in 1983 as a candidate to replace LCO for cathodes [11]. Some key advantages for LMO over LCO are:

- The spinel structures provide battery structure stability at the fully charged state and during charge/discharge cycles.
- The spinel crystal structure of LMO enables a better rate capability (Li-ion can diffuse between different crystal layers) compared to the LCO materials where the Li-ion can only diffuse within the cobalt-oxygen layer in the rock salt structure [11].
- No toxicity and lower cost when using manganese as a raw material.

Disadvantage: low specific capacity (<150mAh/g) and poor high-temperature performance, instability of the electrolyte at elevated temperatures, leading to Mn dissolution and capacity loss under high-temperature storage conditions.

The LMO cathode material is commercialized for mainly for power tool applications where the good high current discharge capability and cost–benefit of this material are fully utilized. For EV applications, LMO is usually blended with higher specific capacity cathode material, e.g., NMC (lithium nickel manganese cobalt oxide) to improve the energy density of overall battery packs. Nissan Leaf and Chevy Volt are two examples which use battery packs made with this approach to provide better capability of high current boost on acceleration while maintaining the comparable energy density as that used NMC without blending.

Lithium Nickel Cobalt Aluminum Oxide (NCA)

As an alternative to the LCO-layered cathode material, lithium nickel oxide ($LiNiO_2$ or LNO), has been intensively studied by researchers because of its higher theoretical capacity compared to LCO. However, the stoichiometric LNO with a Li/Ni ratio of 1:1 is difficult to synthesize because it often results in Li-deficient $Li_{1-x}Ni_{1+x}O_2$ with part of the Ni^{2+} ions in the Li layer [12] due to the similar ionic radii of Li^+ (0.76 Å) and Ni^{2+} (0.69 Å). The stoichiometry can be improved to $LiNi_{1-x}M_xO_2$, where M can be cobalt (Co), manganese (Mn), or aluminum (Al), by heteroatom doping. The heteroatom doping has been proved as the most effective way to maintain the layered structure and enhance the cycle stability [12, 13].

Lithium nickel cobalt aluminum oxide can be prepared by doping both Co and Al into LNO structure. Various doping ratios of Co and Al to Ni have been evaluated. $LiNi_{0.8}Co_{0.15}Al_{0.05}O_2$ is the optimized composition for NCA so far. The advantages of the NCA materials are described as follows:

- The highest specific discharge capacity. With a specific capacity of around 200 mAh/g, NCA delivers the highest specific discharge capacity among the current mature cathode materials for Li-ion batteries.
- Low cost. Since the price of Ni is lower compared to Co, the material cost of NCA is lower compared to those of high Co-content cathode materials.
- Good structure stability compared to LMO at high operation temperature.

Though the price of NCA is still higher than that of LMO, and the volumetric energy density of NCA is lower compared to LCO, the high specific capacity and high gravimetric energy density of this material still make it one of the successful cathode materials which are widely used in EV applications, for example, Model S, Model X from Tesla.

Lithium Nickel Manganese Cobalt Oxide (NMC)

NMC is a group of materials that were studied as alternatives to LCO materials. NMC materials have similar or higher specific capacities due to high Ni contents. They also have similar operating voltages because of the Mn presence in the crystal structures. Various combinations of Ni/Mn/Co ratio have been studied. $LiNi_{1/3}Mn_{1/3}Co_{1/3}O_2$(NMC-111) and $LiNi_{0.5}Mn_{0.3}Co_{0.2}O_2$ (NMC-532) both are commonly used in Li-ion batteries for EV applications due to the following advantages [14, 15]:

- Higher specific capacities compared to LCO and LMO.
- Lower costs compared to LCO or NCA, since Mn is even cheaper compared to Ni.

- Since Mn tends to form spinel structure in the NMC mix, the low overall resistance of NMC makes them the lowest self-heating material during charge and discharge.

The NMC-blended materials will be chosen for the next-generation Li-ion EV battery cathodes. The ratio of Co, Ni, Mn in the blended materials can be tuned based on demands on different aspects: energy density, performance, stability, and cost. The same strategy can also be pursued further for energy storage applications that require longer cycle lives. Some of the Ni-rich cathode materials (NMC-811, NMC-622) have demonstrated the potential for EV applications in the near future because of their even higher specific capacities and lower costs [16, 17]. Nonetheless, the challenges in material stability during the cycle still need to be addressed before their deployments can be successful.

Another group of promising next-generation materials is the high-voltage spinel materials with average operating voltages around 4.0–4.1 V (3.8 V for LCO as reference). With increased operation voltage, the energy density of the Li-ion battery can be further increased.

2.5.2 Anode Chemistry for Li-Ion Battery

Besides the cathode materials, the anode materials are another key component in Li-ion batteries. The ideal anode material should have a low electrochemical potential vs lithium, a high-energy density, a good cycle life, an excellent safety performance, and a relatively low cost. Table 5 compares some of the important characteristics of different anode materials. The detail of each anode material will be discussed further below.

Graphite

Graphite is a crystalline form of the element carbon with its atoms arranged in a hexagonal structure. During the charge process in Li-ion batteries, lithium ions are transported from the cathode to the anode and then inserted into the space between two layers of the hexagonal structure. In theory, one lithium bonds to six carbons at fully charged state, so the theoretical capacity of graphite is 372 mAh/g. The average operation voltage of graphite vs Lithium is 0.2–0.5 V. When the Li-ion battery was

Table 5 Comparison of different anode materials for Li-ion batteries

Materials	Graphite	Si/SiOx	LTO	Metal alloy
Energy density	Good	High	Low	High
Power	High	Good	Good	High
Lifetime	Good	Low	High	Low
Safety	Good	Good	High	Low
Cost (material/manufacture)	Good	Good	Good	High
Industry readiness	High	Good (mix with graphite)	High	Low

first commercialized, nature graphite (from the mining process) was used as anode material due to its excellent electronic conductivity and low cost, although the energy density and cyclability of the graphite were not very good initially. During the first charge–discharge cycle which occurs in the battery manufacturing process, the difference between the charge capacity and discharge capacity of graphite (non-reversible capacity) is relatively high when graphite was initially introduced as anode material for Li-ion batteries. Extensive work has been done to improve the characteristics of this material. For example, artificial graphite, synthesized and further treated from natural graphite, has the properties tuned specifically to address the issue of nature graphite. In recent years, some companies are using artificial graphite blended with natural graphite to lower the cost and also get the benefits of both materials at the same time. Graphite is by far the most successful anode material used in Li-ion batteries.

Lithium Titanate ($Li_4Ti_5O_{12}$ or LTO)

Lithium titanate is a material with spinel structure, which is very stable during charge and discharge in lithium-ion batteries. It was developed initially as an alternative cathode material [18]. However, the operation potential of the LTO vs Li/Li^+ is about 1.2–1.5 V, which is much lower compared to the cathode material discussed earlier (>3 V). When using it as anode material, the high working voltage will definitely lower the energy density of the whole battery cell, but the crystal structure stability of the LTO is much higher compared to graphite, which makes it a strong candidate of the anode material for Li-ion battery packs with high-capacity, better safety performance and cycle life. The high working potential also allows the LTO anode to be used with lower cost aluminum foil as a current collector, which is also very beneficial for large-scale EV applications. Finally, the LTO showed better performance and stability at extremely high- or low-temperature conditions compared to graphite anode material. Due to these advantages, some of the EV buses have used the Li-ion battery with LTO anode.

Nonetheless, several issues for LTO need to be addressed before it can be widely used for EV applications:

- Relatively poor high-power performance. Due to the low electric conductivity of LTO, the discharge/charge performance with high current usage needs to be improved. Nano-sized LTO with conductive surface coating has been used in some of the batteries.
- Relatively lower energy density. As mentioned earlier, LTO has higher working potential and specific capacity (~ 150mAh/g) compared to other anode materials, which lowers the energy density of the whole cell when LTO is used. Matching LTO with new high-voltage cathode material can help. However, new electrolytes with more stable solvent and additives are needed to make the whole system working properly.

Silicon and Silicon Oxide (Si/SiO_x)

The combination of silicon (Si) and silicon oxide (SiO_x) has been investigated extensively in recent years due to its unique property as anode material for lithium batteries [19, 20]. Compared to LTO and graphite, Si forms an alloy with lithium during the charge process. Pure silicon provides supreme theoretical capacity for lithium (~ 4200 mAh/g), which is more than 10 times higher that of the graphite anode. Even the silicon oxide provides about 3 –4 times higher specific capacity compared to graphite. In addition, Si/SiOx has a lower operation potential vs. Li/Li^+ (~ 0.4 V), but still higher compared to graphite. Moreover, silicon is an abundant element source on earth, which can lower the raw material cost.

Similar to LTO, Si/SiO_x also faces some challenges before it can be widely used in EV applications:

- Very large volume change during the charging and discharging cycle in batteries. During lithiation, one Si atom can alloy with four Li atoms. The high-volume change results in significant structure strain and mechanical fracture at the particle level and also causes the loss of active components from the current collector at the electrode level, leading to poor cycle life performance. Using the nano-size silicon particles with hollow structures has been demonstrated as an effective method to accommodate the large volume change, but the manufacturing process cost of the specific nanostructure is too high for large-scale production.
- Unstable surface properties during cycle. When undergoing the huge particle volume change during the cycle, the surface protection layer on Si or SiO_x will result in repeated fracture and reformation, causing faster capacity fading rate and consumption of electrolyte in the battery. Work is ongoing to further stabilize the surface property with polymer-coating and specially designed binders in the electrode.
- Low electric conductivity of Si/SiOx due to the nature of the material. Lower conductivity will impact the high current charge and discharge performance of the batteries. Using nano-sized particles together with conductive material coating can improve the conductivity.

Although various strategies have successfully improved the electrochemical and mechanical performance of Si/SiO_x anodes, they still cannot fully replace graphite for EV applications. However, composite materials with Si/SiO_x mixed with graphite have been used by some EV makers in the market to achieve higher energy density, but the tradeoff is relatively short cycle life.

2.6 Next-Generation Li-Ion Battery for EV Applications

Although lithium batteries are used in EV applications now, there are still lots of challenges related to conventional chemistry. New battery chemistries with higher energy density and lower cost to meet the requirements of the next-generation EV applications are being developed. Some research directions discussed in various media and resources are described as follows.

High-Voltage Cathode

As discussed earlier, the conventional cathodes for EV show lower specific capacities compared to graphite, which is one limiting factor of the energy density of the whole battery. Developing new cathode materials with high average working potentials vs. Li/Li^+ (>4.0 V) is one direction to increase the energy density. High-voltage spinel cathode $LiNi_{0.5}Mn_{1.5}O_4$ and high nickel content NMC-811 are promising candidates in this category. However, there are always challenges to be met [21, 22]:

- Material stability. With increasing working potential, the thermodynamic stability of the material in the fully charged state is dramatically decreased. Less stable material will easily lead to material decomposition or thermal runaway, causing poor cycle life, even safety concern during regular usage. Structure doping and surface coating are possible solutions, but more systematic investigation and theoretical breakthrough are needed before their commercialization.
- New electrolyte materials and formulas need to be developed for new cathode materials. The solvent and lithium salt used in conventional electrolyte for EV applications have been optimized over the last decades in the potential range of below 4.5 V vs. Li/Li^+. Maintaining the stability of the solvent and lithium at high potentials is the key to ensure good cycle life and safety performance.
- Overall system improvement. With a higher voltage cathode, the voltage of the single battery will increase, which may need a new design for the whole battery pack to maintain the same pack properties.

Solid-State Electrolyte and Li metal Anode

As indicated in Fig. 7, conventional lithium-ion batteries use organic liquid electrolyte to provide ionic conductivity and also need polymer separators as electric insulators to prevent short between electrodes. In comparison, solid-state electrolyte (SSE) performs as both separator and electrolyte in solid-state batteries. SSE would be safer and could also possess a longer cycle life, a higher energy density, and impose less requirements on packaging and circuit-monitoring the state of charge [23].

Li metal anodes possess a very high theoretical specific capacity (3860 mAh/g) and the lowest electrochemical potential, both of which are very promising for future high-energy lithium-ion battery (LIB) applications. However, lithium dendrite formation during charging will seriously deteriorate the interfacial stability and further create thermal runaway if the separator is penetrated by the lithium dendrite [8]. With the non-porous SSE, using pure lithium metal as the anode becomes possible, because there is no risk of the lithium dendrite formation during the cycle to cause electric short between the two electrodes through the separator. The combination of high-energy Li metal anode and highly stable and safe SSEs would be a very promising solution to achieve high-energy density for automotive industry applications.

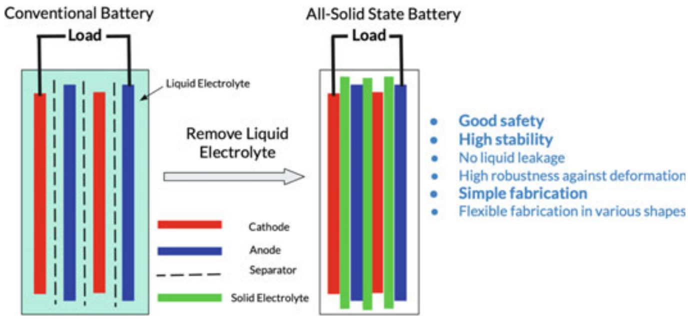


Fig. 7 Structure comparison between conventional battery and all-solid-state battery

Despite the advantages listed above, there are a few issues to be addressed before SSE can be commercialized with lithium anodes for EV applications:

- Issues related to SSE materials: large charge transfer resistance, low lithium-ion conductivity, large interfacial impedance. More research is needed to develop new materials to address those issues.
- Issues related to the manufacturing of solid-state batteries: relatively more complex/costly process and lower consistency. It is a challenge to incorporate the solid-state electrolyte in the battery and maintain the same production rate as that of manufacturing conventional batteries.
- Issues related to poor cycle life performances: there are large volume change and continuous surface morphology change of the lithium anode during the cycle, resulting in continuously increasing dead lithium volume. New interfacial layers between the lithium metal anode and SSE are needed to address such issues.

2.7 Battery Management System

In automotive applications, the power needed to drive the vehicle cannot be provided by single battery cells, battery pack consists of multiple battery cells to be used, and the number of battery cells could be from hundreds to thousands depending on the specific application. Designing a battery manage system (BMS) is one of the key parts of a successful battery pack design for any vehicle and requires a deep understanding of various aspects of battery cells, such as chemistries, performance behaviors, and failure modes. A successful BMS design also enables interaction between hardware and software to provide a reliable control under operation conditions [24].

The following are the three main objectives of a BMS design:

- To ensure the function of the battery could meet the specific application requirement under all conditions.
- To prevent damage on the cells or packs during normal or abnormal usage conditions.
- To extend the battery life (usage time per charge and total usable time).

One or more of the following functions summarized in Table 6 may be incorporated in a special BMS:

Table 6 Summary of different BMS functions and the impact on battery pack control

Functions	Control items and impact to the pack
Cell protection	Avoid operation conditions out of the specified range: • Voltage • Current • Temperature • Abnormal use (e.g., short circuit)
Charge control	Control the charge process per spec and requirement: • Control charge voltage and current at given state of charge (SOC) and temperature • Control voltage and current in given SOC and temperature • Control charge termination in given condition
Demand management	Minimize the current drain with power saving algorithm to extend the driving range between charges
SOC determination	Determine the SOC of the battery pack or the individual cells in the battery for user and charge control
State of health (SOH) determination	Determine the SOH of the battery pack or the individual cell in the battery pack. Provide warning to the customer or manufacturer when abnormality is found or maintenance is required
Cell balancing	Detect the imbalance between individual cells in the battery pack, compensate for weaker cells by equalizing the charge in all the cells in the chain to extend battery life
History record	Monitor and store the battery's history during usage to evaluate the usage condition and estimate SOH
Authentication and identification	Record information about the cell and battery packs: Cell manufacturer, chemistry, batch, pack production date, serial number, etc
Communications	Communicate between battery and charger or test equipment. Allow access to modify BMS control parameters or diagnosis

2.8 Battery Testing Methods and Industrial Standards

For batteries made for any application, multiple tests have to be performed to ensure the electric performance, safety, reliability of the battery can meet the industrial standards (as shown in Table 7) based on the battery application as the minimum requirement.

Beyond the industrial standard, during product development stage, the device or vehicle manufactures usually perform more extensive tests on both battery cells and packs levels for multiple purposes, for example, to characterize the battery performance at different temperature and charge/discharge conditions; to test the battery safety performance at different abuse use conditions to ensure the battery will not cause any safety event even under extreme usage conditions; to run tests to simulate the battery reliability performance after long-term usage in the field. Table 8 shows some main categories with some tests example and purpose related to the tests.

Another important category of battery testing is the tests performed at battery cell and pack manufactures after the mass production starts right after development stage done, extensive ongoing reliability tests (ORTs) are kept being performed to ensure the battery cells and packs produced during mass production can still meet the same requirement as that was set during the product development stage. The ORT sampling size and requirement may vary based on the device manufacture requirement and scale of the mass production. The ORT tests usually are the subset of the tests which were performed during the development stage. And the sample size of each test is usually based on the production volume together with guideline set by standard from quality control organization, for example, ANSI/ASQ Quality Standards Z1.4& Z1.9 can be used as a good reference.

Table 7 Summary of major industrial standard for Li-ion batteries

Standard	Standard name	Application
UN38.3	Certification for safe air transportation	All
UL1642	Standard for Lithium batteries	All
SAE J2464	Secondary lithium-ion cells for the propulsion of electrical road vehicles—part 2: reliability and abuse testing	EV
IEEE 62,660	Secondary lithium-ion cells for the propulsion of electrical road vehicles—part 2: reliability and abuse testing	EV
UL2580	Batteries for use in electric vehicles	EV
UL2271	Batteries for use in light electric vehicle (LEV) applications	EV
IEEE 1725	Standard for rechargeable batteries for mobile phones	Consumer
IEEE 1625	Standard for rechargeable batteries for multi-cell mobile computing devices	Consumer
UL 2054	Household and commercial batteries	Household

Table 8 Examples of battery tests during development stage

Test category	Example tests
Performance tests	Cycle performance-capacity fading rate
	Cycle performance-charge/discharge rate fading
	SOC vs OCV at various temperature
	Charge rate performance at various temperature
	Discharge rate performance at various temperature
	Battery DC/AC Imp at different temperature
Transportation and safety tests	Altitude simulation
	Thermal tests
	Shock tests
	Vibration tests
	External short circuit
	Overcharge (cell)
	Overdischarge (cell)
Abuse and reliability tests	Crush tests
	Impact tests
	Nail penetration tests
	Abnormal charge (pack)
	Force charge and discharge (pack with multi-cells)
	High-temperature/high humidity storage
	High-voltage long-term storage
	Temperature cycles

2.9 Battery Failure Mode and Effects Analysis (FMEA)

Begun in the 1940s by the US military, failure modes and effects analysis (FMEA) is a step-by-step approach for identifying all possible failures in a design, a manufacturing or assembly process, or a product or service. It is a common ***process analysis tool*** [25].

As Li-ion batteries are used in wide range of application, there are reports related to battery safety event or failure in almost every year from last decades. The reports covered all the application for Li-ion batteries, from Samsung Note7 battery safety events report [26] for consumer electronic application, to Telsa battery batteries caught fires during charge [27] and GM recalled Chevrolet Bolt and Bolt EV due to fire risks [28]. All those events make the understanding of the battery failure mode, especially the ones can cause safety events, a very important area for all batteries and devices manufactures.

Very sophisticated FMEA can be done based on specific chemistry used in Li-ion batteries, manufacturing process used to make the battery cells, and various components used in battery packs, and specified application conditions in the devices.

Table 9 shows the battery FMEA reported by Christopher Hendricks which covers some key factors which can introduce potential failure from manufacturing process, aging effect, abnormal or abuse use conditions.

3 Fuel Cells

Unlike battery, which is an energy storage device, fuel cell is energy conversion devices. The chemical energy stored in the fuel is converted directly into electrical energy through electrochemical reactions that happened inside the fuel cells. The electrochemical conversion process has higher efficiencies and lower overall emission compared to traditional combustion engine. For some types of fuel cells, zero emission is achieved when pure hydrogen and oxygen (air) are used as fuels, since the only reaction product is water [30, 31].

A fuel cell usually consists of three main parts: cathode (positive electrode), anode (negative electrode), and ionic conducting electrolyte, which is very similar to other electrochemical devices. During the process to generate electric energy, the fuel is oxidized in the cathode side, and the oxidant is reduced at the anode side, which generates the ion flow through the electrolyte within the fuel cell. As the result, electrons move through the external circuit to keep the static balance on both side of the electrodes. Since there is no energy is stored within the fuel cell, it can continuously provide energy as long as the fuel and oxidant are available.

3.1 Major Types of Fuel Cells

Five major types of fuel cells (based on ionic electrolyte types) have been widely investigated (listed in Table 10): proto-exchange membrane as electrolyte (PEMFC/DMFC), alkaline-based electrolyte (AFC), phosphoric acid electrolyte (PAFC), molten carbonate electrolyte (MCFC), and solid oxide electrolyte (SOFC). Same electrochemical principle is shared by all different kinds of fuel cells; however, the working temperature, mechanical structure, material used, fuel tolerance, and performance vary depending on properties of different electrolytes used in each type of the fuel cell. Table 10 provides the summary of major properties of those five major fuel cells mentioned earlier [32].

PAFC and AFC were investigated extensively and were used as first two types of the fuel cells; however, the advantages of other types of fuel cells listed in the table attract more and more attentions in later years by both industry and academic. For example, SOFC and MCFC show high operation efficiency and good fuel flexibility due to the high operation temperature, and the generated operation heat can also be used for power production by other method. Hydrogen can be used as fuel for

Table 9 Battery FMEA table reported by Christopher Hendrick et al. [29]

Battery component	Potential failure mode(s)	Potential failure mechanism(s)	Mechanism type	Observed effect	Potential failure causes	Likelihood of occurrence	Severity of occurrence	Ease of detection
Anode (active material)	Thickening of solid electrolyte interphase layer	Chemical reduction reaction and deposition	Wearout	Increased charge transfer resistance, reduction of capacity, reduction of power	Chemical side reactions between lithium, electrode, and solvent	High	Low	High
	Particle fracture	Mechanical stress	Overstress	Reduction of capacity, reduction of power	Intercalation stress	Moderate	Low	Low
	Reduced electrode porosity	Mechanical degradation	Wearout	Increased diffusion resistance, reduction of capacity, reduction of power	Dimensional changes in electrode	Moderate	Low	Low
	Lithium plating and dendrite growth on anode surface	Chemical reaction	Wearout	Can cause short circuit if dendrites puncture separator	Charging the battery at low temperatures or high rates	Low	High	Low

(continued)

Table 9 (continued)

Battery component	Potential failure mode(s)	Potential failure mechanism(s)	Mechanism type	Observed effect	Potential failure causes	Likelihood of occurrence	Severity of occurrence	Ease of detection
Anode (current collector)	Free copper particles or copper plating	Chemical corrosion reaction and dissolution	Wearout	Increased resistance, reduction of power, reduction of current density	Overdischarge of the battery	Low	High	Low
Cathode (active material)	Thickening of solid electrolyte interphase layer	Chemical reduction reaction and deposition	Wearout	Increased charge transfer resistance, reduction of capacity, reduction of power	Chemical side reactions between lithium, electrode, and solvent	High	Low	High
	Particle fracture	Mechanical stress	Overstress	Reduction of capacity, reduction of power	Intercalation stress	Moderate	Low	Low
	Reduced electrode porosity	Mechanical degradation	Wearout	Increased diffusion resistance, reduction of capacity, reduction of power	Dimensional changes in electrode	Moderate	Low	Low

(continued)

Table 9 (continued)

Battery component	Potential failure mode(s)	Potential failure mechanism(s)	Mechanism type	Observed effect	Potential failure causes	Likelihood of occurrence	Severity of occurrence	Ease of detection
	Gas generation and bloating of cell casing	Thermally driven electrode decomposition	Overstress	Reduction of capacity	Overcharge of the battery or short circuit	Low	High	Low
Cathode (current collector)	Pitting corrosion of aluminum	Chemical corrosion reaction	Wearout	Increased resistance, reduction of power, reduction of current density	Overcharge of the battery	Low	Moderate	Low
Separator	Hole in separator	Mechanical damage	Overstress	High heat generation due to joule heating, bloating of cell casing, drastic voltage reduction	Dendrite formation, external crushing of cell	Low	High	Moderate
	Closing of separator pores	Thermally induced melting of separator	Overstress	Inability to charge or discharge battery	High internal cell temperature	Low	High	High
Lithium ions	Reduction in lithium ions, thickening of solid electrolyte interphase layer	Electrolyte reduction and solid product formation	Wearout	Reduction of capacity	Chemical side reactions between lithium, electrodes, and solven	High	Low	High

(continued)

Table 9 (continued)

Battery component	Potential failure mode(s)	Potential failure mechanism(s)	Mechanism type	Observed effect	Potential failure causes	Likelihood of occurrence	Severity of occurrence	Ease of detection
Electrolyte salt	Decrease in lithium salt concentration	Chemical reduction reaction and deposition	Wearout	Increased diffusion resistance	Chemical side reactions between lithium, electrodes, and solvent	High	Low	High
Organic solvents	Gas generation and bloating of cell casing	Chemical decomposition of solvent	Overstress	Increased diffusion resistance, and may lead to thermal runaway	High external temperature, overcharging of the cell	Low	High	Low
	Thickening of solid electrolyte interphase layer	Chemical reduction reaction and deposition	Wearout	Increased charge transfer resistance, reduction of capacity, reduction of power	Chemical side reactions between lithium, electrodes, and solvent	High	Low	High
Terminals	External corrosive path between positive and negative leads	Chemical corrosion reaction	Wearout	High heat generation due to joule heating, bloating of cell casing, drastic voltage reduction	Inadvertent shorting of the terminals	Low	High	Moderate
	Solder cracking	Thermal fatigue Mechanical vibration fatigue	Wearout	Loss of conductivity between battery and host device	Circuit disconnect	Low	Moderate	High

(continued)

Table 9 (continued)

Battery component	Potential failure mode(s)	Potential failure mechanism(s)	Mechanism type	Observed effect	Potential failure causes	Likelihood of occurrence	Severity of occurrence	Ease of detection
Casing	Internal short circuit between anode and cathode	Mechanical stress	Overstress	High heat generation due to joule heating, bloating of cell casing, drastic voltage reduction	External load on cell	Low	High	Moderate

Table 10 Major properties summary of five types of fuel cells

	PEMFC&DMFC	AFC	PAFC	MCFC	SOFC
Electrolyte	Proton exchange membrane	Potassium hydroxide in asbestos matrix	Phosphoric acid in SiC	Molten carbonate in LiAlO2	Ceramic oxide ion conductor
Electrode	Carbon	Transition metals	Carbon	Nickel and Nickel oxide	Oxide, oxide/metal cermet
Operating temperature	50–130 °C	50–250 °C	180–200 °C	650 °C	600–1000 °C
Charge carrier	H+	OH^-	H +	CO_3^{2-}	O_{2-}
Catalyst	Pt/PtRu	Pt	Pt	Electrode material	Electrode material
Fuel	Pure or reformed H_2/CH_3OH	H_2 and CH_3OH	Reformed H_2	Reformed H_2 and CH_4	Reformed H_2 and hydrocarbons
CO tolerance	Poison (<50 ppm)	Poison (<50 ppm)	Poison (<1%)	Fuel	Fuel
Electrical efficiency (%)	40–50	50	40	45–55	50–60
Power density (mW/cm^2)	300–1000	150–400	150–300	100–300	250–350
Internal reforming	No	No	No	Yes	Yes
Power range (kW)	0.001–1000	1–100	50–1000	100–100,000	10–100,000
Major applications	Portable, transportation, stationary	Space, stationary, transportation	Stationary	Stationary, transportation	Stationary, transportation
Balance of plant	Low-moderate	Moderate	Moderate	Complex	Moderate

types of fuel cells, but the higher operation temperature provides higher tolerance for impurity and provides possibility to internal reforming hydrocarbon fuels to generate hydrogen, all those advantages make them good candidates for stationary power supplies with less emission. In comparison, PEMFC/DMFC uses proton-exchange membrane as electrolyte, which runs at relatively lower temperature and higher overall energy density, which make them more attractive for portable and transportation applications.

Fig. 8 Summary of electrochemical reactions in PEMFC

Anode reaction: $H_2 \rightarrow 2H^+ + 2e^-$ $E^\circ = 0.00V$ (1.1)

Cathode reaction: $\frac{1}{2}O_2 + 2H^+ + 2e^- \rightarrow H_2O$ $E^\circ = 1.23V$ (1.2)

Overall reaction: $\frac{1}{2}O_2 + H_2 \rightarrow H_2O$ $E_{cell}^\circ = 1.23V$ (1.3)

3.2 Fuel Cells for EV Application

Proto exchange membrane fuel cell (PEMFC) is the only type of fuel cell that is currently used for electric vehicle applications due to its high-power density, rapid start-up, quick response to power demands, and zero emission during operation. Pure hydrogen and oxygen/air are used as fuel and oxidant, respectively; the electrochemical reactions involved are given in Fig. 8.

In PEMFC, pure hydrogen is used as fuel and oxidized to proton and electrons on the surface of the catalyst in the anode electrode. The produced protons flow through electrolyte to the cathode side and get oxidized by the oxygen (air) on the surface of catalyst in the cathode electrode to produce water. The overall potential of a single PEMFC cell is 1.23 V in the standard state.

Figure 9 is a schematic view of inside typical single PEMFC which usually consists of seven different layers. The bipolar plates are the paths to supply the fuel and oxidant, also the terminals connected to external circuits. The membrane electrode assembly (MEA) is the heart of the fuel cell, which includes the gas diffusion layer (GDL) and catalyzed membrane in the center. The GDL is prepared by depositing a mixture of carbon powder and binder (PTFE) onto a carbon paper or cloth and works together with bipolar plates for same function. The catalyzed membrane consists of a proto exchange membrane and two layers of electro-catalyst on each side of the membrane. Nafion (product of Dupot) membrane is the most widely used membrane for PEMFC and DMFC due to its high proton conductivity and good chemical and thermal stability. The catalyst coated on each side of the membrane is a mixture of carbon, Nafion solution, and nanoparticle metal catalyst. The detailed composition and loading (unit weight) of the catalyst vary depending on the type of fuel cell and specific application, most commonly used catalyst is the nano-sized Pt deposited on nanostructure carbon materials.

Different auto-vehicle companies have been working in this area since 1990s like GM, Ford, Honda, Toyota. The most successful example so far is the Toyota Mirai, which was unveiled in 2014 and on-sale in 2015. Mirai uses proton-exchange membrane fuel cells (with hydrogen as fuel) as the main power source and hybrids with built-in battery pack to improve the overall efficiency. The operating range of Miara could reach over 300 miles with a full tank of hydrogen [33]. There are more companies trying to commercialize PEMFC-powered vehicles in transportation applications.

Compared to battery-powered electric vehicles, the PEMFC has relatively higher cost due to the high cost of the catalyst used in the electrode. However, there are also advantages listed as follows:

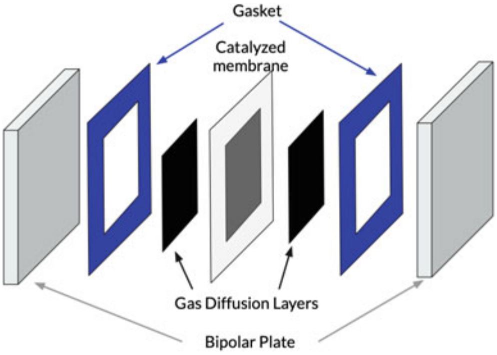


Fig. 9 Schematic of a proton-exchange membrane fuel with seven-layer membrane electrode assembly (MEA)

- Fast refuel. The hydrogen is used as fuel for PEMFC, refuel time needed is comparable to regular gasoline.
- Very high-energy density, hydrogen has the highest gravimetric energy density among all the fuels, and the overall operation efficiency of the fuel cell can be as high as 70%.
- Zero operation pollution: during the operation, hydrogen reacts with oxygen to form water as the only product, no harmful by-product is generated.
- Same power advantage as battery-powered car, with fast response at low speed, and no lags.

There are also challenges related to PEMPC before it can be further commercialized by more automotive manufacturers.

- High material costs for electrode and proton-exchange membranes. Pt-metal is widely used as a catalyst in the electrode fuel cell, and no alternative metal has shown similar reactive activity compared to Pt. Proton-exchange membrane, which consists mainly of carbon fluorine-based polymer, is also a relatively expensive material due to high manufacturing cost and limited demand for related applications in the industry.
- High-pressure storage devices are needed to store hydrogen in vehicles and fuel stations. The manufacturing and maintenance costs for high-pressure storage devices are still too high for a wide deployment of the hydrogen-related technology.
- New hydrogen fueling stations are needed. The infrastructure is not ready. More hydrogen fuel stations are to be built.

4 Capacitors

Similar to batteries, capacitors are energy storage devices with a long history. The working mechanism for capacitors is listed in Fig. 10 [34].

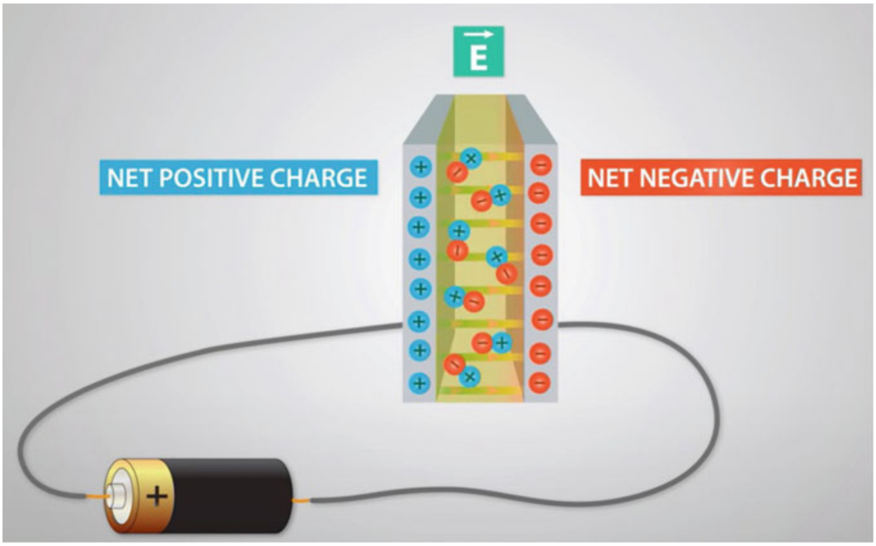


Fig. 10 Working mechanism of capacitors

As shown in the chart, a conventional capacitor consists of two electrodes which are separated by aqueous electrolyte and a separator. During the charge process, driven by the potential, positive and negative charges accumulate at the interface between each electrode and the electrolyte. The accumulated charges will be released when there is load connected to the two electrodes through an external loop. However, different from charges being created by electrochemical reactions during the charge and discharge process in the battery, the charges in the capacitor are physically stored on the surfaces of both electrodes. As a result, the capacitor can provide very high peak power in a relatively short time, but the overall energy density is much lower compared to the battery and the fuel cell.

The porous materials in electrodes need to have very high specific surface area, superb chemical stability, and high electric conductivity, low cost. Carbon materials have been widely used as electrode materials in capacitor production. Over years, there is intensive development of carbon materials with extremely high specific surface area. By using the highly porous carbon material as electrode material, the energy density of the capacitor has increased dramatically. There are also companies around the world trying to commercialize the capacitor with much higher energy density by using newly developed carbon-based materials, like carbon nanotube and graphene.

Another approach to increase the energy density for supercapacitors is using non-aqueous electrolytes. Because of the electrochemical stability of the aqueous solution, the operating voltage of a traditional capacitor has to be below 1.5 V. With a non-aqueous electrolyte such as an organic solvent with lithium salt, the operating voltage of the new capacitor can be more than 3.0 V. The development provides an opportunity to double the energy density compared to conventional aqueous electrolytes.

Because of its low energy density, capacitors can hardly be the single power source for EV applications. However, the supreme high-power performance makes them a strong candidate to be used with fuel cells or batteries as a backup power source to provide extremely high discharge power in a very short period of time. Moreover, there is a report about a new form of electric bus, known as Capabus, which runs without continuous overhead lines (which makes it an autonomous vehicle) by using power stored in large onboard electric double-layer capacitors (EDLCs). The onboard supercapacitor packs are quickly recharged whenever the vehicle stops at any bus stop (under so-called electric umbrellas) and fully charged at the terminus [35].

5 Summary

In this chapter, Li-ion cells and packs, the most popular power sources, are reviewed in terms of the types and structures of the cells, the manufacturing process, the state-of-the-art Li-ion chemistries for EV applications, the development trends, and the battery management systems for battery packs. The industrial battery test standard together with potential failure mode for Li-ion battery was also briefly discussed. The properties, costs, and current status of fuel cells and capacitors for EV applications are also discussed.

There are few key points based on our discussion:

These three electrochemical devices have different properties and characteristics with their own advantages and disadvantages for various usage cases in EV applications. Some key points:

- Batteries are energy storage devices. Li-ion batteries show the highest energy density along with good cycle life, acceptable cost, and reasonable high rate discharge capability. All those properties make it the first choice as the power source for EV.
- Extensive studies are ongoing to further improve the overall performance of Li-ion batteries. Especially, the combination of Li metal anode and highly stable and safe SSEs would be a very promising solution to further improve the energy density of batteries.
- Capacitors are also energy storage devices with excellent high rate charge and discharge capability, but the energy density is much lower compared to batteries. They can be used to boost the high rate discharge capability when combined with Li-ion batteries in EV application.
- Fuel cells are energy conversion devices which convert the chemical energy stored in the fuels to electric energy. The overall energy density for fuel cell is comparable to internal combustion energy system. However, high cost of the fuel cell and less availability of fuel station prevents the usage of fuel cells at current stage.

The future of electrochemical devices for EV applications depends not only on the continuous investigation in electrochemical theories, materials, and manufacturing processes but also on the further commercialization of the new technologies with lower costs, higher energy densities, and better safety.

Acknowledgements Authors would like to thank Wentao Qin from Onsemi and Hualiang Shi from Apple for their critical review of this chapter.

References

1. https://industryeurope.com/sectors/transportation/volkswagen-to-go-all-electric-by-2026.
2. Mathis, T.S., Kurra, N., et al. Energy Storage Data Reporting in Perspective—Guidelines for Interpreting the Performance of Electrochemical Energy Storage Systems, Advanced Energy Materials, 9 (39), 1902007- (2019).
3. https://www.epectec.com/batteries/cell-comparison.html.
4. https://www.sciencedirect.com/science/article/pii/B9780128035818107647.
5. P. Roy and S. K. Srivastava, Nanostructured anode materials for lithium ion batteries, Journal of Materials Chemistry A, 2015,3, 2454–2484.
6. https://www.tmaxcn.com/automatic-pouch-cell-production-equipment-line-for-li-ion-battery-making_p1554.html.
7. https://en.wikipedia.org/wiki/John_B._Goodenough#cite_note-RSC1-1.
8. https://en.wikipedia.org/wiki/Lithium_cobalt_oxide.
9. Ding, Y.L, Cano, Z.P, et al. Automotive Li-Ion Batteries: Current Status and Future Perspectives, Electrochemical Energy Reviews, 2, 1–28 (2019).
10. Padhi, A.K., Nanjundaswamy, K.S., Goodenough, J.B., Phospho-olivines as positive-electrode materials for rechargeable lithium batteries. Journal of The Electrochemical Society **144**(4), 1188–1194 (1997).
11. Thackeray, M., David, et.al, Lithium insertion into manganese spinels. Materials Research Bulletin **18**(4), 461–472 (1983).
12. Yan, W, Huang, Y, et.al, A review on doping/coating of nickel-rich cathode materials for lithium-ion batteries, Jour of Alloys and Compounds, 819, 15308(2008).
13. Ohzuku, T., Ueda, A., et.al, Comparative study of LiCoO2, $LiNi_{1/2}Co_{1/2}O_2$ and LiNiO2 for 4 volt secondary lithium cells. Electrochimica Acta **38**(9), 1159–1167 (1993). doi: https://doi.org/https://doi.org/10.1016/0013-4686(93)80046-3.
14. Choi, J., Manthiram, A.: Role of Chemical and Structural Stabilities on the Electrochemical Properties of Layered $LiNi_{1/3}Mn_{1/3}Co_{1/3}O_2$ Cathodes. Journal of The Electrochemical Society **152**(9), A1714–A1718 (2005). doi:https://doi.org/10.1149/1.1954927.
15. An, S.J., Li, J, et.al, Correlation of Electrolyte Volume and Electrochemical Performance in Lithium-Ion Pouch Cells with Graphite Anodes and NMC532 Cathodes. Journal of The Electrochemical Society **164**(6), A1195–A1202 (2017). doi: https://doi.org/10.1149/2.1131706je.
16. Myung, S.-T., Maglia, F., et.al, Nickel-rich layered cathode materials for automotive lithium-ion batteries: achievements and perspectives. ACS Energy Letters **2**(1), 196–223 (2016).
17. A. L. Lipson, J. L. Durham, et.al, Improving the Thermal Stability of NMC 622 Li-Ion Battery Cathodes through Doping During Coprecipitation, ACS Applied Materials & Interfaces 2020 12 (16), 18512–18518, DOI: https://doi.org/10.1021/acsami.0c01448.
18. Wang, Y.-Q., Gu, L., et.al., Rutile-TiO_2 nanocoating for a high-rate $Li_4Ti_5O_{12}$ anode of a lithium-ion battery. Journal of the American Chemical Society **134**(18), 7874–7879 (2012).
19. Cui, L.-F., Yang, et.al, Carbon− silicon core− shell nanowires as high capacity electrode for lithium ion batteries. Nano Letters **9**(9), 3370–3374 (2009).

20. J. Liu, **W. Li,** and A. Manthiram, “Dense Core-shell Structured SnO_2/C Composites as High Performance Anodes for Lithium Ion Batteries”, Chemistry Communication **46**, 1437 (2010).
21. Kim, J.H., Myung, S.T., et.al, Comparative Study of $LiNi_{0.5}$Mn1.5O4-δ and LiNi0.5Mn1.5O4 Cathodes Having Two Crystallographic Structures: $Fd\bar{3}m$ and P4332. Chemistry of Materials **16**(5), 906–914 (2004). doi:https://doi.org/10.1021/cm035050s.
22. Lu, D., Xu, M., et.al., Failure Mechanism of Graphite/$LiNi_{0.5}Mn_{1.5}O_4$ Cells at High Voltage and Elevated Temperature. Journal of The Electrochemical Society **160**(5), A3138–A3143 (2013). doi:https://doi.org/10.1149/2.022305jes.
23. Janek, J., Zeier, W.G.: A solid future for battery development. Nature Energy **1**, 16141 (2016). doi:https://doi.org/10.1038/nenergy.2016.141.
24. https://www.mpoweruk.com/bms.htm.
25. https://asq.org/quality-resources/fmea.
26. https://pages.samsung.com/us/note7/recall/index.jsp.
27. https://www.washingtonpost.com/technology/2021/08/04/tesla-fire/.
28. https://www.consumerreports.org/car-recalls-defects/chevrolet-bolt-recalled-again-due-to-fire-concerns-a3566085147/.
29. C. Hendricks, N. Williard, et.al, A failure modes, mechanisms, and effects analysis (FMMEA) of lithium-ion batteries, Journal of Power Sources, 297,113(2015).
30. S. Srinivasan, R. Mosdale, et.al, Fuel Cells: Reaching the Era of Clean and Efficient Power Generation in the Twenty-First Century, Annu. Rev. Energ. Env. 24 (1999) 281.
31. B. D. McNicol, D. A. J. Rand, K. R. Williams, Fuel cells for road transportation purposes — yes or no, J. Power Sources 100 (2001) 47.
32. W. Li, Development and Understanding of New Membranes Based on Aromatic Polymers and Heterocycles for Fuel Cells. Doctor dissertation.
33. https://en.wikipedia.org/wiki/Toyota_Mirai.
34. https://howtomechatronics.com/how-it-works/electronics/what-is-capacitor-and-how-it-works.
35. https://en.wikipedia.org/wiki/Capa_vehicle.

In-Vehicle Display Technology

Fen Chen and Jim Kuo

Abstract Visualization technologies are the most vital components of in-vehicle interactions. The shift toward autonomous vehicles and connected cars is bringing a future in which occupants would be needed to monitor the status of the vehicle and its surroundings. Meanwhile, occupants would also spend significantly more time watching displays for entertainment, information, and connectivity on the road. Therefore, the need for in-vehicle displays with better visibility, brightness, viewing angle, resolution, sharpness, and reliability together with larger size and free-form that offer unobtrusive visual information during journeys is on the rise. Superior display with touch technologies can enable a safe, informative, and comfortable driving or riding experience. The applied in-vehicle display products include center infotainment display, rear-seat entertainment display, head-up display, side mirror display, and instrument cluster display. In this chapter, motivations, as well as various architectures of display including LED, LCD, OLED, mini-/micro-LED, TFT, flexible, head-up display, and touch screen, will be introduced. Designing displays into vehicles imposes very different challenges than designing them for consumer applications. This is due to some unique factors associated with vehicle usages, such as the required product life cycles, the extremely harsh environment, frequent mechanical impacts, the stringent EMI/EMC compliance, the required high-level ESD protection, and functional safety requirements. Requirements and challenges of display in-vehicle application, including fabrication, characterization, inspection, quality, reliability, EMI/EMC/ESD, and failure analysis are reviewed.

F. Chen (✉)
Cruise, San Francisco, CA, USA
e-mail: fen.chen@getcruise.com

J. Kuo
Pegatron, Taipei, Taiwan
e-mail: jim_kuo@pegatroncorp.com

Y. Li and H. Shi (eds.), *Advanced Driver Assistance Systems and Autonomous Vehicles*,
https://doi.org/10.1007/978-981-19-5053-7_11

1 Introduction

Not too many years ago, there were hardly any electronic displays in vehicles. In today's vehicle, any new intermediate or luxury models are equipped with multiple displays. And as we will see, the usage of the display is about to expand greatly. Visualization technologies are the most vital components of in-vehicle interactions. The shift toward autonomous vehicles and connected cars is bringing a future in which occupants would be needed to monitor the status of the vehicle and its surroundings. On the other hand, levels 4 and 5 of autonomous driving also provide a lot of free time for occupants. Consequently, occupants will spend significantly more time watching displays than today as vehicles will evolve to mobile in-car offices and gaming suites, which can be used for work as well as leisure while on the roads. Therefore, the need for in-vehicle displays with better visibility, brightness, viewing angle, resolution, sharpness, and reliability together with larger size and free-form that offer unobtrusive visual information during journeys is on the rise. Superior display with touch technologies can enable a safe, informative, and comfortable driving and riding experience.

The applied in-vehicle display products include center infotainment display, rear-seat entertainment display, head-up display, side mirror display, and instrument cluster display. Figure 1 illustrates the futuristic in-vehicle display. Designing displays into vehicles imposes very different challenges than designing them for consumer applications. This is due to many unique factors associated with vehicle usage, such as the required product life cycles, the extremely harsh environment, frequent mechanical impacts, the stringent EMI/EMC compliances, the robust ESD protection, and safety requirements. This may be part of the reason to explain why it has taken so much longer time for the automotive industry to adopt state-of-the-art display technologies integrated into vehicles than the consumer product industry. The cost and development plus qualification cycles are much higher and longer for in-vehicle displays.

Liquid crystals were discovered in 1888 by the Austrian botanist Fredrich Reinitzer. A liquid–crystal display (LCD) is a flat-panel display that uses the light-modulating properties of liquid crystals combined with polarizers. It uses a backlight or reflector to produce images in color or monochrome. During the 1960s, RCA developed the first LCDs based on the dynamic scattering effect [1]. In 1970, the twisted nematic field effect in liquid crystal, on which many current LCDs are based, was the first filed for patent by Hoffmann-LaRoche in Switzerland (Swiss patent No. 532 261). Then LCD quickly dominated in small portable applications. A matrix of row and column electrodes was needed for displays with higher information content. This leads to the development of matrix addressing. To obtain the maximum contrast ratio, a switch was added at the intersection of every row and column to activate the pixel in the matrix. T. Peter Brody was the first to construct so-called active-matrix LCDs (AMLCDs) with thin-film transistors (TFTs) in 1973 [2]. The TFT approach has emerged as the most successful technique for creating an active matrix. In 1987, Sharp manufactured a 14-inch, active-matrix, full-color,

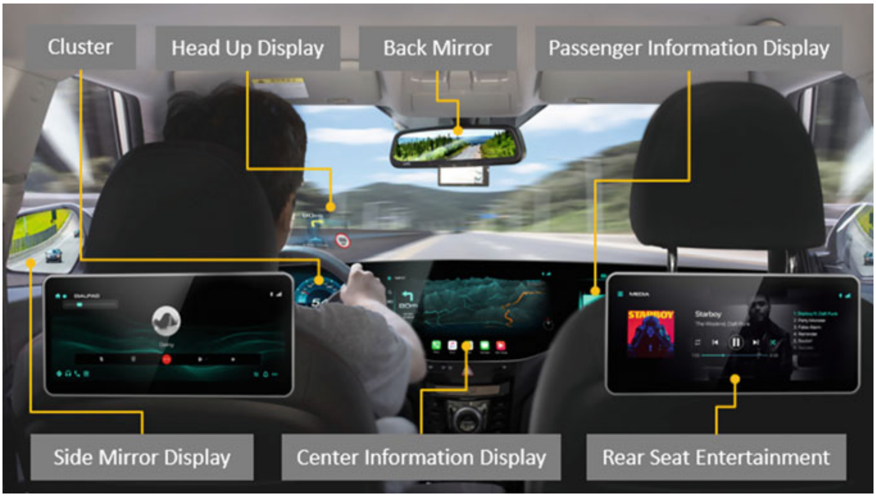


Fig. 1 Futuristic in-vehicle display products

full-motion TFT LCD, which led to the booming of the LCD industry. This industry developed large-size LCDs, including TFT computer monitors and larger LCD TVs in the 1990s and 2000s. By 2008, annual sales of televisions with LCD screens exceeded sales of CRT units worldwide, and the CRT became obsolete for most purposes. Since LCDs produce no light of their own, they require external light to produce a visible image. The light source is provided at the back of the glass stack and is called a backlight (https://en.wikipedia.org/wiki/Backlight). Active-matrix LCDs are almost always backlit. During the 1990s, the LED backlight has been merged as a mainstream backlight for LCD due to its high brightness and uniformity. Today, most LCD screens are being designed with an LED backlight consisting of edge-lit white LED LCD, white LED array, RGB LED array, quantum dot LED, and mini-/micro-LED array technologies. Due to the LCD layer that generates the desired high-resolution images using very low-power electronics in combination with LED-based backlight technologies, LCD technology has become the dominant display technology for vehicle displays nowadays.

André Bernanose and co-workers at the Nancy-Université (https://en.wikipedia.org/wiki/Nancy-Universit%C3%A9) in France made the first observations of electroluminescence (https://en.wikipedia.org/wiki/Electroluminescence) in organic materials in the early 1950s [3]. In 1987, Ching Wan Tang and Steven Van Slyke (https://en.wikipedia.org/wiki/Steven_Van_Slyke) at Eastman Kodak (https://en.wikipedia.org/wiki/Eastman_Kodak) built the first practical OLED device [4]. OLEDs work in a similar way that conventional LEDs do, but instead of using n-type and p-type semiconductor layers. They use organic molecule layers to generate electrons and holes. A simple OLED is made up of six different layers. On the top and bottom, there are layers of protective glass (https://www.explainthatstuff.com/glass.html) or plastic. The top layer is called the seal, and the bottom layer is called the substrate. In between those layers, there is a cathode and an anode. Finally, in between the anode and cathode are two layers made from organic molecules called the emissive layer and the conductive layer. Kodak released several of the earliest OLED-equipped products. In 1999, Kodak entered into a partnership to jointly research, develop, and produce OLED displays. They announced the world's first 2.4-inch active-matrix, full-color OLED display in September of the same year [5]. In September 2002, they presented a prototype of a 15-inch HDTV format display based on white OLEDs with color filters at the CEATEC Japan. For a high-resolution display, a TFT (https://en.wikipedia.org/wiki/Thin-film_transistor) backplane is necessary to drive the OLED display pixels correctly. As of 2019, low-temperature polycrystalline silicon (LTPS) TFT is widely used for commercial active-matrix OLED (AMOLED) displays. Meanwhile, amorphous silicon and IGZO backplanes have also been reported with large display prototype usages. Kodak has licensed its OLED technology to many companies to commercialize OLED display technology and lighting technology. Recently, OLED technology has also been adopted in the automotive display market. A Japanese manufacturer, Pioneer Electronic Corporation, (https://en.wikipedia.org/wiki/Pioneer_Electronic_Corporation) produced the first car stereos with a monochrome OLED display, which was also the world's first OLED product in vehicles [6]. Although competing OLED technology is pushed to the automotive market, the number of automakers using OLED displays is still rare and limited to the high-end of the market. The biggest technical challenge for OLED displays used in vehicles is the limited lifetime of the organic materials.

A touch screen or touch display is the assembly of both an input ("touch panel") and output ("display") device. The touch panel is normally layered on the top of a display of an information processing system (https://en.wikipedia.org/wiki/Information_processor). The first finger-driven touch display based on capacitive touch was invented by E.A. Johnson in 1965 at the Royal Radar Establishment in Malvern, United Kingdom. A capacitive touchscreen panel uses an insulator that is coated with a transparent conductor such as indium tin oxide (ITO) or silver wires. A human finger acts as the conductive part to make a fine electrical conductor. Touchscreens began being heavily commercialized during the 1980s. Hewlett-Packard HP-150 used a 9-inch Sony CRT surrounded by infrared emitters and detectors to send where the user's finger touches the screen. In 1984, BoB Boie of Bell Labs developed the first transparent multitouch screen overlay to bring multitouch technology a step forward. In 1993, IBM and BellSouth teamed up to launch the Simon Personal Communicator, one of the first cellphones with touchscreen technology. At the end of the decade, University of Delaware graduate student Wayne Westerman and his faculty advisor, John Elias, developed a multitouch projected capacitive or pro-cap technology, which has gone on to become a staple feature in modern touchscreen-equipped devices. They formed a company called FingerWorks to produce a line of multitouch gesture-based products. FingerWorks was eventually acquired by Apple in 2005, and many attribute technologies like the iPhone's touchscreen and the multitouch Trackpad to this acquisition. One emerging trend is that touchscreens are becoming increasingly interdependent on display technologies. The growing variety of interactive display technologies allows touchscreens to be widely used.

The rising demand for better touchscreens means users will increasingly expect more performance at a lower cost. Today's touchscreens should be thin, light, visible in varying ambient light conditions, highly responsive, and affordable. Quick response and transparent touchscreens are now critical to a great user experience. Such emerging needs can only be achieved through transparent conductors invisible to the naked eye. A new component to enable touchscreen technology to advance further is the silver nanowire due to its low conductivity, high transmission, superior pattern visibility with no moire issue, lightweight, thinner thickness, and flexibility. Silver nanowires are expected to become the new gold standard in touchscreen displays, delivering a bounty of benefits such as large-area touchscreen, and flexible display at a reduced cost. Automotive manufacturers are integrating touchscreens into center information displays, navigation systems, co-drive displays, cluster displays, and rear-seat entertainment systems. The monitor installed in a vehicle dashboard is usually used for navigation and information display purposes, and most are touch screens. Touchscreens began in the 1980s. Automakers long ago had the idea to transform displays into digital, interactive screens. But they took some time to catch on with the masses. In 1986, the Buick Riviera became the first production car to have a touch screen and infotainment system. Every Riviera came with a 9-inch touch screen, equipped with the graphic control center infotainment system. Over the years, the technology improved, and in-dash touchscreens seemed to grow bigger and bigger (https://www.motorbiscuit.com/the-2020-mazda-cx-30-is-missing-a-key-piece-of-technology/). What began with digital displays for car-stereo equalizers quickly transitioned to full map-view navigation and digital information display. According to The Globe and Mail (https://www.theglobeandmail.com/drive/culture/article-shift-from-huge-in-car-screens-may-be-under-way-but-first-theyll/#:~:text=In%201986%2C%20the%20Buick%20Riviera,to%20have%20a%20touch%20screen) [7], some observers credit the 2001 BMW 7 Series for standardizing the central, dash-mounted touchscreen as the hub for car interactions. Almost all of today's new vehicles offer in-vehicle display touchscreen technologies. The display is often a TFT LCD (https://en.wikipedia.org/wiki/Liquid-crystal_display) or OLED (https://en.wikipedia.org/wiki/OLED) display. Touch display screens are taking up more dashboard real estate and control more vehicle functions. As an example, Tesla has revolutionized its touchscreen capabilities (https://www.motorbiscuit.com/why-is-mazda-removing-all-touchscreens-from-future-vehicles/). Given the Model S's 17-inch screen that controls everything from vehicle speed to the sunroof, it is clear that touchscreens and other digital controls will continue to evolve.

2 In-Vehicle Display Technologies and Architectures

With the rapid development of automotive technologies plus the more and more time users spend on the roads, the demand for cars to provide a superior riding experience has never become higher. At present, cars are no longer just a safe means of transportation. What users expect is a more comfortable, smarter, and less energy consumed on-road experience. These demands are inextricably related to the in-vehicle displays. Currently, in-vehicle display applications and technologies are listed as follows:

a. Center infotainment display (CID) with GPS: The current mainstream technology is TFT LCD and a small number of cars use OLED
b. Dashboard cluster or digital cluster (DC): The current mainstream technology is TFT LCD, and a small number uses OLED
c. Air-conditioning control: The current mainstream technology is TFT LCD, and a small number use OLED
d. Head-up display: The current mainstream technology is TFT LCD, and a small number use a laser diode.

Regarding currently used in-vehicle display technologies, LCD-based display dominates as it is a well-established, robust, reliable, and cost-effective technology. However, the mini-/micro-LED display is rapidly gaining attention in the automotive market due to its greater accuracy of backlighting, greater peak brightness, and higher reliability than OLED. In the short term, although it cannot change the leading position of TFT LCD for in-vehicle display usage, once the supply chain is complete, and the yield rate is rapidly increasing, its potential first to take over OLED, then TFT LCD as the mainstream in-vehicle display technology cannot be underestimated. Overall, the diversified technologies of in-vehicle displays and the demand for larger and more displays inside future vehicles will inevitably enrich the evolution and development of in-vehicle display technologies.

2.1 LCD

The early passive LCD has a wide variety of applied technologies. Per LCD type, it can be divided into twist nematic (TN), high-twisted nematic (HTN), enhanced twist nematic (ETN), super twist nematic (STN), film super twist nematic (FSTN), temperature compensation film super twist nematic (TCFSTN), double film super twist nematic (DFSTN), and others. The passive LCD can be built with different backlights based on the different background colors of the display to enhance the contrast ratio. Per electrical connection type, it can be divided into pin type, conductive rubber type, heat seal type, and FPC type. Per display driving IC position, it can be divided into chip-on-glass (COG), chip-on-film (COF), chip-on-board (COB), etc. Figure 2 shows the various passive LCD for automotive applications.

During the early stage of LCD development, passive LCD has the advantages of low-development cost and high reliability. Later, due to the continuous decline in the cost of thin film transistor (TFT) LCD and the rapid development of wide viewing angle TFT technologies such as in-plane switching (IPS) and multi-domain vertical alignment (MVA), passive LCD is gradually being replaced by TFT LCD for automotive display applications. Currently, even the entry-level vehicle has been

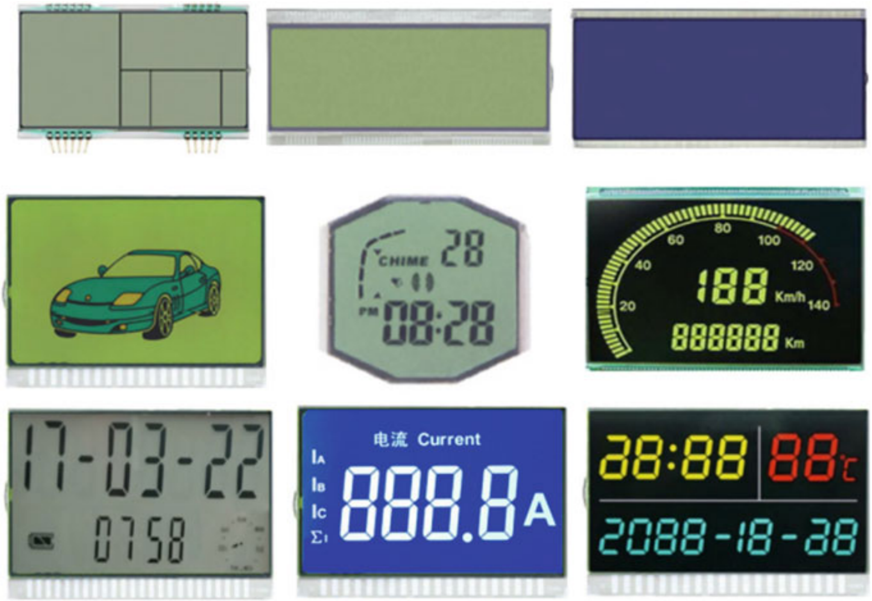


Fig. 2 Illustrations of different glass LCD technologies [8]

using TFT LCD as its standard display. The passive LCD now has been mainly used for industrial control applications.

2.2 TFT LCD

The display technology that emerged after passive LCD was TFT LCD. The TFT LCD structure from top to bottom is constructed by the upper polarizer, color filter, liquid crystal, TFT array cell, lower polarizer, and a backlight module. Depending on the arrangement and rotation mode of the liquid crystal, TFT LCD technology can be divided into TN, vertical alignment (VA), and in-plane switching (IPS). Among them, IPS, which combines stable color over a wide viewing angle and low-power consumption, is the current mainstream technology of TFT LCD as shown in Fig. 3.

According to the different transmission modes, TFT LCD technology can be divided into the reflective mode, transmissive mode, and transflective mode as illustrated in Fig. 4. Among them, the reflective display itself has no backlight source so the power consumption is minimal, but it needs to rely on external light sources to show the display. The reflective display can be applied to wearable watches, electronic paper, or electronic picture frames. The transmissive display needs to be equipped with a backlight source, as it has no reflective area. Its light penetration rate is higher, and it can provide a higher contrast ratio and color saturation in indoor environments.

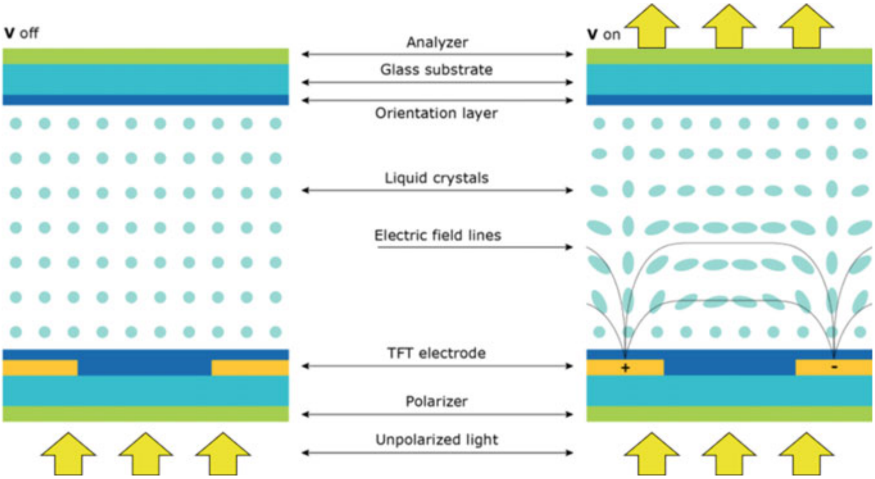


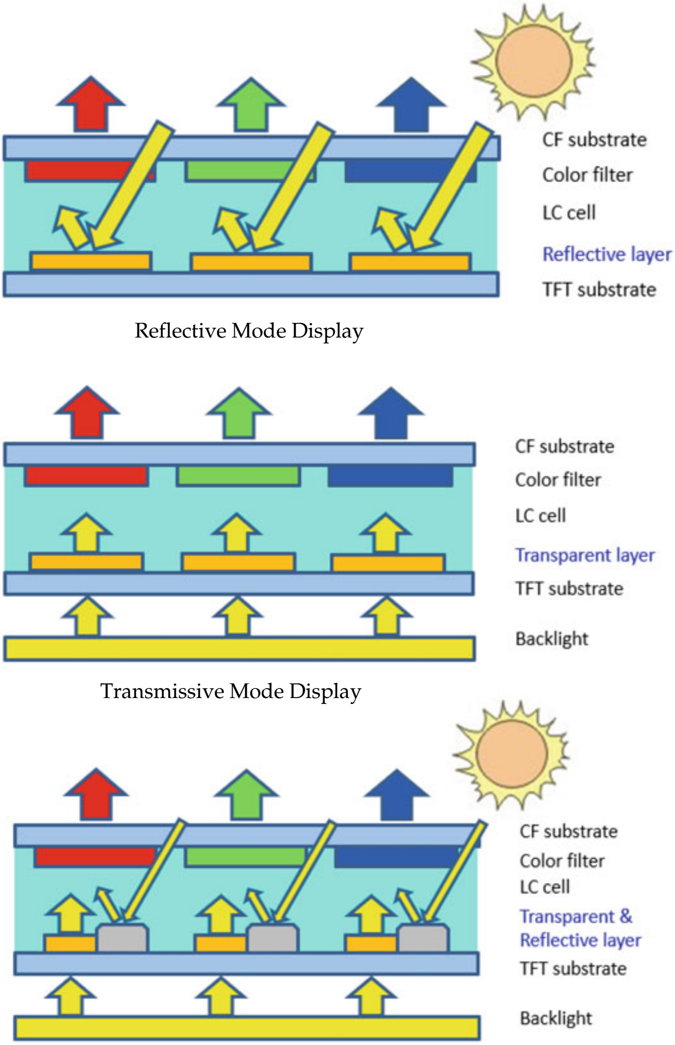
Fig. 3 In IPS technology, the liquid crystals are arranged horizontally. When voltage is applied, they rotate by 90°in the same plane on the same level [9]

For outdoor use in the sun, the brightness of the backlight needs to be increased to improve the contrast ratio, so the relative power consumption is the most. The semi-transmissive display also called transflective is somewhere in between. There is still a backlight auxiliary display in an indoor environment, but the penetration rate is low. Its contrast ratio is not as high as that of transmissive displays. For outdoor use, the reflection of sunlight can be used to improve the contrast ratio and reduce power consumption. Therefore, the transflective visual effect under the sunlight is better than that of transmissive displays.

The function of the backlight system of TFT LCD is to light up the entire display area. Its dark-state brightness mainly determines the contrast ratio of the display. To meet the viewing angle requirements of the display specification for automotive applications, most panel manufacturers will use the photo-alignment (PA) process to reduce light leakage to improve the contrast ratio as shown in Fig. 5. By using the PA process, Sharp demonstrated that the contrast of its UV2A LCD can be as high as 5000:1 with an extremely deep black performance, which is more than 60% higher than the traditional panel [10].

Some in-vehicle display providers will further control the viewing angle of the panel to prevent the display image from being reflected on the windshield to affect the driver's view. To do so, it is necessary to use the light control film (LCF) to limit the viewing angle of the panel as shown in Fig. 6.

In terms of TFT drive substrate technology, as illustrated in Fig. 7, the mainstream has been converted from amorphous silicon (a-Si) substrate to low-temperature poly-silicon (LTPS) substrate because the electron mobility of LTPS TFT is one hundred times higher than that of a-Si TFT (>100 cm^2/Vs). The speed of carrier mobility of each TFT substrate is also highlighted in Fig. 7. Together with the TFT size



Reflective Mode Display

Transmissive Mode Display

Transflective Mode Display

Fig. 4 Illustrations of reflective mode, transmissive mode, and transflective mode displays

and line width reduction, a higher aperture ratio (Fig. 8), a higher resolution, and a narrower border design (Fig. 9) can be achieved. Another substrate technology whose characteristics are between the a-Si and LTPS substrates is indium gallium zinc oxide (IGZO) substrate. Currently only Sharp, Samsung and LG have obtained patent authorization from the Japan Science and Technology Agency (JST). Although the display performance of IGZO is not as good as that of LTPS, its yield rate is relatively high. IGZO is suitable for large-size displays (TVs), which can improve

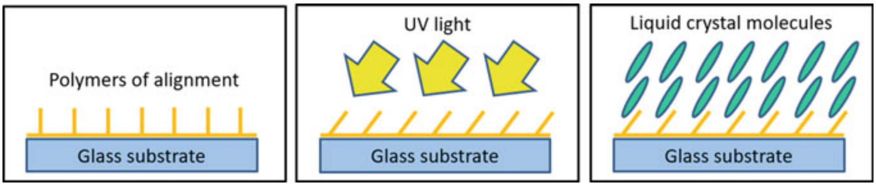


Fig. 5 Left: The glass substrate is coated with a special material developed by Sharp as a photo-alignment film. Middle image: When UV light is irradiated, the polymer of the alignment film is automatically guided to the angle of UV light irradiation. Right: The pretilt angle of the liquid–crystal molecules is automatically oriented to the direction of the alignment film polymer

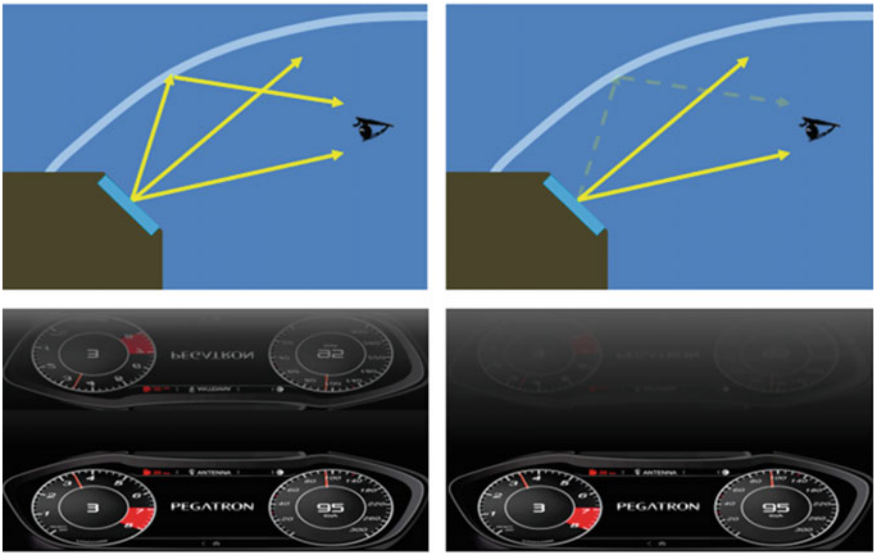


Fig. 6 By installing the view angle control film, the light reflection on the widescreen can be prevented

the panel resolution and reduce the cost. However, IGZO is very sensitive to light, moisture, and oxygen. It is not suitable for automotive display applications which have high-reliability requirements.

2.3 OLED

The light-emitting principle of OLED is to make electrons and holes travel through the transport layer under an applied voltage and then combine in the organic light-emitting layer to generate luminescent excitons. When energy is released, photons are generated, which form the OLED display. Since it is a self-luminous device, it

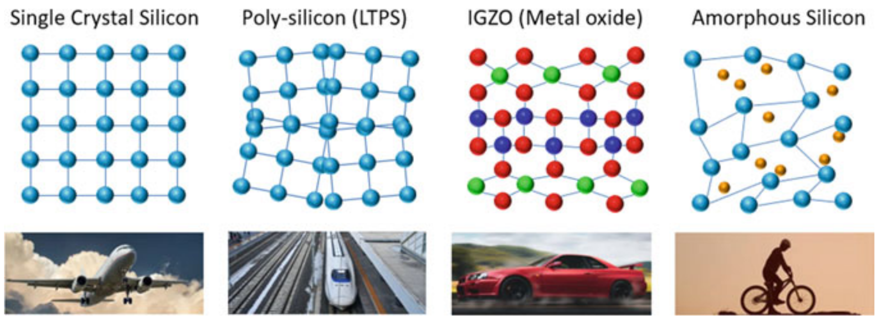


Fig. 7 Polysilicon is a form of silicon that sits between the amorphous silicon of low-cost LCDs and the single crystal silicon of semiconductor chips. The carrier mobility in LTPS TFT is 100 times (>100 cm^2/Vs) higher than in a-Si TFT

- Higher aperture ratio and brightness
- Lower power consumption
- Smaller line width/spacing
- Higher resolution

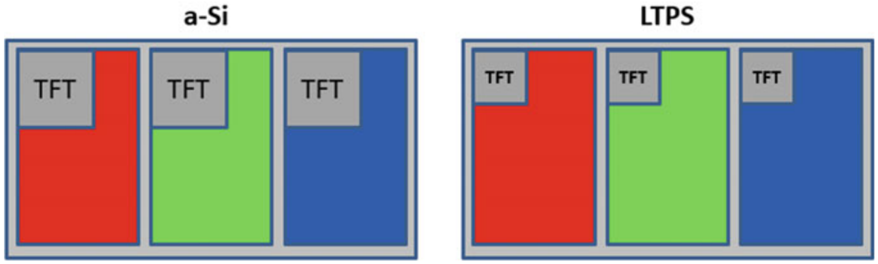


Fig. 8 Display high-aperture ratio demonstration

- Narrow border
- Lower cost

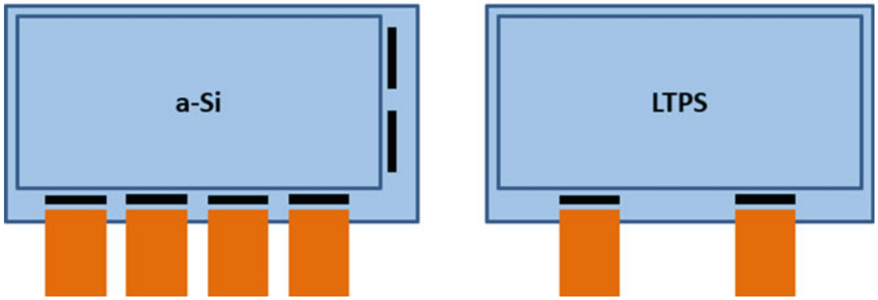


Fig. 9 Border area comparison between a-Si and LTPS

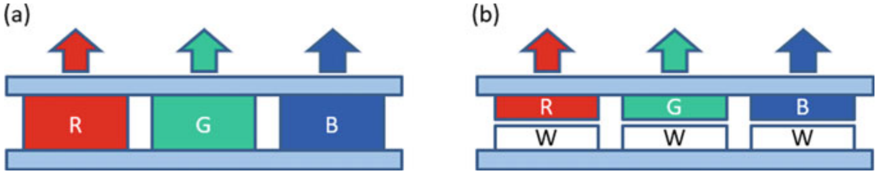


Fig. 10 **a** RGB-SBS versus **b** WOLED pixel arrangement

does not require a backlight module. When its pixel is not lighted up, OLED can achieve the darkest as compared to the TFT LCD which always has a small amount of backlight penetration. Therefore, OLED can have an ultra-high contrast ratio. The other advantages of OLED include low-power consumption, wide viewing angle, fast response speed, and high-color saturation. If the glass substrate is a polyimide substrate, the OLED display can be made flexible and transparent. However, the limitations of OLED are its reliability and durability. Its brightness also cannot meet the automotive display brightness specification. OLEDs are becoming more and more popular for general consumer-grade products such as mobile phones and TV. Although currently, a few car models have introduced OLEDs such as the Lexus 2010 RX with PMOLED displays in the dashboard and Cadillac 2021 Escalade with the first OLED concave appearance infotainment screen [11], all of them adopted a relatively reduced display performance as a trade-off to meet the automotive display reliability requirements. Therefore, changing OLED displays from consumer-grade to automotive is a challenge because to survive in the automotive environment, it has to work under extreme environmental conditions such as from -40 °C to 85 °C plus under high humidity and UV irradiation.

There are currently two major OLED technologies as shown in Fig. 10:

a. RGB-SBS (side-by-side) technology: Arrange RGB pixels through an evaporation process. The advantage is low-power consumption. But the panel size and resolution are limited. The main supplier is Samsung, and it is mainly used for mobile phones.
b. White OLED (WOLED) or WRGB technology: Convert white light OLED into color OLED through the color filter. The advantage of this technology is that it is suitable for large-size and high-resolution OLED production. The disadvantage is that the transmittance and color saturation are low. The main supplier is LGD, and it is mainly used for TV.

In terms of OLED drive substrate technology, the mainstream technology is LTPS substrates. But some mobile phones are combining the advantages of LTPS and IGZO to introduce new low-temperature polycrystalline oxide (LTPO) substrates in 2021. LTPO has higher mobility so a faster response speed, and lower power consumption than LTPS (saving 5–15%). The low-power consumption is a very important advantage for future 5G applications.

2.4 LED, Mini-/Micro-LED

Compared with the organic materials of OLED, which have relatively poor reliability and durability performance, mini-LED and micro-LED displays use more reliable inorganic materials, which can be said to be born for vehicle display technology. Partnering with quantum dot technology (QLED), LED display color saturation can be further increased. QLED uses a photoluminescent quantum dot (QD) inorganic material on top of blue LEDs. Part of the blue input light is then converted into green and red output light within narrow spectral bandwidths. The color gamut of a QLED display can be very similar to that of an OLED display. New display technology which is on a fast rise is mini-LED and micro-LED. Between the mini-LED and micro-LED displays, 75 um pitch size is used as the dividing point. The size larger than 75 um is called mini-LED, and the size smaller than 75 um belongs to the scope of micro-LED. A large number of LEDs are mounted directly to a common backplane to comprise the display. Each LED is a sub-pixel component. The manufacturing process is similar to semiconductor IC, involving epitaxial growth of LEDs, and pick-and-place mass transfer steps for the LEDs placed onto a backplane. Two applications use mini-LED in the display, passive or active matrix. For the active fully-array backlight application, whose brightness can be further controlled by 2D local dimming (Fig. 11) to improve the active contrast ratio in dark scenes to more than 100,000:1 by dimming backlight zones. The lights behind the LCD layer are calculated to adapt to the picture pattern displayed, drastically improving the contrast ratio of LED display to be comparable to OLED. As an example, the 2021 version of the Apple 12.9-inch iPad Pro uses nearly 10,000 mini-LED chips as the backlight. At the current stage, the cost of a mini-LED display is still higher than US$100, and the cost of a mini-LED backlight module accounts for more than 60% of the display cost (Fig. 12). Therefore, there is room for price reduction in future with the technology advancing and the manufacturing process maturing. Mini-LED and micro-LED could be game-changer for general display applications if industrialization succeeds. As the RGB pitch continues to shrink, mini- and micro-LED have the opportunity to enter into medium-sized display markets, which is the main size for automotive displays. Mini-/micro-LED's high-brightness and high-reliability characteristics are especially suitable for vehicle display applications. Figure 13 illustrates the concept of mini-LED in-vehicle curved, free-form display demonstrated by Innolux Corporation.

At present, micro-LED display products have been launched one after another. For a display application, due to the very small size of LEDs, the yield rate of mass transfer and defect unit repair is still the bottleneck that determines product cost. Figure 14 illustrates the LED mass transfer techniques. Take full HD resolution as an example, the total number of LEDs is $1920 \times 1080 \times 3(\text{RGB}) = 6{,}220{,}800$ (6.22 million). The yield rate after mass transfer and repair must reach 99.9999% to achieve a goal of single-digit defective dots. Then, let us take 8 K/4 K resolution as another example. The total number of LEDs is: $7680 \times 4320 \times 3(\text{RGB}) = 99{,}532{,}800$ (99.5 million). The yield rate after mass transfer and repair must reach

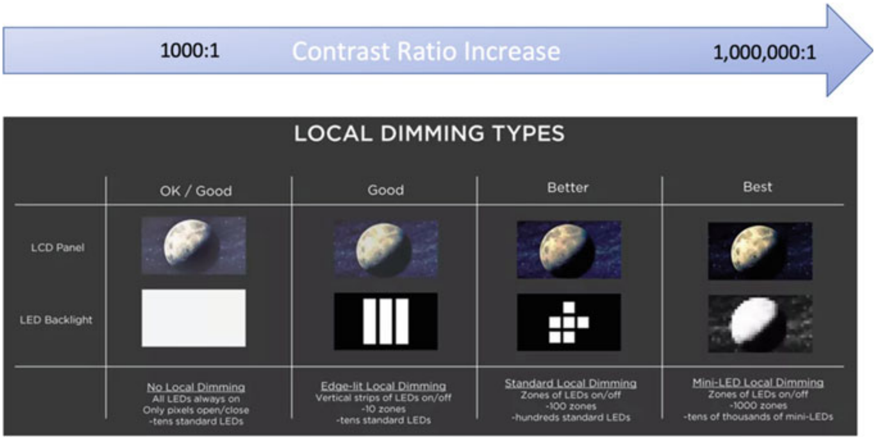


Fig. 11 Contrast ratio versus number of local dimming area [12]

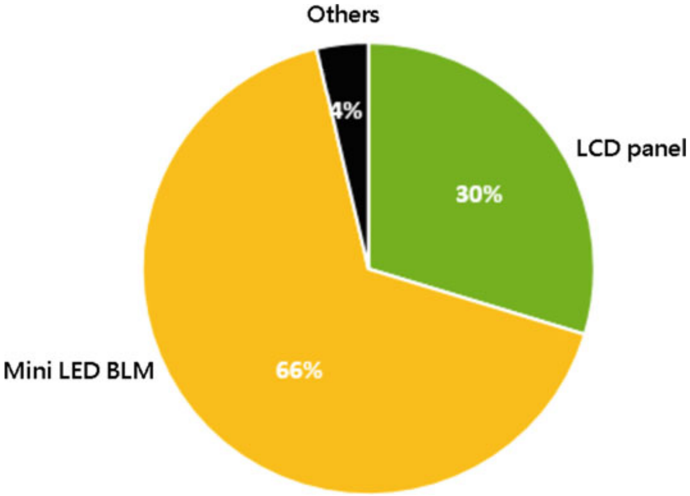


Fig. 12 Mini-LED backlight module cost proportion in the entire display. *Source* TrendForce, May, 2020

99.99999% to reach a goal of single-digit defect dots. In addition to the potential unstable yield management challenge, other manufacturing challenges of micro-LED display include complicated manufacturing processes, unstandardized processes and equipment, and multiple participants in the supply chain. It should be noted that there is still a lack of long-term field reliability data for micro-LED displays as compared to traditional LED and even mini-LED. Therefore, a comprehensive investigation of its long-term reliability for vehicle usage is critically needed within the industry.

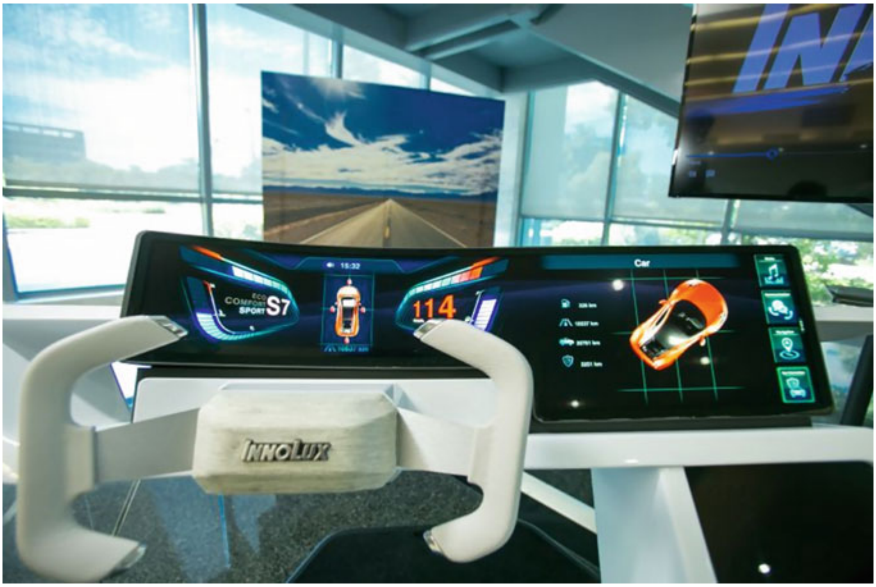


Fig. 13 Innolux Corporation mini-LED in-vehicle curved, free-form display [13]

2.5 *Head-Up Display*

In terms of head-up display (HUD), according to display type, there are combiner HUD and windshield HUD as shown in Fig. 15. Traditionally, OEMs still like to use windshield HUD based on esthetics and safety considerations. But windshield HUD must use the wedge-shaped glass to eliminate ghost images (i.e., two virtual images), which results in an increased cost. According to the light source, the HUD display can be divided into TFT LCD, DLP, laser MEMS, etc. Among them, laser MEMS has the smallest size, lowest power consumption, and highest imaging quality. One serious challenge for HUDs is the highly constrained packing space within the dashboard of the vehicle. As the field of view and virtual-image-distance (VID) increase, it requires a larger concave mirror, increased distance between the concave mirror and projecting unit, higher optical magnification (concave mirror more curved), and increased luminous flux to maintain luminance. A further challenge is limited readability with polarized sunglasses, which is because the angle between the tilted windscreen and the driver's line of sight is in the vicinity of the polarizing Brewster's angle for a typical vehicle [15]. It should be noted that HUD displays are also required to be designed to be used in extremely high-temperature conditions (>100 °C) and high illumination by sunshine and backlight, which is another challenge.

In the AR-HUD application as illustrated in Fig. 16, a further VID is required to be combined with the surrounding road condition, a brighter light source to compensate for the outside sunlight, and a wider field of view (FOV). AR-HUD imposes the additional challenges of requiring real-time monitoring of the driver's eye position

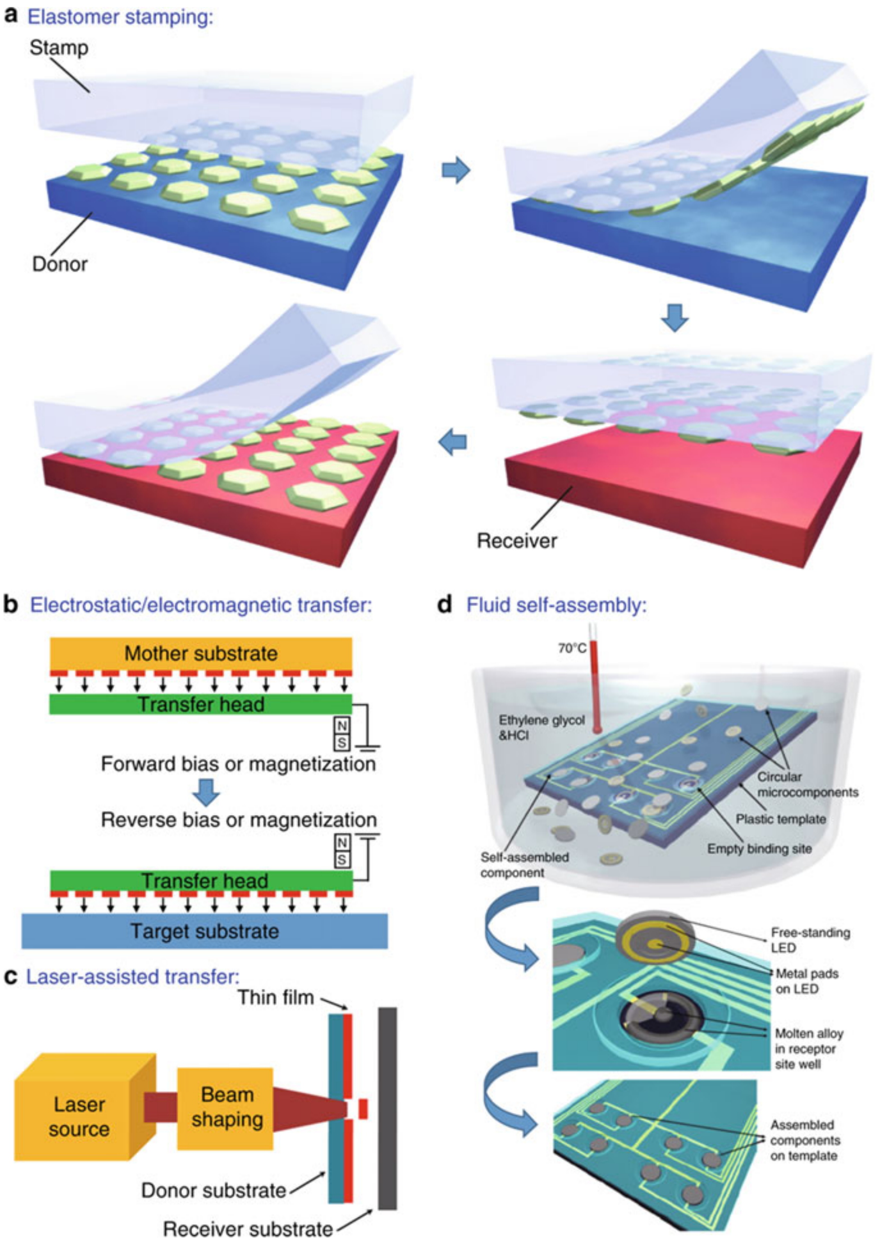


Fig. 14 Micro-LED mass transfer techniques. Schematics of **a** elastomer stamping, **b** electrostatic/electromagnetic transfer, **c** laser-assisted transfer, and **d** fluid self-assembly [14]

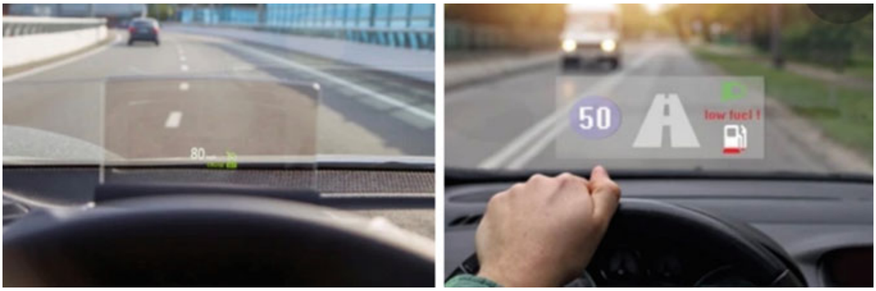


Fig. 15 Combiner HUD versus windshield HUD

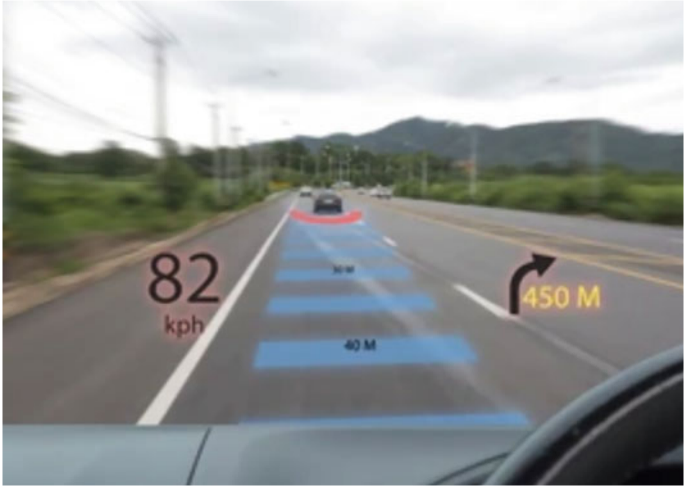


Fig. 16 AR-HUD by Panasonic to provide superior driving assistance

to perform a real-time calculation of the correct placement of AR graphical elements. Currently, the major Tier 1 HUD players are Nippon Seiki, Continental, and Denso.

2.6 Flexible and Free-Form

In the automotive space, generally, non-flat interior surfaces exist. Since OLED, mini-/micro-LED is a self-luminous display without a backlight module, they have a great advantage in making flexible and free-form displays. As long as a polyimide substrate is adopted, good flexibility can be achieved. In contrast, the traditional TFT LCD needs to go through a cell thinning process to further achieve a smaller deflection curvature, and it must be equipped with a curved backlight module. Then during the assembly, great attention must be paid to the compatibility of the two curvatures to avoid Mura problems caused by mechanical interference and uneven built-in stresses. Free-form design must be accomplished by customized circuit and sealant design, a special-shaped cutting and CNC manufacturing process, and an assembly process of a custom-shaped backlight module. Free-form in-vehicle display typically is a thoroughly customized display. This type of display is generally used in complying with car instruments per interior fitting and appearance design requirements. Examples are shown in Figs. 17 and 18.

The popularity of adopting electric vehicles leads to the rapid development of automotive electronics. The display of 3D curved surface design and multi-screen laminated together for automotive displays seems to have become a foreseeable trend

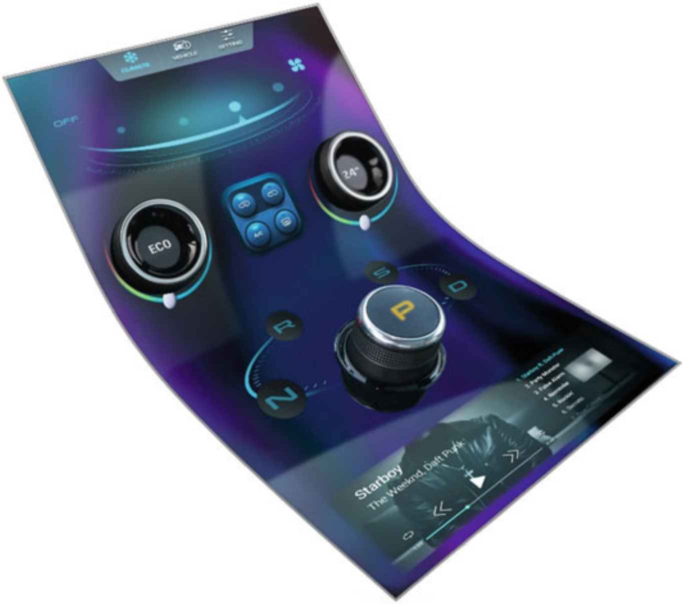


Fig. 17 Concept of a car center console featuring a curved organic LCD

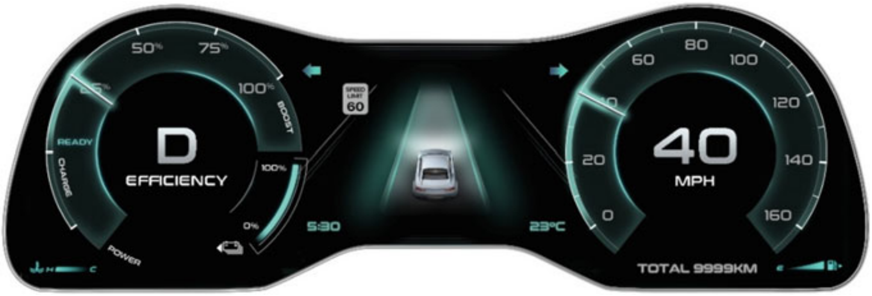


Fig. 18 Sharp free-form in-vehicle display with IGZO technology

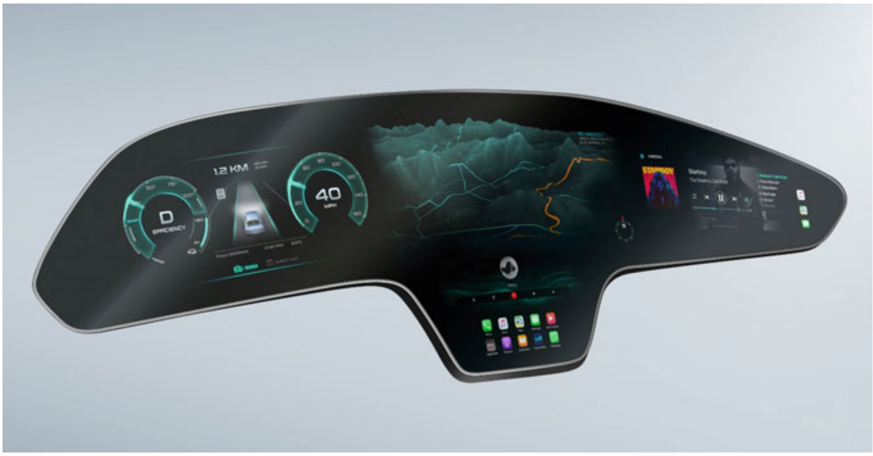


Fig. 19 Intelligent display design concept

in the automotive industry. In the past, the mechanical control units in the driving cabin are the dominating players to control the vehicle's functionality. The car display acts only as a supporting role for driving. Now, thanks to the good user experience of the touch function provided by the mobile phones for almost all the users, people are happy and adaptive to the modern display with embedded touch functions for car applications. Many mechanical control devices have been replaced by touch functions in display. Therefore, the in-vehicle display with touch is becoming a dominating role not just for information display but also for vehicle control and even aesthetic purposes. The intelligent display design concept is shown in Fig. 19.

2.7 Touch Technology

With the maturing of autonomous ride-sharing driving, driver identification and personal settings by fingerprint recognition and/or voice recognition would be required features for future AV ride-sharing vehicles. For such vehicles, displays will no longer be just an output device, but an input and output system.

Touchscreen-based smartphone applications have successfully led to the rise of capacitive touch technology. The current touch technology used in cars as the major interface between a driver and vehicle infotainment system is also mainly based on capacitive touch technology. Touch display serves as a very important human–machine interface (HMI) in the vehicle so that the display not only exhibits important driving information to occupants but also transfers driver's orders to the vehicle. The sensitivity and accuracy of touch sensing play an extremely role in the interaction of human beings and vehicles if voice control is not widely adopted. Some display features even incorporate haptic feedback, gesture control, and hover touch functions

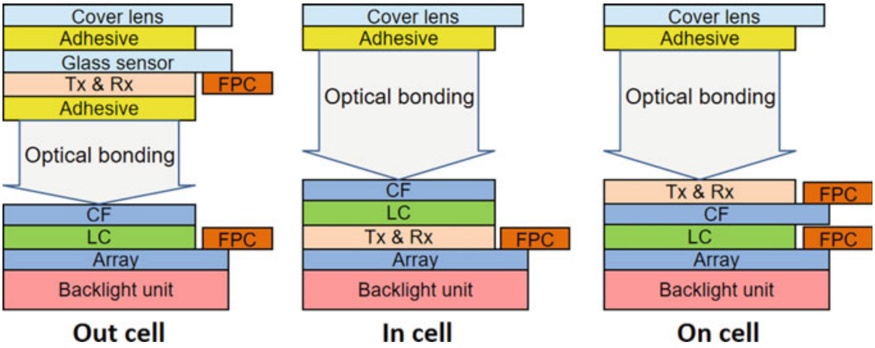


Fig. 20 Illustrations to show out-cell, in-cell, and on-cell touch architecture

similarly to a smartphone so that the driver can confirm the operation free from distraction. Depending on the location of the touch sensor inside the display module, the touch technology can be roughly classified as out-cell touch, on-cell touch, and in-cell touch, as shown in Fig. 20. The cover material of out-cell touch can be divided into glass type (commonly designed by European and American car manufacturers) and plastic-type (commonly designed by Japanese car manufacturers). Based on the use of different substrate materials of the touch sensor, it can be divided into glass type (corresponding to high-reliability requirements) and film type (corresponding to 3D surface design requirements), respectively. If the in-vehicle display adopts film type, the material must be selected to avoid the rainbow effect due to birefringence. The common combinations of out-cell touch for in-vehicle usage are one glass solution (OGS), glass lens + glass sensor (G/G), glass lens + film sensor (G/F), and plastic lens + film sensor (P/F). The Tx/Rx circuit of the touch sensor is not limited to just a single-layer substrate, and a double-layer substrate can also be implemented. The combination of glass and film may have G/F (Tx/Rx on the same side of the single-layer film), G/F2 (Tx/Rx on the upper and lower sides of the single-layer film), and G/FF (Tx/Rx each on one film) as shown in Fig. 21.

However, with the evolution of display technology, touch technology has gradually evolved from out-cell touch to on-cell touch and in-cell touch. In on-cell touch and in-cell touch, the Tx/Rx circuit is integrated above or inside the display. Compared with out-cell touch, in-cell touch has the following advantages:

a. Cost reduction: When the touch and display technology are integrated, the driver ICs of the two can also be integrated into a touch with display driver integration (TDDI). Therefore, the separated touch module conceptually is eliminated in the structural stack, which brings a simplified process with at least one less bonding step. Even if the yield difference is not substantial, the overall BOM cost can be significantly reduced.
b. Thickness reduction: Due to the absence of out-cell touch module on the structural stack, the total thickness of the touch display module can be reduced.

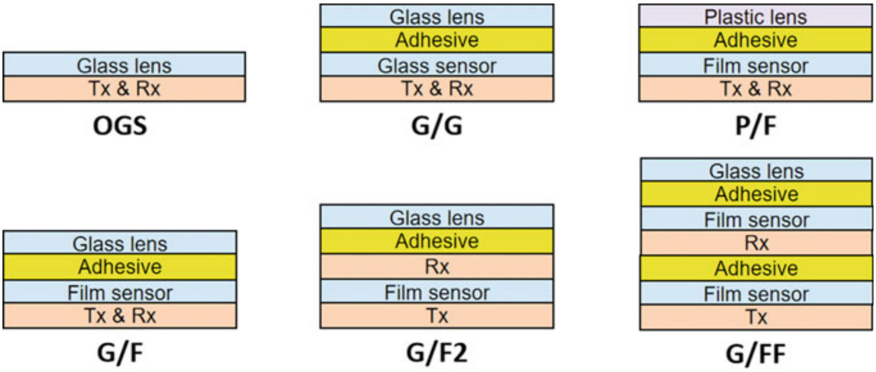


Fig. 21 Combinations of glass and film for various in-vehicle tough applications

c. Improved optical performance: Since there is no touch circuit outside of the display structure stack, the overall display can achieve higher transmittance, lower reflectivity, and less color shift.

d. Glass strength: Compared with one glass solution (OGS) which is the same single-layer glass and has only the front and back sides strengthened by chemical, in-cell touch's cover glass (CG) is a six-sided chemical strengthening glass. The test data shows that the strength of the glass of six-sided chemical strengthening glass is superior to two-sided chemical strengthening glass, and it has good resistance to glass fracture after head impact test (HIT). Although G/G also can have 6-sided chemical strengthening glass, with the same design thickness, in-cell touch CG thickness is equivalent to G/G glass total thickness. Therefore, even compared with G/G, in-cell touch can provide a higher glass strength to pass HIT than G/G can. Therefore, in-cell touch will likely become the mainstream touch technology for automotive displays in the next few years. On the other hand, touch gesture recognition, proximity sensing, and hovering sensing are expected to grow as the future sensing options for vehicles.

The design of an automotive-grade capacitive touch screen needs to consider meeting the reflectivity below 4% and superior grayscale requirements. The most important evaluation point of the surface treatment effect of the capacitive touch screen or the optical treatment effect between the touch screen and the LCD screen is the reflectance. As the vehicle infotainment display touch screen serves as a human–machine interface, in general, it is located at the center of the instrument cluster at a position higher than other vehicle control components and directly under the windshield. Therefore, the intensity of incoming light will be very high-under sunlight. To ensure that the driver can see the contents of the screen clearly without being affected by light reflections, optical compensations such as using PC/PMMA-based anti-reflection film on the cover glass surface and anti-glare/anti-reflect (AGAR) optical coating between ITO and foam are implemented.

3 In-Vehicle Display Requirements

In-vehicle display in general has more stringent requirements than consumer-grade display in regrading to performance and reliability. Performance specifications required for in-vehicle displays are brightness, power consumption, chromaticity, screen size, narrow border, visibility, and cockpit design compatibility. There are also specific reliability, EMC/EMI, and functional safety requirements for in-vehicle displays. Visibility under sunshine has been a top priority in automotive displays. The trend of increasing display size and the rapid adoption of electric vehicles trigger the demand for lower power displays. During night operation and while driving in the tunnels, a higher contrast ratio requirement is needed for in-vehicle display as the drivers may see annoying luminance on a black screen if the contrast ratio is not high enough, i.e., 1500:1. High-color purity may also be needed in case the warning icons are integrated into the display. Due to the vehicle's potential extreme weather condition exposures, display performance at high and low temperatures will be the key reliability requirement. Response time at lower temperatures and LCD sheet wrinkle at higher temperatures are all needed to be seriously considered. All in-vehicle displays shall be able to still work normally under the conditions of electromagnetic interference in the car and large fluctuations of operating voltage.

3.1 Optical Performance Requirement

Given the vigorous development of the automotive display, the four major European car manufacturers, Volkswagen, Audi, BMW, and Porsche jointly formulated the display specification for automotive application (DSAA) in 2009. This is the only specification available in the industry covering some basic display performance requirements such as brightness, chromaticity, contrast ratio, viewing angle, response time, reliability, and inspection specifications to meet the needs of OEMs for in-vehicle display. This specification reflects the common intention of the automotive OEM working group as a standard to ensure optimum customer experience and maximum compliance. Table 1 shows the general optical performance requirements for in-vehicle display based on DSAA standards.

3.2 Appearance

In-vehicle display appearance requirements include display size, resolution, seamless, and shape. Table 2 shows the mainstream sizes and resolutions for different in-vehicle displays from J. Liu et al. [16].

Table 1 Optical performance requirement

Item	Description	Temperature	Required value	Information
1	Viewing angle area		A+-area: H ± 10°, V + 8/−4° A-area: H ± 40°, V + 20/−10° B-area: H ± 50°, V + 20/−10°	A+ zone is a narrow center zone where driver is the primary viewer. A and B zones are wider to address center information displays where the passenger and driver are both viewers
2	Sunglasses		Polarization angle of display between 12:00 and 01:30 allowed	Readability with polaroid sunglasses must be guaranteed
3	Color white	+ 25 °C	White point reference point xref = [0.307 – 0.318]; yref = [0.318 – 0.321];	
4	Color red	+ 25 °C	$\lambda_{\mathrm{dom}} = 623 + 7 -$ 5 nm, Sat > 85%	Ref: (x = 0,661; y = 0,306) all pixels in red condition
5	Color green	+ 25 °C	$\lambda_{\mathrm{dom}} = 549 + 5 -$ 5 nm, Sat > 80%	Ref: (x = 0,298; y = 0,662) all pixels in green condition
6	Color blue	+ 25 °C	$\lambda_{\mathrm{dom}} = 469 + 5 -$ 5 nm, Sat > 90%	Ref: (x = 0,137; y = 0,067) all pixels in blue condition
7	Color black	+ 25 °C		target: equivalent to color white
8	Tolerance white (calibrated)	+ 25 °C	x/y = ± 0.005	Valid only for system supplier after calibration. Luminance from 1 cd/m2 up to maximum luminance
9	Tolerance white	+ 25 °C	x/y = ± 0,03, target value: x/y = ± 0.025	Valid for the whole module (display panel + backlight); measured perpendicular and in center
10	Tolerance black	+ 25 °C	to be agreed during development	

(continued)

Table 1 (continued)

Item	Description	Temperature	Required value	Information
11	Tolerance white (high-temperature drift)	+ 85 °C	x = +0/−0.03; y = +0/−0.03 based on initial value	Valid for the whole module (display panel + backlight)
12	Tolerance white (lifetime)	+ 25 °C	x/y = ± 0.01	
13	Color depth		Minimum 3 × 8 bit	
14	Luminance over viewing angle within A + area	+ 25 °C	Minimum 800 cd/m^2	Luminance values valid after white calibration
15	Luminance over viewing angle within A area	+ 25 °C	Minimum 450 cd/m^2	Luminance values valid after white calibration
16	Luminance over viewing angle within B area	+ 25 °C	Minimum 300 cd/m^2	Luminance values valid after white calibration
17	Luminance	+ 85 °C	Minimum 50% of values item 14,15 and 15	
18	Luminance	+ 70 °C	Minimum 80% of values item 14,15 and 15	
19	Luminance lifetime		Minimum 80% of values item 14,15 and 15	
20	Dimming minimum brightness		< 1 cd/m^2	Full white image
21	Contrast	+ 25 °C	Minimum 1200:1 for item 14 Minimum 650:1 for item 15 Minimum 350:1 for item 16	
22	Contrast	+ 85 °C	Minimum 80% values item 21	
23	Gamma uncalibrated	+ 25 °C	г = 2.2 ± 0.3	
24	Gamma calibrated	-30–85 °C	г = 2.2 ± 0.2	
25	Uniformity white luminance	+ 25 °C	>= 75%	For whole display area
26	Uniformity black luminance	+ 25 °C	>= 50%	For whole display area
27	Black Mura	+ 25 °C	Relative luminance gradient: $Z\alpha_{,\mathrm{relW,\ max}}$0.02% /mm	Normalized to the white screen

(continued)

Table 1 (continued)

Item	Description	Temperature	Required value	Information
28	Reflection (systems without TP)	23 °C ± 2 °C	Rdi < 1% −6.0 < a* < 0 −6.0 < b* < 0 Rde > 0.3*Rdi	Reflectance and color measured separately for specular included (di) and specular excluded (de) (ASTM E 1164–09), Y, a*, b* (CIE 15.2), Reflectance R is given by Y(D65) (DIN 5033–4)
29	Reflection (systems with capacitive TP, measured is TP + Display)	23 °C ± 2 °C	Rdi < 1.7% −6.0 < a* < 0 −6.0 < b* < 0 Rde > 0.3*Rdi	

Note All values have to be fulfilled with cover glass, touch panel, LCF—if required—and after white calibration. Warm-up time before measurement: 30 min. Ambient temp: +25 °C. Backlight (if not specified): 100% PWM. All values will be measured perpendicular if there is no viewing angle (area) specified

Table 2 Automotive display sizes and resolutions

Application	Display size	Resolution
Cluster		
Mixed cluster	3.5” to 8”	800 × 480
Full digital	10.25” to 12.3”	1280 × 480, 1920 × 720
Central information display		
Entry	6” to 8”	800 × 480, 1280 × 720
Middle	9” to 15”	1280 × 720, 1920 × 1080
High	12.3” and above	1920 × 720, 1920 × 1080
Head-up Display		
Level 1	1.8”	640 × 320
Level 2	3.1”	8000 × 480

3.3 Integration and Fabrication

Traditional vehicle’s display and touch system integration methods mainly are LCD panel manufacturers integrate open cell and backlight to form LCD module, and touch panel (TP) panel manufacturers integrate cover glass and touch sensor to form touch module. Then, LCD module and touch module are integrated by bonding at final assembly, testing, and packaging (FATP) manufacture. The FATP manufacturer can be a TP manufacturer, a panel manufacturer, a laminating manufacturer, or Tier 1 supplier. Tier 1 suppliers are those companies in the automotive supply chain that can directly conduct business with car OEMs. At last, Tier 1 will integrate other systems' electrical and mechanical components to form an on-board display system, and then deliver this system to car OEMs. However, today's display and touch system integration methods have changed, and the future trend is toward the following two integrated methods:

a. Display and TP makers are integrated to the back end: Display maker laminates cover glass and in-/on-cell touch display together in house or TP maker laminates cover glass, touch sensor, and LCD module together in a house.

   The electronic and mechanical parts of other components are then integrated with touch and display modules to form a complete vehicle-mounted display system. Then, this system is delivered to the car OEMs. In other words, the display/TP manufacturers directly play the role of Tier 1.
b. Tier 1 is integrated into the front end: Tier 1 obtains an open cell from the panel supplier, a touch sensor from the touch supplier, and backlight from the LED supplier. Then, Tier 1 assembles display, touch, and backlight using a full lamination process in house. After the assembly is completed, Tier 1 then integrates the other electronic and mechanical components into the touch and display module to form an on-board display system for car OEMs. In other words, Tier 1 has undertaken most of the assembly work of display and TP manufacturers.

The details of touch and display integration and fabrication process flow are listed in Fig. 22.

In addition to the aforementioned changes in the role of system integration, changes have also begun in the design of on-board displays. Example 1: The metal frame of the car display is integrated with the metal structure of the system, which can simplify the part structure to achieve a narrow border design. It also can obtain a better heat dissipation and achieve a lower manufacturing cost. The obstacles to

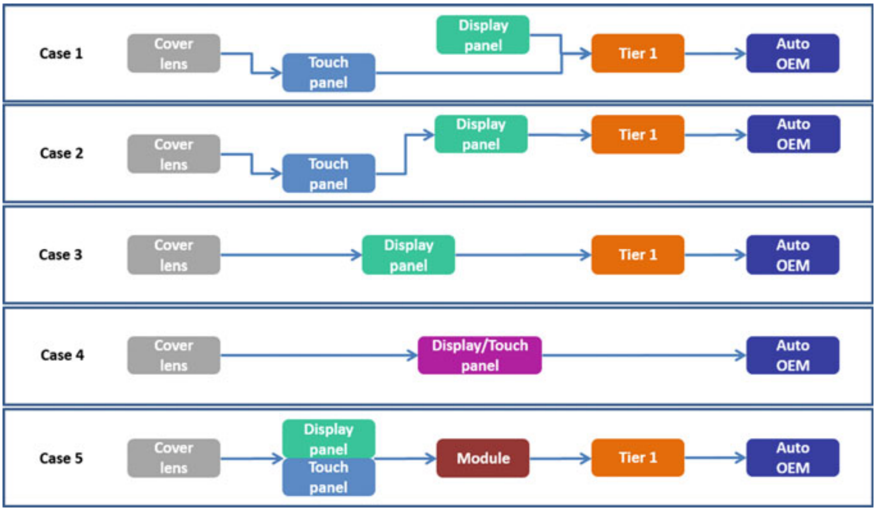


Fig. 22 Touch and display integration and fabrication process flow

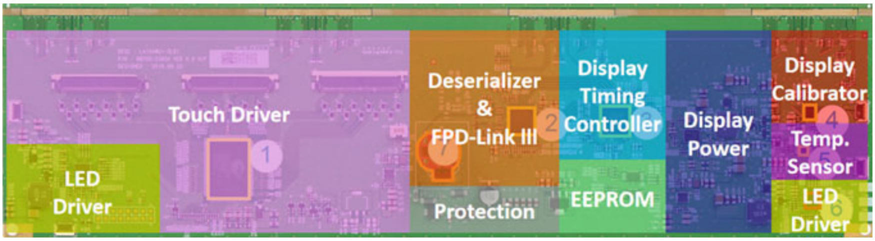


Fig. 23 Display module component of the Tesla model 3 center touchscreen display

doing such a design approach change are its higher degree of customization and an increased integration difficulty. Neither the panel maker nor Tier 1 wants to lose the dominance of design. Example 2: The driver board of the vehicle display system is integrated with the driver board of the display, which can simplify the drive structure, reduce the number of connectors and FPCs, and obtain a better signal transmission efficiency. A practical case is illustrated in Fig. 23. The touch driver has been integrated with deserializer and FPD-link III connector, display timing controller, display power, display calibrator, temperature sensor, LED driver, EEPROM, and protection on one single driver board. Although the integration of display and touch is not a surprise, the deserializer and FPD-The integration of the Link III connector means that the driver board is ready to be connected with the electronic control unit (ECU) from the car OEMs. This integration will greatly remove the boundaries between the panel maker and the car OEM. Even the panel maker has the opportunity to replace Tier 1 by doing more system integration. Some new car manufacturers have already adopted this approach. The specific benefits of doing so are that it not only can simplify the supply chain and improve management efficiency but also can reduce parts, materials, and assembly costs.

Vehicle displays have high-quality requirements. The defect rate usually cannot exceed 100 PPM. Suppliers are generally required to establish vehicle-specific assembly and testing production lines or areas and special personnel and special stations for ensuring the quality of the products per IATF-16949. To meet the required process capability index, Cpk, a total quality management system is adopted, and the corresponding key specification parameters need to be statistically controlled. The inspection is more stringent than consumer electronic products. Automatic optical inspection (AOI), manual functional inspection, and a 100% aging process are all required to be implemented to achieve the quality goal of zero defects.

3.4 Color Measurement and Characterization

Referring to the requirements of automotive application display specifications, different OEMs have different chromaticity coordinate requirements. Tier 1 must

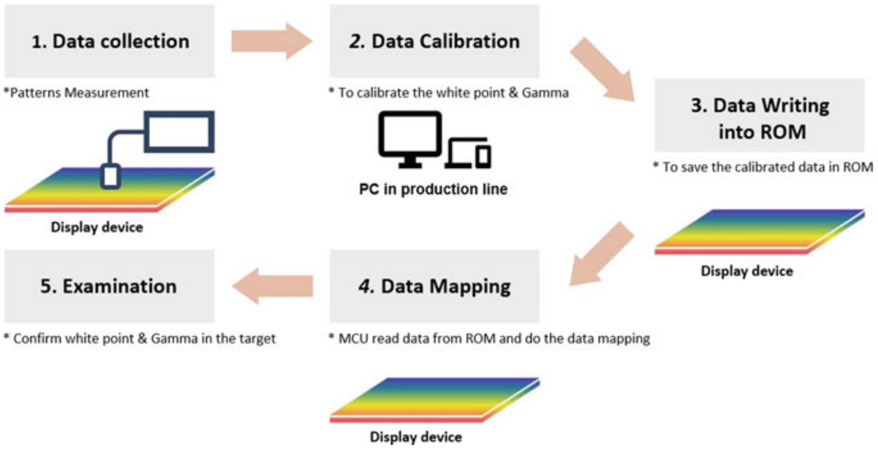


Fig. 24 WPC adjustment steps

perform white point calibration (WPC) to conduct color measurement and adjustment on the display. The reason for doing so is to make sure that the multiple displays installed from different panel manufacturers will exhibit consistent chromaticity performance in the same car.

Generally, the content of WPC includes adjustment of gamma curve and RGB table. However, in the process of RGB gradation correction, the maximum brightness of the white screen will be sacrificed more or less to meet the requirement. WPC steps are illustrated in Fig. 24 as a reference.

An example of the WPC results is as follows. As shown in Fig. 25, from the measurement results of the chromaticity coordinates, it can be seen that the white point (WP) has moved from the original red dot position to the green dot position. The WP chromaticity coordinates are still within the tolerance of WP (green box) per the customer's request. The chromaticity coordinates of WP, therefore, move from bluish to yellowish. It can be understood that this is the result of lowering the blue gradation of WP, which also causes the brightness of WP to decrease.

3.5 Mura, Defect, Inspection, and Demura

Mura is a Japanese word that means unevenness, irregularity, or blemish. For display, this word is referred to irregularities, non-uniform luminance, and "clouding" effects seen on display screens. Mura's effects on a display screen impact a user's viewing experience and can degrade display performance or functionality. There are many possible reasons for the formation of Mura. Some of the most common types of Mura and display defects include high-brightness backlight leads to increased TFT leakage current, light leakage, poor TFT film quality uniformity, non-uniform TFT thickness,

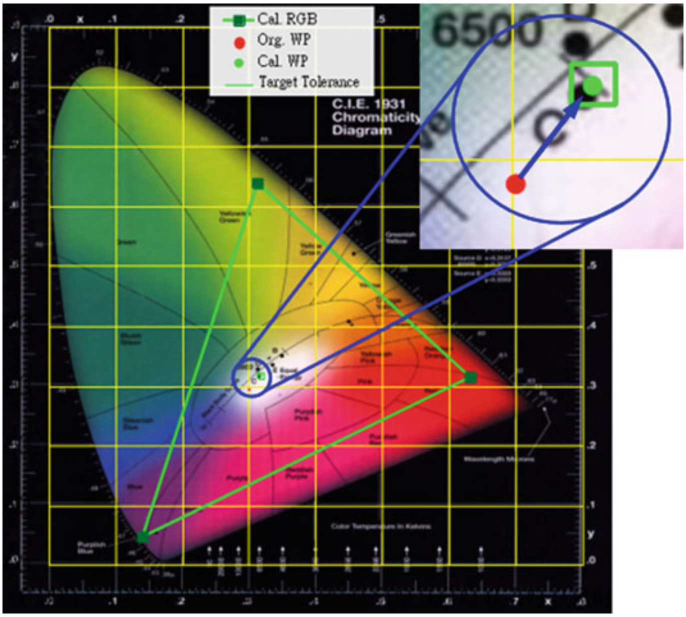


Fig. 25 After correction, the white point chromaticity coordinates will enter the range of the target value

uneven gaps between substrates due to uneven full bonding pressure, poor flatness, various interferences caused by assembly process, impurity or foreign particles in the liquid–crystal matrix, uneven aging characteristics of display components, flaws in the LCD cells, the non-uniform luminance distribution of the backlight, non-uniform color in color filters, warped light guide plate or film, environmental and mechanical stresses induced Mura. Mura can be detected and measured by an automatic imaging colorimeter. The general types of display Mura and defects are bright dots (red/green/blue), milky-way weak bright dots (red/green/blue), black dots, jointed black dots, vertical or horizontal stripes, RGB lines, and white line defects, tiny bright dot, gray spot, edge Mura, color Mura, sheet wrinkle, LCD contamination/foreign material in the display, LED dark spot, LCD bright band, frame Mura, polarization deformation, black Mura, light leakage, bubbles, corner light/bright corners, white dot, spot Mura, function defect, no image, abnormal operation, afterimage, flickering, RGB timing error, image sticking, wrong color/discoloration, lower brightness, no backlight, glass crack visible, transmission reduction, reflection increase, linked pixels, stuck pixel, rainbow coloring, yellow border, blue edges, glass scratches visible, and glass dents visible. Some examples are shown in Figs. 26 and 27. Among them, black Mura consists of large blemishes grouped and is most commonly analyzed per quality parameters for automotive displays. It is strongly impacted by mechanical stress on the panel, and IPS displays are quite sensitive to this issue [17].

The automotive display black Mura requirements specification is relative luminance gradient: Zα, relW, max < 0.02%/mm. An example is shown in Fig. 28.

During the display manufacturing, the incoming quality control and in-line Mura inspections are mostly 100% screened and judged by human eyes. The questionable parts are then checked by an image color meter combined with a neutral density (ND) filter for the second selective screen. Different panel manufacturers may have different Mura specifications for passing and failing. In general, the Mura specifications are approximately between 2 and 6% of the ND filter as shown below in Table 3.

For OLED display, brightness uniformity and image burn-in or retention are still the two main problems it faces at present. To solve these two problems, in addition to

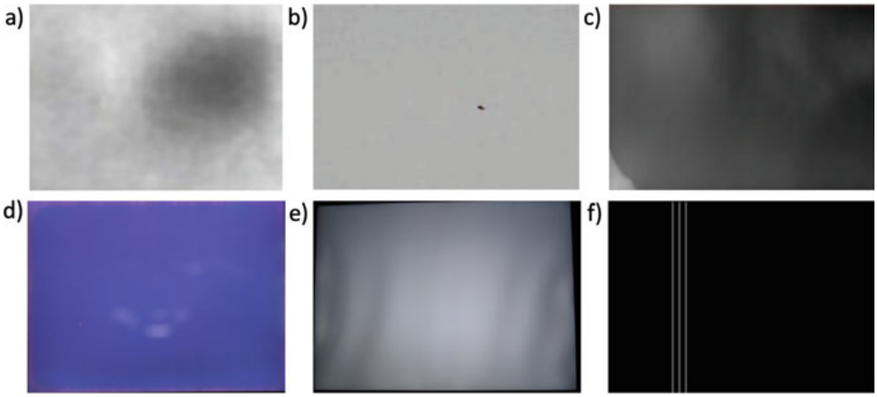


Fig. 26 Examples of various display Mura defects. **a** black Mura, **b** LCD contamination, **c** corner light, **d** frame Mura, **e** sheet wrinkle, **f** vertical lines

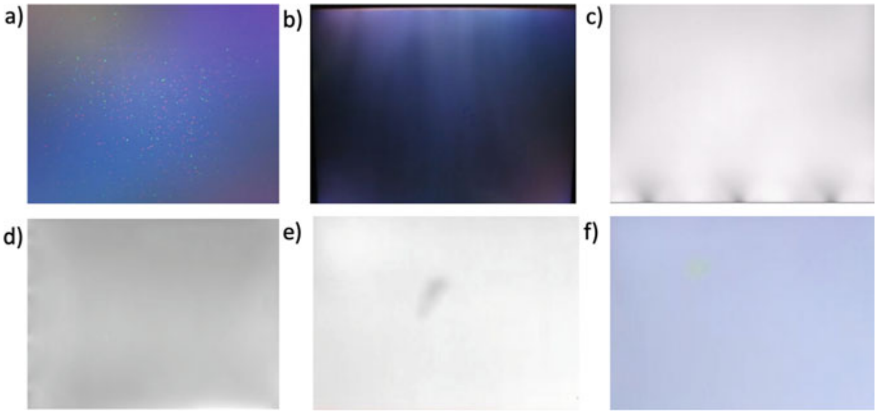


Fig. 27 Examples of various display Mura defects. **a** Tiny bright spot, **b** LCD bright band, **c** LED dark spot, **d** light leakage, **e** spot Mura, **f** color Mura

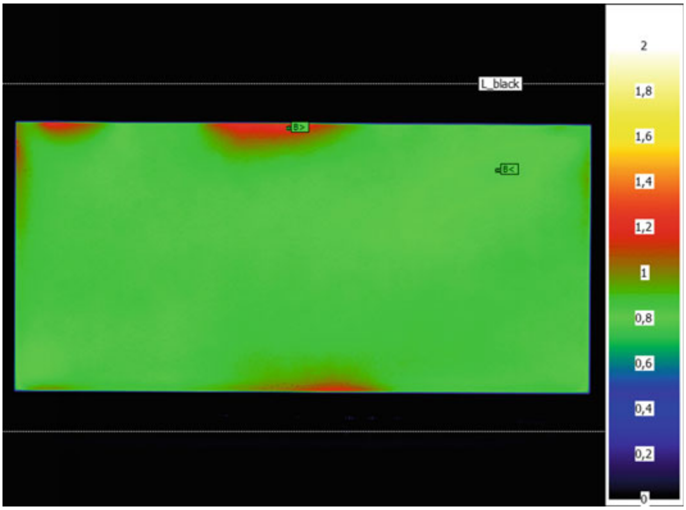


Fig. 28 Black Mura analysis example

Table 3 Mura specification examples [18]

ND-LCD type

- Size: 7.5 cm × 7.5 cm
- Depth: 90 ± μm
- Material: TAC (Tri Acetyl Cellulose)
- Transmittance: 1% ‘ 2% ‘ 3% ‘ 4% ‘ 5% ‘ 6% ‘ 8% ‘ 10%

Type and transmittance (%)

ND-LCD-1%	0.75–1.25%	ND-LCD-2%	1.60–2.40%
ND-LCD-3%	2.55–3.45%	ND-LCD-4%	3.60–4.40%
ND-LCD-6%	5.40–6.60%	ND-LCD-5%	4.50–5.50%
ND-LCD-10%	9.00–11.00%	ND-LCD-8%	7.20–8.80%

the improvement of the process, compensation technology including optical Demura must be considered. Compensation methods can be divided into two categories: internal compensation and external compensation. Internal compensation refers to a method of compensation using sub-circuits constructed by TFTs inside the pixel. External compensation refers to a method of sensing the electrical or optical characteristics of a pixel through an external drive circuit or device and then performing compensation.

Generally, the luminous brightness of OLED is directly proportional to the current, and the current is provided by the TFT, which is related to the characteristic parameters of the TFT. The current is usually expressed as

$$I = k\mathrm{Cox}(\mathrm{Vgs} - \mathrm{Vth})2(1 + \lambda\mathrm{Vds}) \qquad (1)$$

where *k* is a parameter related to the mobility of the TFT, and Vgs and Vds are related to the power supply voltage and the OLED driving voltage.

It can be seen that the parameters that affect the current are TFT mobility, threshold voltage, OLED drive voltage, and power supply voltage. The main purpose of compensation technology is to eliminate the influence of these factors, and finally make the brightness of all pixels reach the ideal value. Figure 29 shows a typical internal compensation circuit, which consists of 7 TFTs and 1 storage capacitor, so it is referred to as a 7T1C structure for short.

There are many similar circuit structures such as 6T1C, and 5T2C as internal compensation circuits. Its general working idea is to store the threshold voltage Vth of the TFT in its gate-source voltage Vgs first in the compensation stage. During the light-emitting stage, it converts Vgs-Vth into the current. Because Vgs already contains Vth, when it controls the current, the effect of Vth variation will be canceled, thereby achieving the consistency of the current across the entire panel. However, due to the influence of parasitic parameters and driving speed, Vth cannot be completely offset, that is, when the deviation of Vth exceeds a certain range (usually ΔVth ≥ 0.5 V), the consistency of the current cannot be guaranteed, so its compensation range could be limited. On the other hand, for such internal compensation circuits, their operation needs to undergo three stages: reset, compensation, and light emission. As a driving cycle must do those extra three things, there is a substantially increased driving load for the entire panel. With continuous research and development in recent years, the topological structure of the internal compensation circuit has almost been exhausted, and it seems to be difficult to have furthermore practical structural innovations.

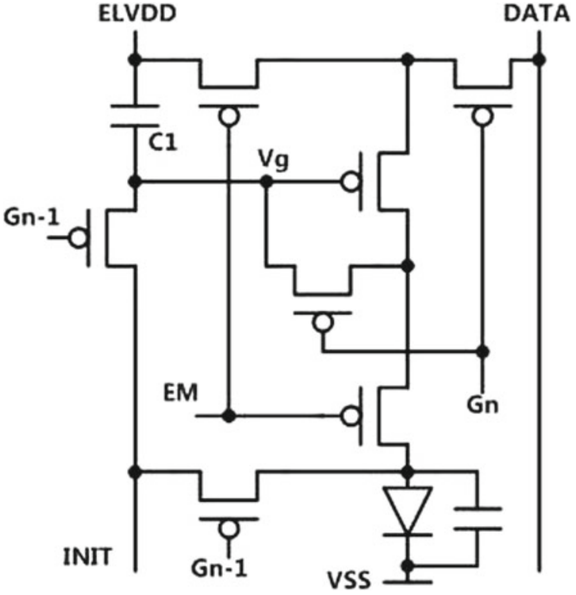


Fig. 29 7T1C internal compensation circuit illustration

External compensation can be divided into optical extraction and electrical extraction according to different data extraction methods. The optical extraction type refers to the extraction of the brightness signal through the optical CCD camera after the display is lighted. The electrical extraction type refers to the extraction of the electrical signals of the TFT and OLED through the sensing circuit of the driver chip. Because the types of signals extracted by these two methods are different, the data processing methods are also different. The optical extraction method has the advantages of its simplicity and flexibility, so it is widely used at the current stage, which is what we usually call Demura.

The general steps of Demura are:

1. Driver IC lights up the panel and displays several screens (usually grayscale or RGB).
2. Use a high-resolution and high-precision CCD camera to take pictures of the above-mentioned images. Analyze the pixel color distribution characteristics according to the data collected by the camera, and identify Mura based on the specific algorithm. There are currently two international standards for Mura detection such as the German Flat Panel Display Forum and IDMS (former VESA).
3. Demura data is generated according to Mura data and the corresponding Demura compensation algorithm.
4. Burn the Demura data to the flash ROM, re-measure the compensated screen, and confirm that Mura has been eliminated.

Figure 30 demonstrates the Demura effect. For OLED Demura technology, Samsung and LG are currently in the leading position. The Demura technology is very complicated and challenging. The difficulties of Demura are summarized as follows:

1. How to use a CCD camera to quickly and accurately capture the color of each pixel?
2. How do identify different types of Muras as some Muras are not visible from the front view but visible from the side view?
3. How to make fast and efficient compensation, to avoid the loss of productivity due to extra manufacturing process time?

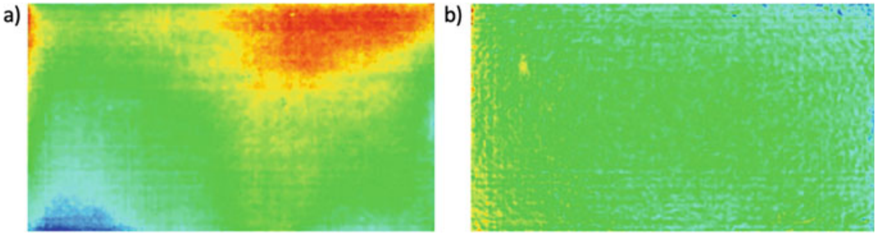


Fig. 30 Demura effect illustration. **a** before Demura, **b** after Demura

3.6 Visibility in Bright Light and Complete Darkness

The automotive display is subject to a wide array of ambient lighting conditions ranging from bright sunlight days to moonless nights. Inside the vehicle, to recognize symbols and letters clearly on the display under direct daylight, the display's visibility under bright light is necessary. Contrast ratio (CR) under bright light can be expressed as in Eq. 2, where the outdoor light strength is *S*, luminance brightness from the display is *L*, black brightness of the display is *B*, and reflectance of the display is *R*. In real calculation, the *B* can be nearly neglected.

$$\mathrm{CR} = (L + R * S)/(B + R * S) \tag{2}$$

From Eq. 2, to ensure good visibility under bright light, we can either decrease the reflectance R or improve the brightness L. As an example, by assuming direct daylight of 45000 lx is radiated per ISO16505 to ensure visibility of 3:1 per ISO15008, which is visible under direct daylight, the specular reflection ratio under 0.9% will be required.

To improve the brightness of the display, increasing electrical input power, improving backlight brightness, improving the transparent ratio of the cell, and using local dimming are the common ways. Automotive displays demand brighter backlighting. High-brightness LEDs used to backlight displays are playing a major role in this unprecedented demand. One benefit of using LED backlighting is its capability of wide dimming ratio provided by a high-performance LED driver IC. The interior of a car is subjected to a very wide variation of ambient lighting conditions, ranging from direct sunlight to complete darkness with every variation in between. Since the human eye is very sensitive to minor perturbations in light output, the screens need to dim or brighten correspondingly. Therefore, the LED backlighting system must be capable of very wide dimming ratios from 1000:1 to as high as 30,000:1. Although the minimum goal for in-vehicle display's contrast in bright light is 3:1 which is visible under direct daylight, the future requirement could be 6:1 or even 20:1 using newspaper contrast as a benchmark. Therefore, more low reflection and brightness improvement will be required.

3.7 Improvement of Image and Touch Quality

With the conversion of TFT drive substrate technology from a-Si to LTPS, coupled with the demand for large-size displays, the resolution of car monitors has also greatly increased, from the early 7″ WVGA (800 × 480) with about 133 pixels per Inch (PPI), to above 50″ with PPI increased to 800 or even 1000 due to the demand for 3D displays and AR-HUD. The image quality shall be outstanding. In addition, the quantum dot material continues to increase the color saturation of the display, and the mini-LED backlight is adopted to increase the brightness. The local dimming

Fig. 31 Concept of turning the cabin into a second living room

design improves the contrast to make TFT LCD exhibit a comparable visual effect to OLED while maintaining its high reliability. If the touch function is further built into the display to become an in-cell touch display, it cannot only reduce the display gap but also reduce the reflectivity and increase the transmittance due to without one glass substructure and two ITO layers outside of the display, which impacts the visibility of the display under strong light. Many car manufacturers have designed the concept of turning the cabin into a second living room as illustrated in Fig. 31. To provide a better user experience in such a living room, the imaging quality of the in-vehicle display will play an important role.

Touch quality cannot be visually observed like display image, but users can instantly sense it through touch operation. Regarding touch quality, we can discuss it from two basic aspects. The first one is the hardware design. As an example, when the circuit grounding design is not robust, the non-uniformity of raw data on the entire touch surface could increase significantly due to the fluctuations of the reference ground level. This phenomenon could lead touch IC to misjudge the touch event and cause the ghost touch problem. “Ghost touch” is what happens when your touch screen device begins performing actions by itself. The screen seems to react to non-existent touches, or apps open without you having done anything. To reduce the ghost touch problem, the hardware design can adopt a dual-loop design (redundancy) to reduce its occurrence rate. The second one is the software design. If the above-mentioned hardware problems occur in the touch panel, the firmware coded with a correction algorithm inside the touch IC shall be able to detect and then self-correct the anomalies. In general, the improvement of touch quality cannot rely solely on hardware or software but must rely on the coherent integration of software and hardware to achieve the best outcomes.

3.8 Reliability and Durability

Different vehicle manufacturers may have different reliability requirements for in-vehicle displays. In general, DSAA specification serves as a basic requirement for display manufacturers. OEM and Tier 1 usually will require a specific reliability validation plan according to their specific design and mission profile. An example of an in-vehicle reliability requirements document (RRD) is shown below. A typical storage temperature range for vehicle is −40 °C and +90 °C, and operating temperature range is −40 °C and +85 °C. Further requirements involve moisture, humidity, and vapor pressure changes. Tests are performed with up to 95% humidity in combination with heat cycles. Wet heat soak and heat humidify cyclic tests are to ensure that display components not only can resist moisture ingression for electrochemical corrosion but also the impact of moisture and vapor pressure on sealing and optical films. The display modules must be able to pass all tests. The displays need to be fully functional without exhibiting any unallowable changes in optical performance. The sample size per test shall be designed with zero failures to demonstrate a specific reliability target with a specific confidence level. The lifetime requirement for typical automotive displays is 15 years or 10,000 h of operation and the mechanical mileage requirement is 150,000 miles. For autonomous vehicles, its mission profile usually requires 40,000–50,000 h of operation and 300,000–400,000 miles although even with its reduced vehicle lifetime of 5–6 years.

- Non-op high- and low-temperature storage:
  - Tamb = + 95 °C, 24 h, 2 cycles of a 12 h each; Tamb = −40 °C or −55 °C, 24 h, 2 cycles of a 12 h each
- Life test—high-temperature endurance test:
  - Typical requirement temp = +85 °C; 1217 h based on the following vehicle temperature mission profile

Temperature (°C)	Distribution (%)
−40	6
23	20
40	65
75	8
85	1

- Low-temperature operating:
  - Tamb = −30 °C or −40 °C, 48 h, per functional status
- Op humid heat, constant:
  - Tamb: 40 °C; RH = 95%; severity 1, 504 h
- Non-op humid heat, constant:
  - Tamb: 65 °C; RH = 95%; severity 1, 504 h
- Temperature shock (without housing):

  - -40 °C/+85 °C, 300 cycles; transfer time < 10 s

- Humid heat, cyclic:
  - −10 °C to +65 °C, RH = 93%, 6 cycles = 144 h

- Mechanical shock:
  - Directions: X, Y, and Z axes; 10 repeats per axis;
  - Peak acceleration = 50G;
  - Pulse duration = 6 ms;
  - 0.5 sine wave

- Vibration test with TC −40 °C to 85 °C or heat soak 65 °C/90%RH:
  - Directions: X, Y, and Z axes; duration: 6 × 8 h;
  - Vibration-profile D (severity 1)
  - Peak acceleration = 20G;
  - Frequency = 5–2000 Hz

- UV test/sun radiation:
  - 830 W/m^2, profile Z-IN-2, 25 days (15 days dry, 10 days humid)

- Salt mist with the operation:
  - 35 °C, 8 h spray and 4 h rest per cycle, 2 test cycles

- Power temperature cycling:
  - −40 °C/+85 °C, begin power-on for 100 s and sleep/off for the 20 s from the start of the positive thermal transition until the next negative thermal transition begins. 339 cycles.

- Impact test:
  - Ball impact is an acrylic ball with 165 mm diameter and 6.8 kg weight. Impact at 24.1 km/h or 19.3 km/h with airbag
  - Pendulum impact. Impact at 24.1 km/h or 19.3 km/h
  - Deacceleration curve: 80 g continuously for less than 3 ms

- Sweat test:
  - Apply acidic sweet, wait for 24 h at room temperature, and place DUT in 85 °C/85%RH for 196 h operational. Repeat for alkaline sweat

- Chemical ingression test:
  - Apply applicable chemicals (cleaning agents, coffee, coke, sunscreen, vaseline lotion, fruit juice, etc.) wait for 24 h at room temperature and place the device under test (DUT) in 85 °C/85%RH for 196 h operational.

- Harmful gas:
  - Flowing mixture gas (H_2S, SO_2, NO_2, Cl_2), 30 °C, 75%RH, 500 h

- Dust and water ingression test:
  - IP5k2 or IP6. Note: for water ingression, the goals are to mimic (1) water dripping due to opened doors/windows/sunroof during rainy days, (2) water spills from drinks by passengers/drivers

Failure modes are usually defined in either absolute or relative terms. If a spec is defined in absolute terms, the measurement of the parameter must always fall within the absolute limits set by the display specification. If a spec is defined in relative terms, the shift in the parameter is calculated as a percentage change relative to the value at time zero. The functional measurements and cosmetic inspection should be conducted during incoming inspection, at the end of the test, and interim read points during the entire course of reliability testing. For mechanical shock and vibe tests, no rattling noise due to internal components is allowed.

3.9 Functional Safety

In recent years, more and more electronics have been installed in a vehicle to bring a lot of conveniences for drivers. However, all in-vehicle electronic components have a safety measure called functional safety. The level of a component that ensures driver's safety by its function is regulated per automotive safety integrity level (ASIL). There are 4 levels in ASIL, from rank A to D, D is the most stringent standard. Generally, center information display is unregulated, but ASIL-B is required for cluster display systems as it provides drivers with information such as speed, rotation speed, signal warning, navigation direction, etc. Therefore, some in-vehicle displays play an extremely important role in vehicle safety. Their development process must fully comply with the ISO 26262 international vehicle functional safety standard. The life cycle of ISO26262 is shown in Fig. 32. This standard applies to automotive electronic systems, including software and hardware components, and defines the safety-related functions used in the development process, as well as the requirements for processes, methods, and tools that need to be met. The ISO 26262 standard is to avoid unexpected safety risks caused by any software or hardware abnormalities and to ensure that the vehicle meets the expected safety requirements during the life cycle.

As an example, when the display backlight fails, the content of the display cannot be read. As a consequence, the driver cannot know any driving-related information. For a less critical scenario, the driver may miss the driving direction so use more time to reach the destination. For a more severe condition, the driver may react improperly to cause a car accident. According to ISO 26262 defined as the functional safety management principle, when any one of the software and hardware fails, the other side should have a remedial mechanism to reduce safety risks. In other words, the hardware design needs to adopt a redundancy design or an on-board prognostic design to reduce the impact of failure on safety. For the redundancy design concept, we can design the light bar of the backlight module driving with two independent LED

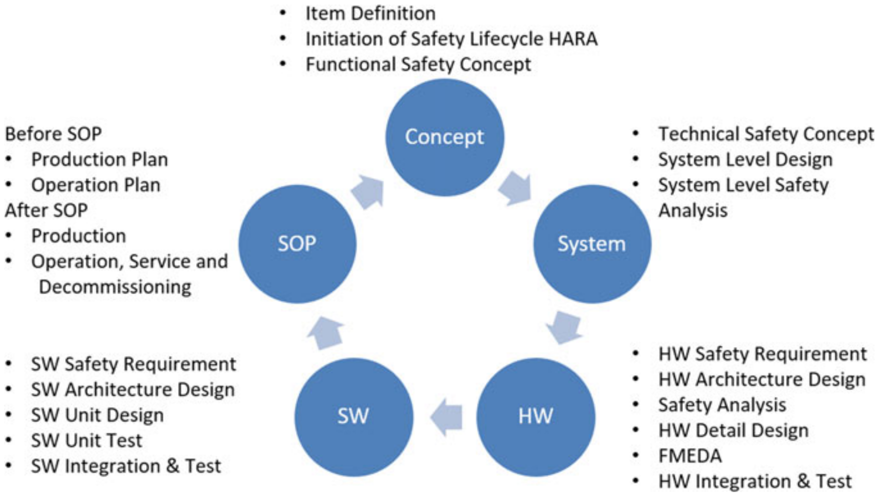


Fig. 32 ISO 26262 functional safety management and safety analysis

drivers. When one of them fails, the display can still maintain a certain brightness, so that driver still can read the information content. For the on-board prognostic concept, a display with a self-failure diagnostic and repair function can be implemented per ASIL-B. As shown in Fig. 33, Panasonic Corporation developed a failure detection display. Liquid crystal has a feedback electrode to confirm electro-conduction for each line, and scan all lines by the confirmation signal from the driver IC in constant time. By doing so, when there are electrical defects such as disconnecting or an electrical short circuit, a warning signal will be displayed.

4 In-Vehicle Display Challenges

Although display technology has been advancing substantially in recent years, there are still quite a few display challenges in practical applications for vehicle usage. The automotive environment imposes more stringent requirements on reliability and product lifetime than any consumer and industrial environment. In addition, strongly varying ambient lighting situations together with the inability of the occupants to freely position the display puts higher demands on designing the display for readability in high ambient lighting conditions. In-vehicle display challenges include how to address the unique specification and functionality of vehicle display, reliability and validation, EMC/EMI, ESD, and high-transient voltage challenges. In general, the scale of these challenges depends on car manufacturers' requirements for specifications and functions, field use scenarios, and relevant regulatory/compliance requirements. To meet these challenges, in-vehicle displays have to adopt different designs

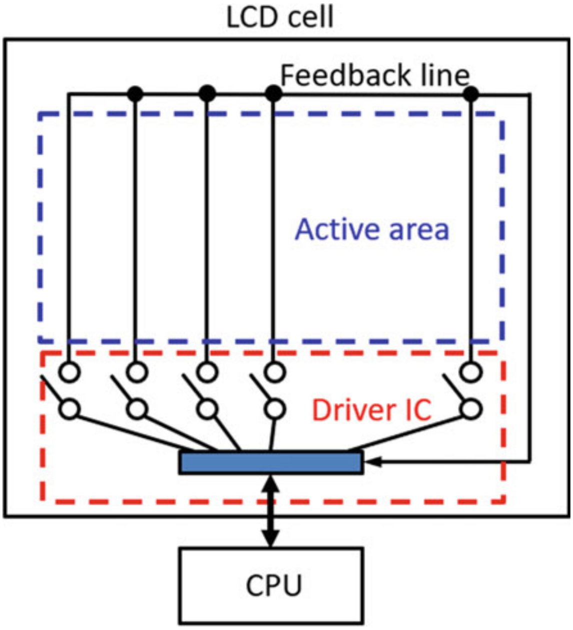


Fig. 33 A self-defect diagnostic display

from the consumer displays. Therefore, it is well known that making displays for vehicle usage has a higher technical threshold than making displays for consumer usage.

4.1 *Specification and Functionality Challenges*

With the improvement of automotive display technology, current in-vehicle displays have been able to meet most of the automotive display specifications and the requirements defined by OEMs. However, there are still certain difficulties for some use cases.

a. **Ultra-Low Reflectance (ULR)**

Due to the increase of brightness together with browless dashboard design for in-vehicle display, optical reflection management becomes more challenging and more important. Windshield reflection, especially unsafe nighttime reflection, can distract drivers to cause serious safety concerns. Without some special engineering methods, the reflectance of LCDs could be as high as 10–15%. The current automotive display specification requirements for ULR are.

System without TP: Rdi < 1%; Rde > 0.3*Rdi
System with capacitive TP: Rdi < 1.7%; Rde > 0.3*Rdi.

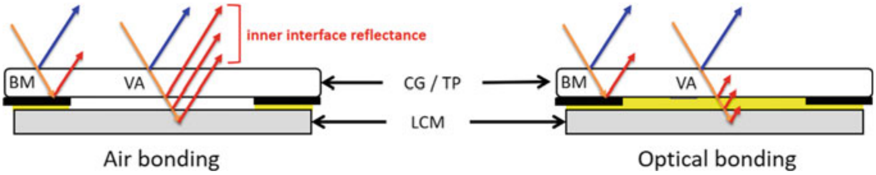


Fig. 34 Air bonding versus optical bonding with different internal reflections

Where Rdi is reflectance measured for specular include, and Rde is reflectance measured for specular excluded.

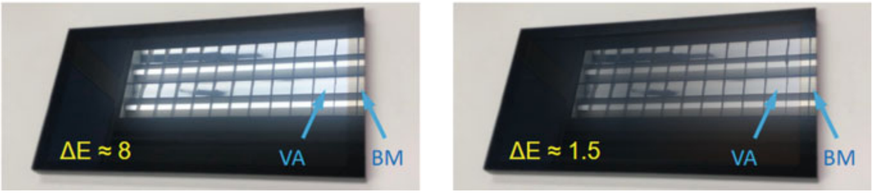
Based on the above automotive display reflectance specifications, there is a need to reduce all the reflection sources such as surface reflection, interface reflection, and internal reflection to achieve the required reflection specifications. Multiple solutions are available to balance brightness and windshield reflectivity. Using light control film (LCF) to minimize light leakage at high-vertical angles (e.g., beyond 35 degrees) is one of the popular approaches. The working principle of LCF is based on louvers, which allow transmission in a perpendicular direction but block light under large angles. To reduce the interface reflection such as between cover glass and LCD, an optically clear light control film can be used to fill the air gap between them to reduce the characteristic glare. For surface reflection, using an anti-reflecting coating on cover glass is common. A super low-reflection surface treatment technology developed by Panasonic was reported to reduce the diffusion reflection under a 0.1% reflection ratio. For the internal reflection between TP and LCD, a reflection light canceling technology such as circular polarization can be used. Alternatively, a full bonding process (optical bonding) to reduce the internal interface reflection of the viewing area also helps as shown in Fig. 34. Even with all the light control techniques, it is still very difficult to achieve the ULR specification, especially for the out-cell and on-cell touch displays as the external TP circuit layer will increase the reflectivity. Therefore, more and more Tier 1 adopts an in-cell display by placing the TP traces inside the display to reduce reflectivity.

b. **Seamless Design**

At present, the specifications of the seamless design for all OEMs are between ΔE≦1.5 ~ 2.0, which is defined as when the display backlight is not lit, the reflected brightness and chromaticity differences between the active viewing area and the black ink printing black matrix (BM) area as shown in Fig. 35. When the difference is smaller, the user is less likely to perceive the boundary between the two areas, and then, a uniform black can be visually achieved. The calculation formula is as follows:

$$\Delta E_{ab}^{*} = \sqrt{\left(L_2^{*} - L_1^{*}\right)^2 + \left(a_2^{*} - a_1^{*}\right)^2 + \left(b_2^{*} - b_1^{*}\right)^2} \qquad (3)$$

It can be seen from the formula that important parameters include reflectance L* and chromaticity coordinates a*/b*. Therefore, it is necessary to take into account the



Optical bonding visual performance

Optical bonding with seamless design

Fig. 35 Optical bonding versus optical bonding with seamless design visual performance comparison

optical matching of reflectance and chromaticity simultaneously to make the seamless black reach the expected target value. Most Tier 1 and panel manufacturers have invested considerable resources in this field to develop their know-how techniques. The common ways are.

a. The full bonding process is used in the structure stacking to reduce the interface reflectivity of the viewing area (VA). When the refractive index of the two objects is closer to the interface reflectivity, the lower the interface reflectivity (Snell's law).
b. Use smokey material on the top of the display to reduce the transmittance of VA. But the side effect is it will reduce the transmittance of the display system, therefore reducing the brightness of the display screen.

c. **Fast Response Time**

The response time of the LCD is determined by the speed of pixels responding to the input voltage signals. For in-vehicle display, fast response time under various temperature conditions becomes a safety-relevant aspect, especially for dashboard clustering information display, rear camera display, and digital mirror including central mirror and side mirror since real-time content needs to be displayed. According to Eq. 4, there are four ways to improve the response time (t_{on}) of LCD including decreasing the viscosity coefficient of liquid crystal, decreasing the cell gap, increasing the driving voltage, and increasing the dielectric constant of liquid crystal.

$$t_{on} = \frac{\gamma}{\varepsilon_0 \Delta\varepsilon E^2 - \frac{\pi^2 K}{d^2}} \tag{4}$$

where γ is the viscosity coefficient of the liquid crystal, K is the torsional elasticity coefficient, ε is the dielectric constant, E is the voltage applied to the liquid crystal, and d is the cell gap. X, Li et al. simulated response time relationship with γ and d. They found that when the γ was decreased by 10 mpas, the t_{on} would be improved by 11%. When the cell gap was decreased by 0.2um, the t_{on} would also be improved by 11% [19]. As LCD technology switching time is highly temperature-dependent, from Eq. 4, the LCD with a lower dielectric constant would have a faster response time, especially under low-temperature operations (400 ms@−30 °C). OLED technology has a temperature stable switching time smaller than 2 ms as compared to LCD technology.

4.2 Quality, Reliability, and Validation Challenges

Automotive TFT LCD has been developed for a long time. Suppliers already have developed automotive-grade components sufficient to withstand the stringent reliability requirements. However, for emerging technologies such as OLEDs, their organic material reliability and durability are still questionable. To trade-off between 3D curved displays or free from display designs, automakers may have to tailor the reliability specifications to meet the current technology of OLEDs. Although the reliability of OLED has been improved recently, there is still a big gap between OLED versus TFT LCD. Therefore, TFT LCD likely will still be the mainstream product of car display soon. Regarding TFT LCD reliability, although it has been studied for many years, there are still some challenges.

It is required that all automotive display suppliers have to follow IATF-16949 for development and process. The display panel vendors need to adopt some special quality control methods in the manufacturing process such as dedicated assembly line, ACF bonding quality inspection, vibration test, and burn-in test. Among them, bonding quality is a key process that determines the reliability of automotive displays. Usually, display panel manufacturers will confirm the pressing state and the effective number of anisotropic conductive film (ACF) particles to ensure that the circuit can have good electrical conduction. Different bonding pressure will cause different ACF particles deformation shapes and reliable results. If the bonding pressure is not optimized, a result as shown in Fig. 36, the ACF particles will not have proper ball breakage shapes, therefore the effective number of conductive particles could be too small. The electrical resistance and its reliability performance have a strong dependence on the effective number of ACF conductive particles as shown in Fig. 37. Although the display can still light up normally during short-term testing or usage in some cases, the ACF electrical resistance can increase after a period of use due to high-temperature and high-humidity environment, which will cause abnormal images during lifetime usage. The display panel manufacturers will count the effective ACF number (well deformed and show a proper ball breakage shape) to judge its bonding quality for automotive application.

ACF Particle Breakage

One unique validation test for automotive displays is the headform impact test (HIT). HIT is a very important safety-related, regulatory verification test item where cover glass performance and system-level architecture play important roles. The purpose of this test is to confirm whether the infotainment system will cause safety concerns such as the splash of glass fragments and sharp broken glass edges to cause physical injuries to the occupants in the event of a vehicle frontal collision. The HIT testing

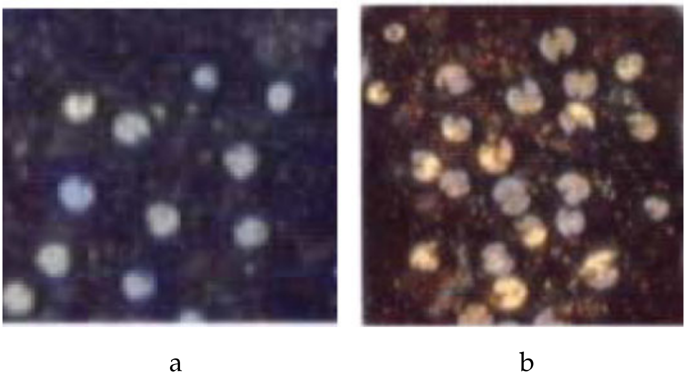


Fig. 36 **a** Low-bonding pressure so ACF particle breakage is not complete, **b** a good bonding pressure to induce a good

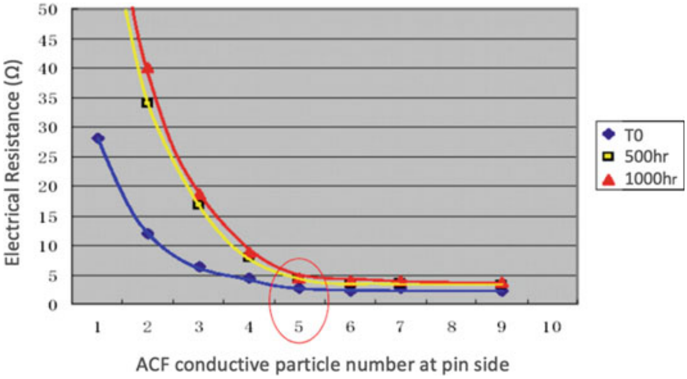


Fig. 37 ACF electric resistance and its time-dependent behavior versus its conductive particle number [20]

specifications are defined by FMVSS201 in the USA, TRIAS34 in Japan, ADR21/00 in Australia, GB 11552 in China, and ECE R21 in the EU. The test configurations and requirements are similar to those standards, where the total headform deceleration must not exceed 80G for longer than a 3 ms duration. The head form, an impact object of the test, has a mass of 6.8 kg and a diameter of 165 mm, which represents a human head. The impact velocity varies from 19.0 to 24.1 km/h depending on the market. The resulting kinetic energy at impact is 94 J and 152 J, respectively. Although no clear pass/fail is specifically defined in those standards, in general, the auto industry desires no glass breakage during HIT. HIT is a destructive test. Ideally, it should be tested at the system or vehicle level with the car body as the outcomes depend on not only a cover glass but also supporting structures. Therefore, conducting HIT tests could be costly and time-consuming. For Tier 1 suppliers, it is not easy to obtain the car body or conduct tests at the vehicle level. Alternatively, Tier 1 suppliers sometimes rely on making a surrogate fixture to accommodate the HIT in-vehicle condition and also rely on using computer-aided engineering (CAE) in the early design phase to assess the risks. Layouni et al. [21] developed a HIT testing setup at the panel-level to represent aggressive biaxial bending cases. Their surrogate setup has higher deceleration than the regulation limit. The CAE finite element modeling simulation can help to better understand the kinematics at a crash and discover the strength deficiencies in the display mounting structural design to enable mitigations. An example is shown in Fig. 38. In general, the lack of extensive reliability testing data makes it difficult to accumulate HIT design verification experiences. Different vehicle models have different appearances and therefore different requirements for mounting positions of the display. The appearance of the in-vehicle is important as it makes the displays achieve a working harmony with vehicle interiors. Therefore, Tier 1 and panel suppliers usually have to develop customized display designs combining appearance, performance, and touch function to accommodate different vehicle appearance requirements, which once again increases the difficulty of HIT design to meet regulatory requirements. As the vehicle display module is located at the top surface of the infotainment system, applying a glass strengthening technique to improve cover glass fracture toughness is mandatory. Engineering glass structures through chemical and microstructure improvements, reducing surface and bulk flaw densities, and introducing surface compressive stress all can be adopted. As glass is a brittle material, it breaks before the yield. This means that glass breaks instantly when the stress is above the yield defined by the material constants and defects (size and density). This also suggests that the glass strength is a statistical parameter. For the HIT, a large sample size usually is required to demonstrate its impact resistance reliability at a higher confidence level. Furthermore, it was found that reducing the stress concentration point from product integration dashboard design is also critical. In general, improving cover glass strength and dashboard system design are the keys to the success of passing the HIT.

The location design of the infotainment system is different. It can be divided into in-dash and on-dash as illustrated in Fig. 39. The main difference in the structural design of the two is that the in-dash display is located inside the dashboard, while the on-dash display is extended above the dashboard. On-dash must be designed to take

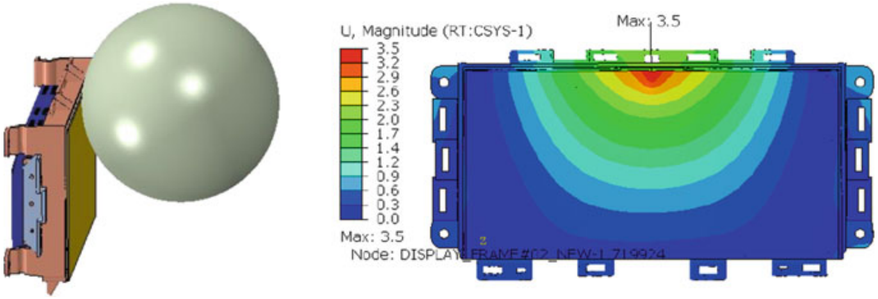


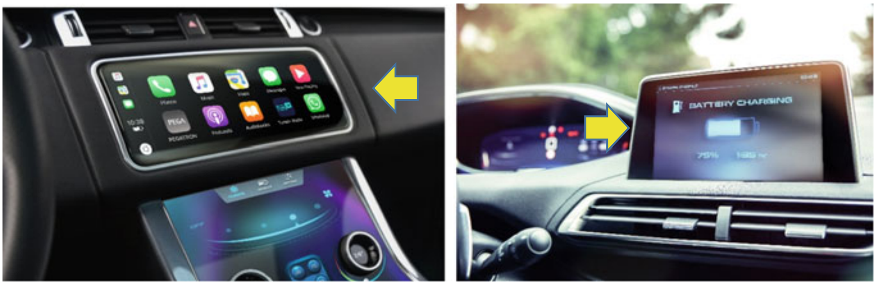
Fig. 38 CAE FE simulation of HIT point and its corresponding deformation

into account the fixedness of the whole display after head impact to avoid causing a second injury to the bodies of occupants.

The test conditions of HIT are listed in Table 4 based on the ECE R21 standard. The resultant acceleration at the center of gravity of the dummy head must be such that the expression shown in Eq. 5. The key is that the deceleration curve should not exceed 3 ms after the vehicle has passed 80G to avoid excessive impact on the human body. In terms of buffer mechanism stiffness design, if the deformation of the whole display is too small to absorb the impact, it will easily cause the G value to rise, and to fail the requirement of less than 3 ms. If the deformation of the whole display that can absorb the impact is too large, it will easily cause the support structure to deform or break. Then, the cover glass could severely crack, which may form sharp edges to induce physical injuries of the occupants. How to balance them is a challenge for in-vehicle display system design.

$$\mathrm{HIT} = \left[\frac{1}{t2 - t1} \int_{t1}^{t2} \mathrm{adt} \right]^{2.5} (t2 - t1) \qquad (5)$$

From the extreme environmental usage conditions side, it is a challenge to make an in-vehicle touch display work with full functionality for extremely high and low temperatures such as 85 °C and −40 °C. Measurements show that the on-current of TFTs changes substantially with temperatures due to the changes in mobility and



a
b

Fig. 39 **a** In-dash infotainment system and **b** on-dash infotainment system

Table 4 HIT test conditions

Regulation	ECE R21
Coverage	For Europe
Test head	165 mm
	6.8 kg
Test speed	24.1 km/h (with Airbag: 19.3 km/h)
Deceleration curve	80 G/less than 3 ms

threshold voltage. For in-vehicle display, it seems cold temperature wake up, and operating is more unique. Yang et al. [22] proposed an integrated gate driver circuit using a separated transfer electrode and the output electrode driving TFTs with two independent bootstrapping nodes to improve display reliability under an extremely cold environment. The other environmental challenges are large temperature swings, vibration frequencies, humidity changes, and variations of atmospheric contaminants such as H_2S and COS.

4.3 EMC/EMI Challenges

Due to the widespread adoption of electronic components in automotive electronic modules, especially with the booming development of electric vehicles, the proportion of electronic components in a vehicle is getting more and more. It is well known that electromagnetic compatibility (EMC) impacts the integrity of sensors, data transmission, control systems, navigation equipment, etc., which can result in driving safety concerns. This issue can be exacerbated with the rapid deployment of 5G networks. Therefore, EMC has become a very important topic in automotive electronic design. EMC includes:

1. Electromagnetic Interference (EMI)—component cannot produce electromagnetic interference to other equipment. In general, the EMI test can be divided into radiated emissions test and conducted emissions test according to the characteristics of the disturbance source.
2. Electromagnetic Susceptibility (EMS)—component can resist electromagnetic interference from other equipment. In general, EMS testing is (1) conductivity tolerance test method; (2) radiation tolerance test method; (3) conducted transient withstand test method; (4) electrostatic discharge ESD test method.

The internationally accepted automotive EMC regulations include ECE R10 regulated by the United Nations Economic Commission for Europe (ECE), 97/24/EEC and 95/54/EEC regulated by the European Union, and CISPR (French: Comité International Spécial des Perturbations Radioélectriques), Society of Automotive Engineers (SAE), Japanese Automobile Standards Organization (JASO), and ISO. Generally speaking, EMC/EMI is tested according to customer requirements and specifications in the state of the whole system. However, for automotive displays, the above-mentioned regulations can still be used to verify the display module. For EMC testing laboratories, the major U.S. automakers have requested that EMC testing of all components must be performed in a laboratory accredited by automotive EMC laboratory recognition program (AEMCLRP). An example of an EMC testing setup in a lab is shown in Fig. 40. There are two-stage procedures for obtaining AEMCLRP certification: (1) obtain an A2LA certificate according to the AEMCLRP procedure; (2) submit test data to the major US automakers such as GM and FORD for approval.

For display EMC, the main consideration is radiated emissions (RE) protection methods. The periphery of the display screen is generally composed of a metal frame

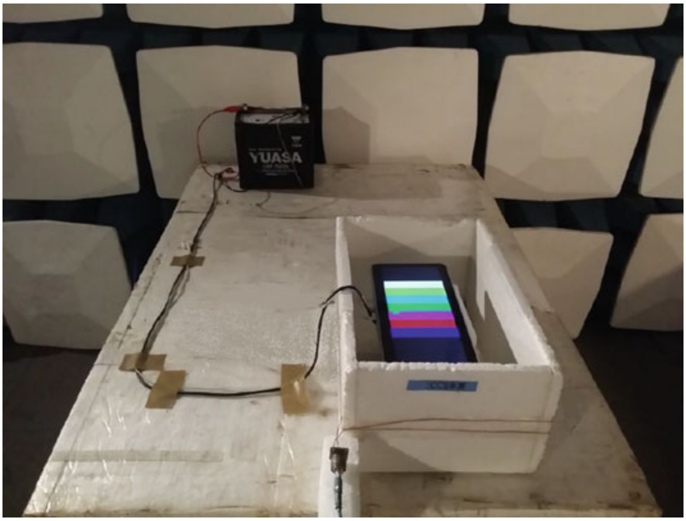


Fig. 40 EMC testing lab EMC test setup illustration

and a metal backplate. In the design, the metal frame and the metal backplate should be effectively connected to the same ground. The floating metal parts without a ground loop will cause static electricity to run away and cause the display to show flickers, black screens, and garbage colors. RE is a difficult test for display components. Common mitigation methods are.

1. Spread-spectrum clocking—spread spectrum technology is a commonly used radio communication technology. It is divided into two ways: frequency hopping technology and direct sequence. Frequency hopping technology has been used for in-vehicle display mostly. The main clock of the low-voltage-differential-signaling (LVDS) communication signal on the display screen is usually between 50 and 100 MHz. The peak energy of the pulse could be strong, which will cause space electromagnetic interference. The spread spectrum technology reduces the electromagnetic interference generated by pulse clipping. Therefore, the electromagnetic intensity of the peak point could be evenly distributed to the surrounding frequency points.
2. Differential impedance matching—is mainly used on LVDS signal transmission lines so that high-frequency limit signal can be transmitted to the load point, and no signal will be reflected to the source point. The internal resistance of the signal source is equal to the characteristic impedance of the connected transmission line. The characteristic impedance of the transmission line is equal to the impedance of the connected load. There are many interconnects between the LVDS communication signal of the display screen and the host display processing chip. Therefore, only the matching impedance can reduce the excessive energy in the connection process to minimize the excessive radiation.

3. Mirror ground—the mirrored ground is a layer of copper-clad plane adjacent to the signal layer. The main functions are to reduce the return noise, reduce the radiation intensity, and prevent signal reflection. The mirror ground is used in many ways in the in-vehicle display design. For example, the display FPC cable has two layers, a signal layer and a layer of copper as a mirror ground. The bottom space of the LVDS communication line on the PCB is filled with conductive film as mirror ground.
4. Shielding cover—add shielding cover to some components with relatively high-radiation interference intensity. The basic principle is to use a complete metal shield to seal the components showing relatively large electromagnetic interference in the near-field RE test. Therefore, the external electromagnetic field strength is lower than the allowable value. It is also possible to seal the electromagnetic sensitive circuit with a wave-absorbing material so that the internal electromagnetic field strength is lower than the allowable value.

In summary, EMC protection needs to be carefully carried out during the design of the in-vehicle display system. By utilizing the platform-based EMC design rules, it can be shown that EMC protection can make displays and various other electrical equipment co-operate in the car but without interference with each other to cause function and performance degradations.

4.4 ESD and High-Transient Voltage Challenges

Electro static discharge (ESD) refers to the transfer of unbalanced charges on the surface of an object. When the charge voltage difference is higher than a certain level, the insulating medium will undergo an electrical breakdown process, which will cause a localized conductive path to form inside the insulating medium. Such localized conductive path can induce high-current passing through. The main destructive force of electrostatic discharge is the thermal effect from the instantaneous peak current, which can easily cause the electronic components to be broken down or burned and then cause malfunction of the entire electronic system. The malfunction could be either partial vehicle function loss or a complete automobile failure.

For safety concerns in automotive electronics, automotive ESD compliance standards have higher voltage test limits than commercial electronics. Unique test setups simulating a car chassis are typically required. Common automotive ESD standards are ISO 10605, GMW3100/3091, FMC 1278 CI 280, ISO 14304, SAE J1113/13, SAE J1455 A94, VW 80972, JASO D001-94, IEC 61000-4-2 etc. In terms of ESD testing, verification is mainly conducted by customer requirements and international specifications. The vehicle display will be validated in the following two discharge modes to confirm the ESD tolerance of the driver IC.

a. Human-body model (HBM)
b. Machine model (MM).

Table 5 shows the ESD test level for SAEJ1113/13. All tests are to be performed in positive and negative polarities.

The area of damage caused by ESD could be very small so it is not easy to be found on the unit under test (UUT) surface just by visual inspection. Only the lower layer structure may be damaged while the surface is still intact. Therefore, ESD damage usually requires additional equipment such as a low-light microscope and light emission microscopy to assist in locating the damage point. The fault level can be divided into four levels: A, B, C, and D.

Level A: The system functions normally after the ESD test.

Level B: After the ESD test, the system will crash, but the system can automatically recover to its original function after reset.

Level C: The system will crash after the ESD test, but the system can be recovered to its original function after a manual reset.

Level D: System function failure after ESD test.

ESD cannot be 100% prevented because too many external static power sources exist in our vehicle system. We can only do our best to minimize the accumulation of electrostatic charge to reduce the occurrence of ESD events. The following design prevention countermeasures can be taken into in-vehicle display design:

a. Absorption: exposed copper
b. Discharge: add conductive material to discharge
c. Blocking: add protective pads and anti-static components.

In the design of the in-vehicle display and touch panel, an ESD ring is generally added at the outermost edge to the ground as shown in Fig. 41. When there is a high

Table 5 ESD test level

Level	Category 1 (kV)	Category 2 (kV)	Category 3 (kV)
Direct contact discharge			
1	2	2	5
2	4	4	6
3	6	8	8
4	8	8	15
Indirect contact discharge			
1	2	2	4
2	4	4	8
3	6	8	15
4	8	15	20
Direct air discharge			
1	2	4	6
2	4	6	8
3	8	8	15
4	15	15	25

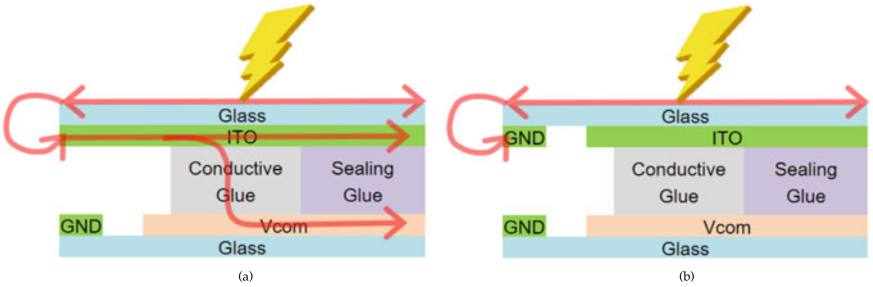


Fig. 41 Working mechanism of the ESD ring for in-vehicle display. **a** Without an ESD ring, charges are transported into the internal circuit, and **b** with an ESD ring, charges are prohibited to transport into the internal circuit

amount of static electricity on the surface of the display, it can prevent the surface charge from transporting into the internal circuit, and then destroy the internal driver IC. In addition, ESD protection components such as transient voltage suppression (TVS) diode shall also be added to the circuit design.

5 Common LED LCD Reliability Testing Failure Modes and Effects Case Studies

As LED LCD seems to be the mainstream of in-vehicle display technology currently, in this chapter, some common LED LCD reliability failure modes and how to model them for accurate field risk assessments will be discussed. We are going to demonstrate how to use a technically sound reliability modeling methodology which consists of a continuous probabilistic model, physics of failure acceleration model, parameter degradation model, and realistic mission profile to achieve some best estimates of LCD reliability.

5.1 FOS Spotlighting Failure Mechanism and Risk Assessment

During the display system-level high-temperature high-humidity (HTHH) reliability test, front of a screen (FOS) spotlighting failure was observed as shown in Fig. 42. This issue only occurred at the display system-level (with full system enclosure) test, panel-level, and LED light bar level tests did not reveal a similar issue. Under the same temperature and humidity level, a non-operating test was worse than the operating test. The failure rate was found to be scaled by testing temperature and humidity levels. Interestingly, wintertime test results were worse than other seasons' results although the display panels were kept the same.

Failure analysis (FA) found that the root cause of spotlighting was due to degraded brightness or no light of some LEDs. All of the problematic LEDs exhibited discolored encapsulant/housing. Some discolored LEDs were out of V_f specification. However, some LEDs with slight discoloration passed V_f specification but at low current. Silver (Ag) and copper (Cu) were detected in the discolored area. A small amount of sulfur was also detected on the failed LEDs. Watermarks were found at the LED housing. Nearly, all failed LEDs were the LEDs at the high-voltage end of the light bar as shown in Fig. 43. So based on FA findings, we can infer the failure mechanism as Ag plated Cu lead frame electrochemical migration (ECM) corrosion accelerated by the presence of sulfur. During the corrosion process, Ag+ ions migrated from anode to cathode (M-chassis) forming a short path. Figure 44 illustrates the physics of failure with all the potential contributing factors for LED spotlighting failure during the HTHH test.

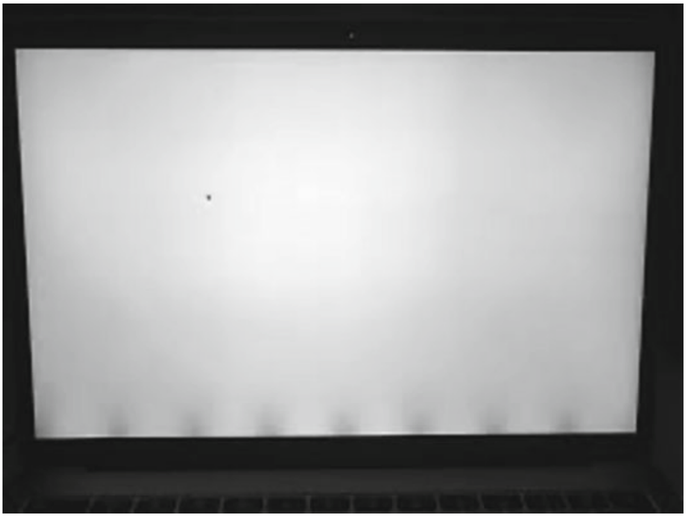


Fig. 42 Display showing spot lighting

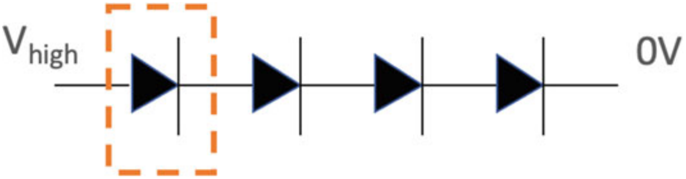


Fig. 43 LED light bar voltage connection illustration

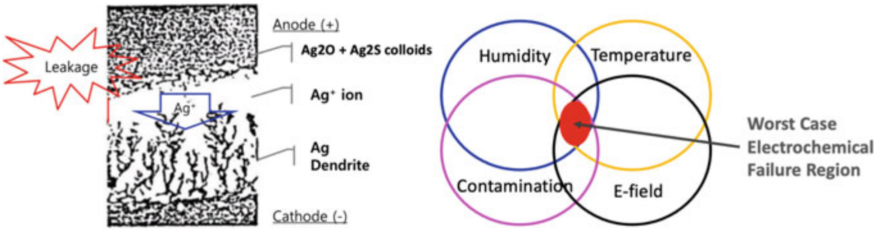


Fig. 44 Silver electrochemical migration physics of failure model

To explain why only system-level tests exhibited Ag ECM while panel level and LED lightbar level tests did not, we proposed a capillary condensation theory as illustrated in Fig. 45. Since the system enclosure offered a confined space, under a high-humidity environment, the distances between vapor phase particles were very small. This increases the van der Waals forces of attraction between particles. Hence, less pressure is required for condensation inside a capillary to occur. This theory was supported by the evidence of watermarks found inside the system housing. Based on Fig. 46, we also could explain why the non-operating test was worse than the operating test. LED operating would generate a significant amount of self-healing. Such heating would expel moisture out from the system instead of allowing it in, therefore the internal moisture level during display operation was limited. On the contrary, during the non-operating stage, external moisture could easily ingress into the display system to be absorbed by the air cavity around LEDs and LED polymers. The worst-case corrosion risk window was when meniscus had formed during system cooling down from a long-term non-operating wet heat soak, while LED was starting to operate for an electrical check as illustrated in Fig. 46.

Mathematically, we developed a two-step ECM corrosion model to quantitatively model the LED dark spot time-to-failure. The first step is the moisture absorption step. The worst case of this step was under-display non-operating mode. Water absorption depended strongly on the exposure temperature, relative humidity level, time, and material properties described by Eq. 6:

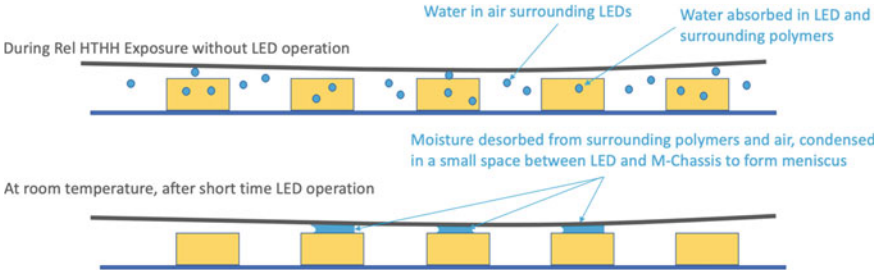

Water in air surrounding LEDs
Water absorbed in LED and surrounding polymers
During Rel HTHH Exposure without LED operation
Moisture desorbed from surrounding polymers and air, condensed in a small space between LED and M-Chassis to form meniscus
At room temperature, after short time LED operation

Fig. 45 Capillary condensation model to explain water formation inside display system

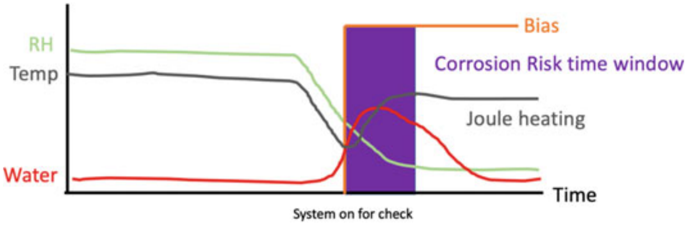


Fig. 46 LED dark spot risk time window illustration

$$M(T, \mathrm{RH}, t) = M_{\max}\left[1 - c\left(\frac{\mathrm{Dt}}{h^2}\right)^{0.5}\right] \tag{6}$$

$$D = D_0 \exp\left[-\frac{\mathrm{Ea}}{kT}\right]$$

where M is the moisture content at any given point in time (t), $M_{\max}$ is the maximum moisture content at a fixed temperature and RH level, c is a constant representing material properties, D is the diffusivity of the material through the thickness, h is the thickness of the material, and Ea is the activation energy of moisture diffusion. $M_{\max}$ can be experimentally obtained by a water weight gain experiment. The second step is LED lead frame metal corrosion under electric field and meniscus formation. This step could be further catalyzed by the presence of sulfur during LED on stage. The time-to-fail (TTF) can be described by Eq. 7 as follows.

$$\mathrm{TTF} = \frac{Q_{\mathrm{crit}}}{J_{\mathrm{ion}}} = A \times c^{-w} \times \frac{1}{I_S(e^{\frac{V}{nV_T}} - 1)} \times M^{-n} \times \mathrm{Exp}\left[\frac{Q}{k_B T}\right] \tag{7}$$

where Q is the critical total ion charge to trigger a dark spot fail, J is the ion flux, A is a constant, Is is LED reverse bias saturation current, V is the applied voltage, V_T is the thermal voltage, Q is the Ag/Cu corrosion activation energy, n is Peck's moisture exponent, c is contamination level, and w is the contamination exponent. After a comprehensive sulfur source investigation, it was found that it came from our testing atmospheric environment. It was well known that atmospheric contaminants such as H_2S and carbonyl sulfide (COS) promote the corrosion of silver and copper. When silver is exposed to an environment containing a minimum concentration of these contaminants (<1 ppm), the corrosion product of Ag_2S can be formed. A similar process of sulfidation occurs for copper to produce Cu_2S. As shown in Fig. 47, our observed failure rate at a fixed 65 °C/90%RH HTHH test condition scaled with air quality index (AQI) number very well, confirming our atmospheric corrosion conclusion. Furthermore, it was found that air pollution in our specific testing location showed a clear seasonal dependence pattern based on over 10 years of AQI data analysis. The winter season was the worst season for air pollution, and January was the worst month over a year for air pollution as shown in Fig. 48.

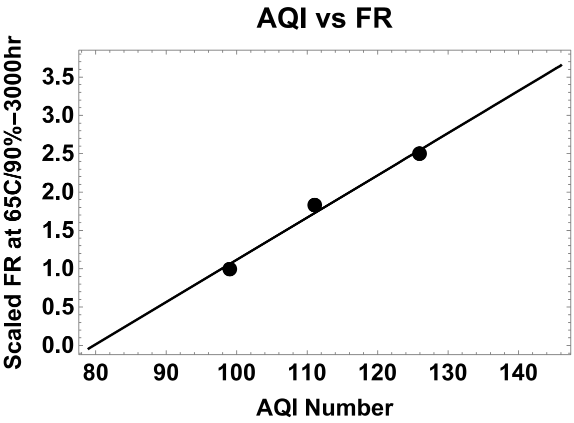


Fig. 47 FOS spotlighting failure rate versus AQI number

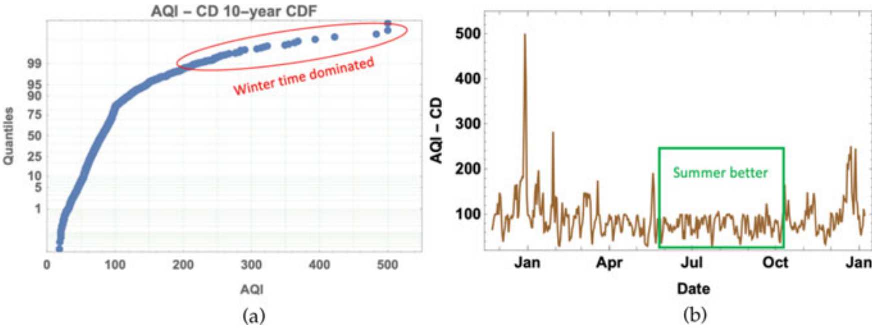


Fig. 48 AQI data analysis. **a** CDF analysis of over 10-year AQI data, high-percentile worse AQI data mostly from wintertime. **b** AQI yearly data to show a seasonal dependence

After the physics of failure was completely understood, next, we designed a specific display power cycling DoE to conduct a failure rate projection for a worst-case user use scenario. The new DoE was developed based on a 2-step ECM corrosion model as shown in Fig. 49. Different temperatures, humidity levels, and duty factors were implemented in the DoE. Figure 50 shows FOS spotlighting time-to-failure (TTF) distribution under a fixed humidity and duty factor test condition. The thermal activation was extrapolated to be 0.95 eV, which was consistent with the published Al corrosion process.

After all the critical acceleration factors were determined experimentally, failure rate and therefore risk assessment could be obtained. Realistically, the field sulfur contamination effect was significant only during half of the year per historical AQI data. Failure was generated during a short period of screen-on time in a high-humidity environment after a long heat soak. Effective moisture content was different at different temperatures and humidity levels. By considering all the influencing

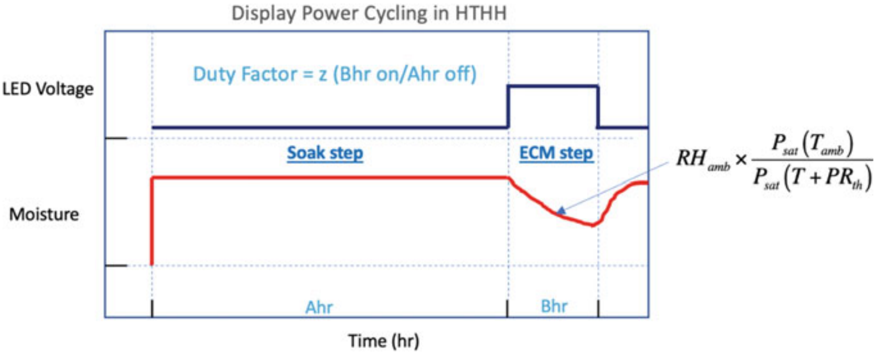

Display Power Cycling in HTHH
LED Voltage
Duty Factor = z (Bhr on/Ahr off)
Soak step
ECM step
Moisture
Ahr
Bhr
Time (hr)
$RH_{amb} \times \frac{P_{sat}(T_{amb})}{P_{sat}(T+PR_{th})}$

Fig. 49 Illustration of display power cycling DoE in high-temperature high-humidity (HTHH) environments. The moisture level was reduced during the ECM step due to self-heating according to the equation listed in the figure

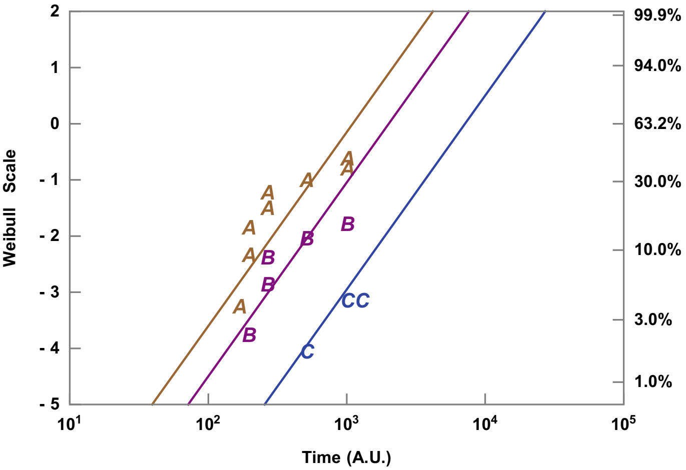


Fig. 50 FOS spotlighting TTF Weibull distribution from A, B, and C three different test conditions

factors, a reasonably conservative mission profile for a worst-case use condition could be established. Based on Eq. 7 and Weibull statistics of TTF analysis from our DoE, a failure rate of FOS spotlighting could be estimated, and its field risk could be assessed. New LED component design, new display system frame design, and improved process control of better LED to chassis gap uniformity were implemented to successfully mitigate FOS spotlighting failure.

5.2 BLU Film Buckling/Waving/Wrinkle Failure Mechanism Study

Display BLU film buckling/waving/wrinkle issues are commonly observed (Fig. 51) during display reliability testing driven by mechanical instability. Winkles occur when the residual stresses exceed a critical value, which can be induced by heating (coefficient thermal expansion driven), moisture absorption (coefficient hygroscopic expansion driven), mechanical stretching/compression, residual stress, and capillarity from bilayer thin films of dissimilar mechanical properties. BLU film stack usually consists of a light guide panel or film (LGP or LGF), ESR, diffuser, and prism. All of them can undergo dimensional changes during high-temperature and high-humidity reliability testing. Such changes are different between adjacent films and are additive. If generated dimensional changes are constrained by something such as metal chassis and display FOS wrinkle could occur.

The LGP or LGF are made from either PMMA or polycarbonate. It can undergo expansion at elevated temperatures and humidity. Bucking will occur when critical buckling strain is achieved. If the expansion of the LGP is much greater than the chassis, the edges of the LGP may interfere with the chassis leading to in-plane compressive stress that exceeds its critical buckling strain, and subsequent LGP buckling. The thermal buckle and hygroscopic buckle resistance to the critical stress can be calculated by Eqs. 8 and 9, independently.

$$\sigma_{\text{cr-CTE}} = \frac{k\pi^2 E}{12\left(1-\nu^2\right)(b/t)^2} = \frac{\alpha \mathrm{ET}_{\text{cr}}}{2} \tag{8}$$

$$\sigma_{\text{cr-CHE}} = \frac{k\pi^2 E}{12\left(1-\nu^2\right)(b/t)^2} = \frac{\beta \mathrm{ET}_{\text{cr}}}{2} \tag{9}$$

where k is buckling coefficient, E is the elasticity modulus, ν is Poisson's ratio, b is the width of a specimen, t is the thickness of a specimen, T_{cr} is critical temperature, α is coefficient of thermal expansion (CTE), and β is coefficient of hygroscopic

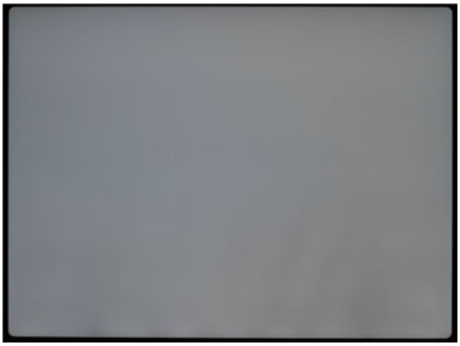


Fig. 51 Display with FOS sheet wrinkle issue

expansion (CHE). It was found that CTE and CHE were intercorrelated as shown in Fig. 52. It was also found that LGP or LGF longitude and latitude CHEs are different, and latitude CHE was about 3× higher than longitude CHE. To prevent LGP buckling under Rel testing conditions, the additive CTE and CHE induced dimension changes need to be incorporated in the design of the gap budget between chassis and LGP or LGF. A rigid gap design reliability rule for LGP or LGF buckling prevention was developed as shown in Eqs. 10 and 11.

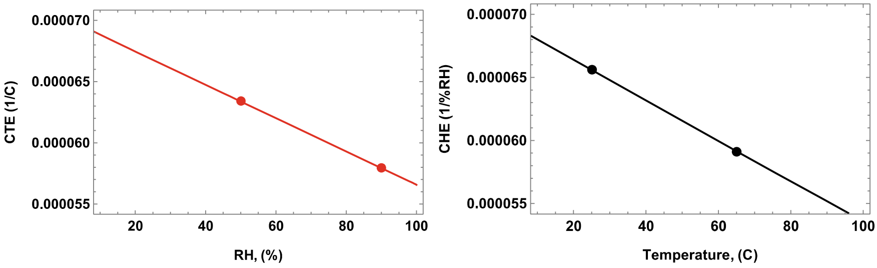
$$L_{\text{gap}} = \{(M_L - \theta_L)(\alpha \times \mathrm{dT} + 1) - (L_L + \lambda_L)(f(\mathrm{dT}, \mathrm{dRH}) + 1)\}/2 \tag{10}$$

$$W_{\text{gap}} = \{(M_w - \theta_w)(\alpha \times \mathrm{dT} + 1) - (L_w + \lambda_w)(f(\mathrm{dT}, \mathrm{dRH}) + 1)\}/2 \tag{11}$$

where

M_{L}: nominal M-chassis length
θ_{L}: 3 sigma tolerances of M-chassis length
L_L: nominal LGP length
λ_L: 3 sigma tolerances of LGP length
dT: temperature change
dRH: relative humidity level change
$\mathrm{M_w}$: nominal M-chassis width
θ_{w}: 3 sigma tolerances of M-chassis width
$\mathrm{L_w}$: nominal LGP width
λ_{w}: 3 sigma tolerances of LGP width
f(dT, dRH) is determined experimentally from LGP thermal and hygroscopic expansion DoE

The above equations assumed that the chassis or frame has negligible expansion from moisture absorption, which in general is true. Knowledge of (dT, dRH) would enable meaningful tolerance stack-up analysis to verify if, by design, enough gap margin exists to allow unrestricted expansion of the LGP given possible extreme field conditions, and "within spec" manufacturing variations. Similar to LGP, it was

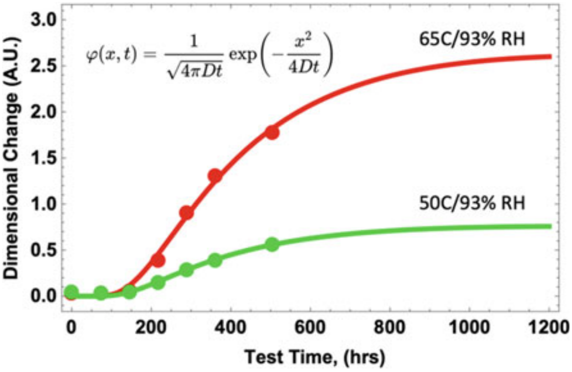


a)

(b)

Fig. 52 **a** Measured CTE dependence on RH, **b** measured CHE dependence on temperature, for LGP found experimentally that diffuser film would expand in length but shrink in width under HTHH test conditions. Both expansion and shrinkage would stay stable after 3-day of the HTHH 65 °C/90%RH test. Interestingly, prism film shrunk along with both longitude and latitude directions after HTHH. It shrunk more along longitude than latitude and the changes were stabilized after 3-day of the HTHH 65 °C/90%RH test as well. Therefore, for a full BLU film stack up, there are variations among the films, which not only affect the nominal dimension but also increase deviation in the tolerance stack up. The shrinkage of films also could result in wrinkles after HTHH exposure. All the PET-based optical films can be saturated with water during the HTHH test. It was reported that PET becomes brittle and would fail when about 0.55% hydrolysis of ester bonds [23]. Water vapor in the PET permeates into the film cavity with air space and then migrates toward the film edge during the post-HTHH dry-out period, resulting in a non-uniform drying process, which would cause film wrinkle. The time to reach a final equilibrium with wrinkle disappearance varies due to film size and display design. It should be noted that thinner film will expand or shrink faster because water diffusion throughout the entire thickness will be quicker. Thermal and hygrometric effects are assumed to be additive. Hydrometric expansion played a more important role in film dimensional changes. As an example, at 25 °C/50%RH, there is 16 torr of water vapor pressure. At 25 °C/90%RH goes up to 21 torrs. When soaking the BLU at 65 °C/90%RH, there will be 170 torr water vapor pressure. Not only 65 °C/90%RH condition will induce more water in the air to be absorbed readily by the polymer, but the higher temperature itself will accelerate the diffusion kinetics. The moisture diffusion-induced dimensional change, in general, can be modeled by Fick's second law. It was found that an incubation time existed, and it took time for the water to diffuse into polymers as shown in Fig. 53. The diffusivity was strongly modulated by temperature. By fitting the data to extrapolate the diffusivity, with the established CTE and CHE values, based on Eqs. 8 and 9, TTF of BLU film for film buckling/waving/wrinkle, therefore, can be quantitatively estimated for a fixed design at fixed testing or test condition.

Fig. 53 Polymer film thermal and hygrometric expansion modeling by Fick's second law



5.3 Metal Oxide TFT Panel-Level VGH and VGL Reliability Modeling

Compared with traditional silicon-based TFTs, metal oxide semiconductors usually contain more chemical elements. Therefore, they are more susceptible to the influences of the manufacturing process and use field environment, which leads to the deterioration of TFT performance. As an example, the oxygen content in metal compounds could be easily impacted during manufacturing processing, which leads to oxygen vacancy defects and uneven oxygen content [24, 25]. In addition, hydrogen in the process and from the working environment can easily diffuse into and out of the oxide semiconductor acting as a fast diffuser. Therefore, the stability of metal oxide TFT and how to model it has become a critical concern to the reliability of oxide TFT-driven LCDs. As the operation time increases, the gate TFT V_{th} is going to drift due to the positive bias temperature stability (PBTS) effect, which would cause VGH margin reduction resulting in color mixing, flickering level increase, and portions of panel failing to scan. On the other hand, the pixel TFT threshold voltage is going to drift due to the negative bias temperature illumination stability (NBTIS) effect, especially for display continuous operation under a wet environment. The negative pixel TFT V_{th} shift would cause VGL margin reduction, which results in an abnormality of panel FOS or latent defect.

As shown in Fig. 54, TFT device level ion reduction and Vth shift under PBTI demonstrated good correlations to panel-level VGH shift, confirming PBTS instability induced by the increase of density of electron trap states inside gate insulator or at the gate insulator (GI), and IGZO interface is the driving force to cause panel VGH margin shrinks. At the device level, the TFT V_{th} shift can be modeled by stretched-exponential (S-E) equation as described in Eq. 12 and shown in Fig. 54. With the S-E model, the V_{th} shift rate will be slowing down with increasing stress time. However, at the panel level, we found that VGH shift under stress conditions could be better modeled by a fraction power-law model (FPL) as described in Eq. 13 and shown in Fig. 54, instead of the S-E model. The FPL model is more generic since the S–E model is just an FPL into an exponential function. Therefore, FPL can also fit device level V_{th} shift equally well as S–E does. However, adopting FPL is simpler, generates a more conservative long-term result, and can guarantee fitting converging as compared to using the S–E model.

$$\Delta V_{\mathrm{th}} = (V_G - V_{\mathrm{th0}})\left\{1 - \mathrm{Exp}\left[-\left(\frac{t}{\tau}\right)^{\beta}\right]\right\} \tag{12}$$

$$\Delta \mathrm{VGH} = (V_{\mathrm{GH}} - V_{\mathrm{GH0}}) \times t^{n} \tag{13}$$

where β, τ, and n are fitting constants, and n must be less than 1.

A similar FPL model can be used to fit the panel-level VGL shift under NBTIS stress conditions as well. With the long-term stress data (Fig. 55), it was confirmed that FPL could offer excellent predictability for both PBTS and NBTIS stresses,

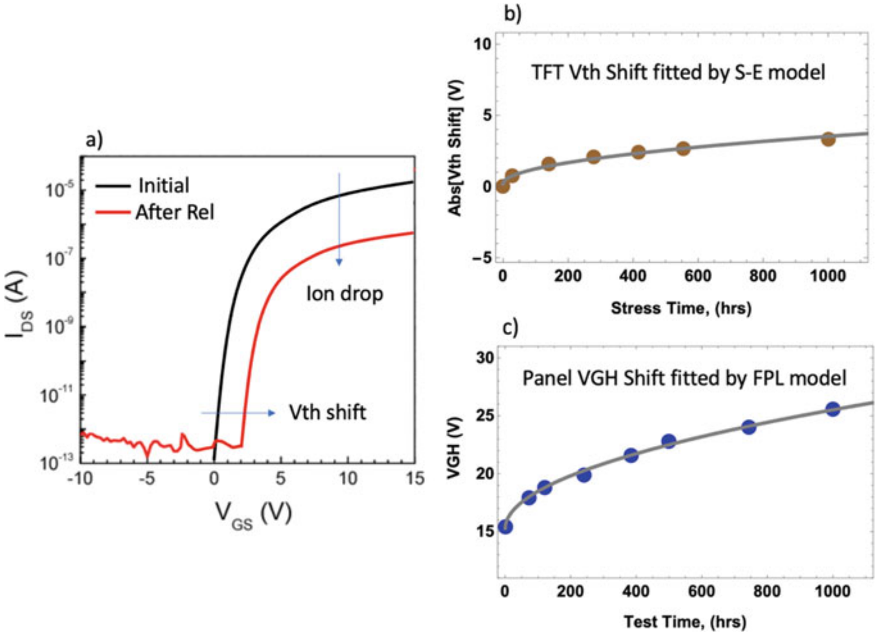


Fig. 54 **a** Device Ids versus Vgs curves to illustrate ion drop and Vth shift before and after PBTI stress condition, **b** plotted TFT Vth shift versus panel stress time fitted by S–E model, **c** plotted panel-level VGH shift versus panel stress time fitting by FPL

suggesting a similar failure physics such as a common defect reactivation and generation process exists for both cases. The defects are hypothesized to be dominated by excess oxygen (oxygen dangling bonds) and hydrogen. Optimization of IGZO and GI interface to minimize excess oxygen and hydrogen is the key step to reducing overall TFT defect density. As shown in Fig. 55, the solid line was the data fitting line based on the earlier testing data points, and the dashed line was the prediction line from the solid line. A good agreement of dashed line and measured data indicated good model predictability.

After establishing a good parametric shifting model, with known VGH and VGL threshold targets, VGH and VGL time-to-failure values can be obtained. It was found that VGH was sensitive to temperature, while VGL was sensitive to temperature, humidity, and illumination. It was also found that VGL degradation had a substantial recovery effect when the panel was off, which was associated with moisture re-distribution. Moisture seemed to play an important role in VGL margin degradation. Moisture could be trapped at T0 inside panel during panel process, which caused a prompt shift of VGL degradation. Moisture also could come from the moisture ingression from the external environment, which caused an enhanced VGL shift behavior. A duty factor modulation effect was needed in the acceleration model to account for moisture in and out diffusions. Lastly, Eqs. 14 and 15 were developed based on the physics of failure to model VGH and VGL on-time lifetime, respectively.

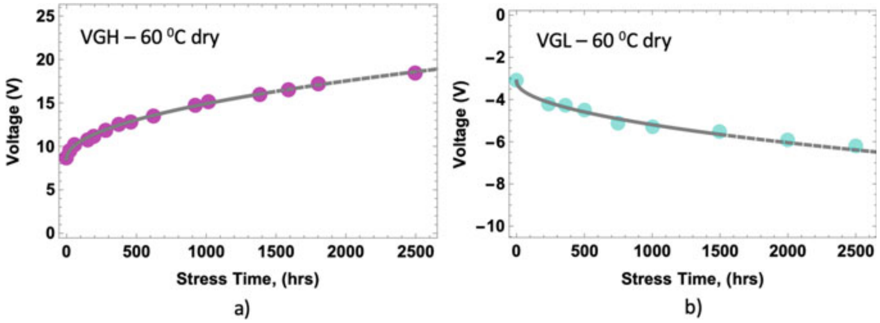


Fig. 55 **a** VGH shift data fitted by FPL model and its predictability demonstration, **b** VGL shift data fitted by FPL model and its predictability demonstration

$$\mathrm{TTF}_{\mathrm{VGH}} = A \times \left[\frac{V_{\mathrm{spec}} - V_0}{R}\right]^a \times \frac{1}{\mathrm{DF}} \times e^{\left[\frac{-\mathrm{Ea}}{k_B T}\right]} \tag{14}$$

$$\mathrm{TTF}_{\mathrm{VGL}} = B \times \left[\left|\frac{V_{\mathrm{spec}} - V_0}{R}\right|\right]^a \times \mathrm{RH}^{-n} \times \left(\frac{1}{\mathrm{DF}}\right)^m \times e^{\left[\frac{-\mathrm{Ea}}{k_B T}\right]} \times L^b \tag{15}$$

where V_{spec} is specification voltages, V_0 is starting voltage, R is degradation rate and a is power from the FPL model, DF is duty factor, Ea is the thermal activation energy, RH is relative humidity, n is Peck's exponent, L is light intensity, m is duty factor adjustment for recovery effect, b is luminance exponent, A and B are pre-factor constants.

For both VGH and VGL TTF distributions, it was found that neither 2-parameter Weibull nor Lognormal was applicable to fit the observed distributions as shown in Fig. 56. Since both Weibull and Lognormal statistics assume every sample comes from the same constituted population, both of them can be interpreted to describe the sample-to-sample variation based on the Poisson yield model. Poisson yield model assumes a constant defect density as shown in Fig. 57a. However, as mentioned before, during metal oxide TFT panel production, oxygen and hydrogen contents likely will not be uniformly distributed across the entire mother glass panel. Therefore, each panel likely will have its unique defect number as shown in Fig. 57b. A new statistical model based on non-uniform defect distribution is needed to be considered. C.H. Stapper proposed a negative binomial (NB)-based clustering defect model by assuming defect density is a random variable [28]. Based on that concept, a three-parameter time-dependent NB clustering model was adopted as shown in Eq. 15. When the clustering factor, α is infinity, the cluster model naturally reduces to Weibull as a special case so it was confirmed that Weibull and clustering model naturally are consistent.

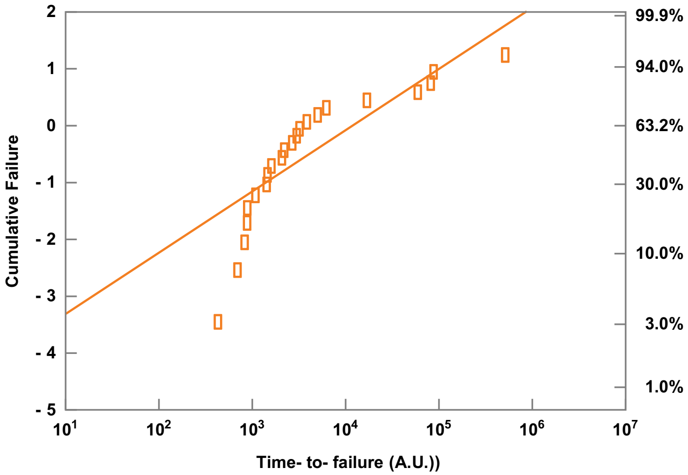


Fig. 56 Panel VGH TTF distribution fitted by Weibull. Due to its severely distorted shape, the fitting is not good

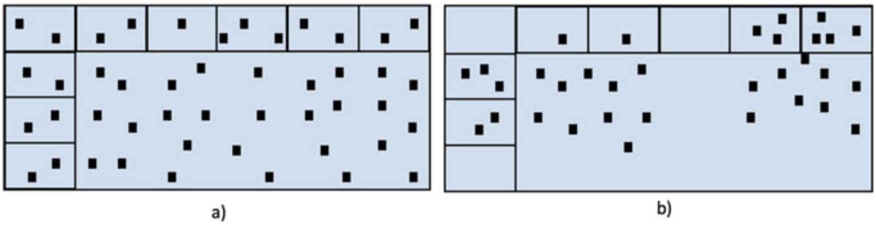


Fig. 57 Illustrations of **a** uniform distributed defects, **b** non-uniform distributed defects. In (**b**) defects are clustered

$$f(t) = 1 - \left[1 + \frac{1}{\alpha}\left(\frac{t}{\lambda}\right)^{\beta}\right]^{-\alpha} \tag{16}$$

where α is the clustering factor, β is the early failure population shape factor, and λ is the characteristic life. α provides a measure of the degree of non-uniformity or variability. The smaller α is, the more variability exists. The clustering model describes the failure rate to saturate toward high percentiles for a long time so it is the late failures that distort the shape, not the early failures. From our comprehensive panel VGH/VGL and device TFT level investigation, we found most early failure panels were from the mother panel (MP) corner locations. Figure 58 shows that by applying the clustering model, we can successfully model the entire distribution of panel VGH TTF.

Table 6 is an example of a comprehensive VGH or VGL reliability validation testing plan. According to Eqs. 14 and 15, based on the clustering statistical model,

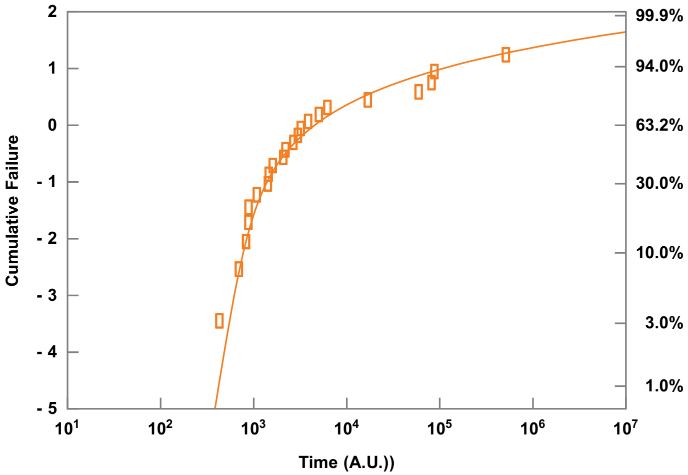


Fig. 58 Same data in Fig. 56 fitting by clustering model, which yields a perfect fitting

Table 6 VGH or VGL reliability testing plan example

Test Cell	Temperature (°C)	RH (%)	Frequency (Hz)	Duty factor (%)	Luminance	Sample size
A-C	50, 60, 70	Dry	60	100	100% Brightness	10 per cell
D-F	60	70, 80, 90	60	100	100% Brightness	10 per cell
G	60	90	60	100	50% Brightness	10 per cell
H	60	90	120	100	100% Brightness	10 per cell
l-J	60	90	60	25, 50	100% Brightness	10 per cell

a lifetime or failure rate projection can be made as illustrated in Fig. 59 per a specific mission profile. In Fig. 59, all the stress condition data were normalized to a fixed stress condition based on Eq. 14 for VGH. By using our panel-level VGH and VGL reliability qualification methodology, we can soundly assess LCD TFT-related reliability performance.

5.4 LCD Panel UV Irradiation Aging Reliability Modeling

LCD panels and polarizers utilize organic compounds that are susceptible to high heat and light energy stress. These organic compounds will eventually break down if

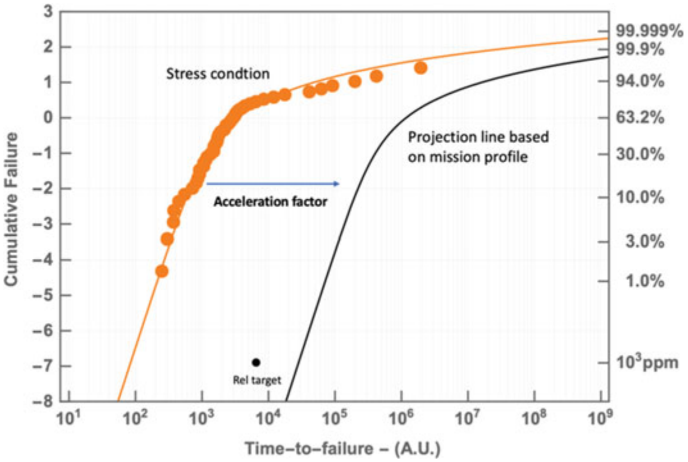


Fig. 59 An illustration of panel-level VGH reliability lifetime modeling based on clustering statistical model and Eq. 14 acceleration formula

they are deployed in high-stress environments. One such contributing factor to LCD panel failure is solar irradiation. The solar irradiation not only adds excess heat that may overheat the liquid crystal and prevent them from working properly, but it also adds UV band energy that is destructive to organic compounds. In most applications, UV irradiation effects are observed to be brightness uniformity shift, color shift, washed-out images, and an observable rise in the darkness levels produced by a damaged LCD panel. LCD sheet wrinkle failure also could be generated after UV irradiation.

There are many UV testing standards available for automotive parts UV testing such as SAE J2527, SAE J2413, ASTM D4459, IEC 60068, GMW14650, and DSAA V5.1. DSAA defined in-vehicle display UV specification is 830 W/m^2, profile Z-IN-2, 25 days (15 days dry, 10 days humid). Spectral power distributions (SPD) from Atlas UV accelerated testing chamber with controlled artificial light source are plotted, one for indoor and one for outdoor, respectively, in Fig. 60a. It was confirmed by the UV dosimeter in the dashboard that for in-vehicle displays, the worst-case radiant energy exposure is very close to the indoor SPD shown in Fig. 60a. The irradiation intensity can be used as an acceleration factor to test the LCD panel for its solar aging performance. Based on the reciprocity breakdown concept, the time to reach a certain level of material degradation is inversely proportional to the rate at which photons attack the material. The photon rate is directly correlated to total light energy. Furthermore, we can use a simple additivity approach to determine the light intensity by assuming the photo-effectiveness of a light source is equal to the sum of the effectiveness of its spectral components. Then, a simple UV acceleration reciprocity rule can be deduced as: the amount of degradation in a weathering exposure is strictly proportional to the product of irradiance and time. This simple rule may not be 100% correct but it can be used as 1st order estimate. As shown in Fig. 60b as an example, LCD color shifts under outdoor and indoor UV irradiation testing were observed. Based on the UV acceleration reciprocity rule, the acceleration factor difference from outdoor to indoor was calculated to be about 7.

For display UV testing, it is found that the air cavity between CG and panel could plan an important role to trap heat to increase panel temperature during UV irradiation. Therefore, whole system-level UV testing is recommended. As far as we can model the LCD color, brightness, and luminance uniformity changes versus UV stress time, per DSAA V5.1 color specifications such as du‘v’ ≤ 0.005, brightness reduction $\leq 20\%$, and uniformity $\geq 75\%$ at the end of vehicle life, we can extrapolate the lifetimes for each specification category. The master UV reliability acceleration model was developed as shown in Eq. 17.

$$\mathrm{AF(TH+UV)} = \mathrm{DF} \times \left[\frac{\mathrm{RH}_{\mathrm{test}}}{\mathrm{RH}_{\mathrm{use}}}\right]^{n} \times \mathrm{Exp}\left[\frac{Ea}{k_B} \times \left(\frac{1}{T_{\mathrm{use}}} - \frac{1}{T_{\mathrm{test}}}\right)\right] \\ \times \left[\frac{\int_{\lambda_1}^{\lambda_2} I_{\mathrm{test}}(\lambda)\lambda e^{-B\lambda}\mathrm{d}\lambda}{\int_{\lambda_1}^{\lambda_2} I_{\mathrm{use}}(\lambda)\lambda e^{-B\lambda}\mathrm{d}\lambda}\right]^{p} \tag{17}$$

where I is intensity, λ is the wavelength, p and B are wavelength-dependent modeling constants. Figures 61, 62 and 63 demonstrate the LCD panel UV aging reliability assessments for chromaticity change, brightness change, and luminance uniformity change, respectively. From this example, it was obvious that UV irradiation changed LCD panel chromaticity and degraded LCD panel brightness substantially. However, panel luminance uniformity was improved a little bit simply due to a substantial reduction of the panel brightness after UV. It should be noted that as radiation can increase the temperature of the LCD panel, a true panel temperature corrected from chamber temperature plus irradiance energy should be used in Eq. 17. The realistic UV exposure mission profile for in-vehicle can be estimated based on worldwide radiant exposure data. As an example, the 99th percentile of worldwide radiant exposure is 1269 kWh/m^2 for 15 years. For in-vehicle display, based on an average

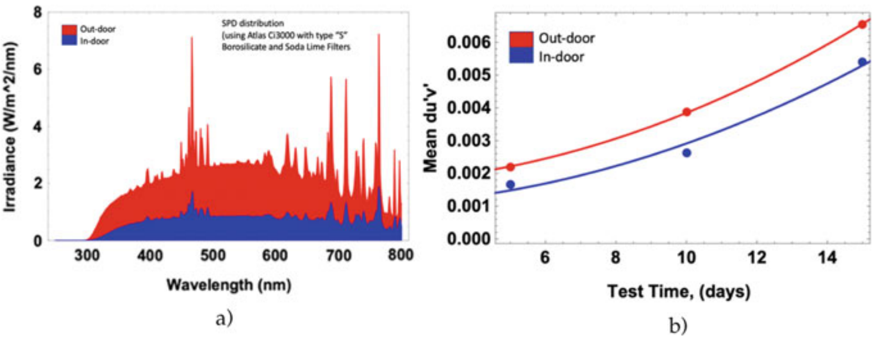


Fig. 60 **a** UV testing chamber SPD distributions, **b** LCD color shift under two UV testing conditions defined in **a**

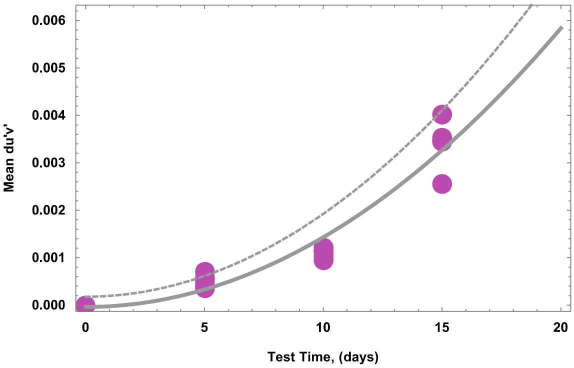


Fig. 61 LCD panel chromaticity change under UV irradiation modeling. Some shift variations were observed. The solid line is the modeling line based on mean change, and the dashed line is the modeling line based on the worst-case points (upper envelope fitting)

2 h per day driving under sunlight and indoor display application, 1269 kWh/m^2 can be scaled down to 52.875 kWh/m^2 as an in-vehicle display lifetime accumulated radiant energy exposure requirement. If we use irradiance of 100w/m^2 for our testing, we need to test a minimum of 528.75 h to fulfill the lifetime UV exposure requirement.

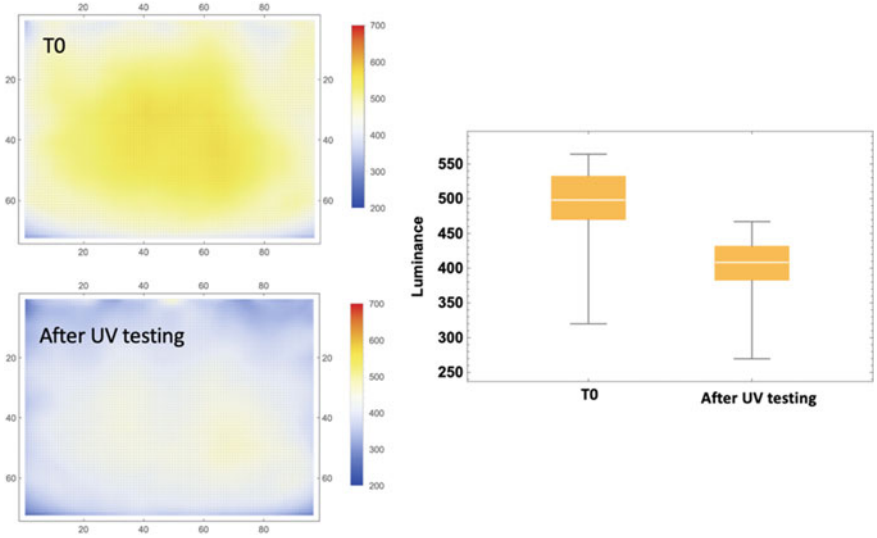


Fig. 62 LCD panel brightness change after UV aging study. A substantial brightness reduction was observed at the end of UV testing

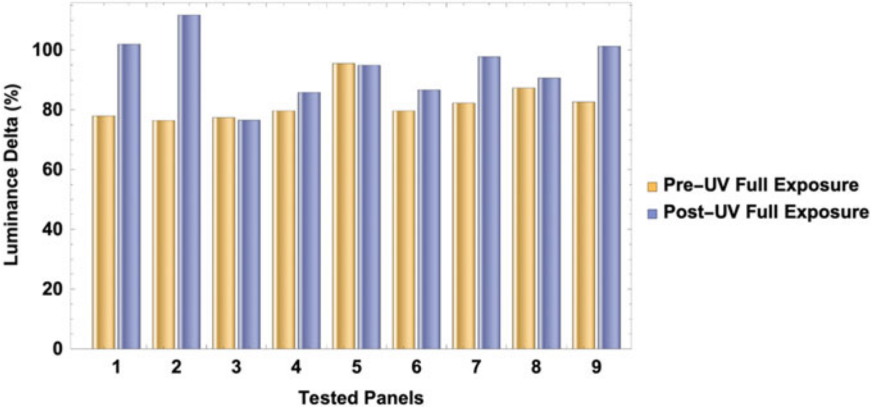


Fig. 63 LCD panel luminance uniformity change after UV aging study. UV aging did not degrade luminance uniformity

5.5 *Polarizer Edge Bleaching Failure Mechanism and Reliability Modeling*

Polarizing film is one of the core elements of LCD technology. It allows the display to show its image by controlling the amount of light. It converts unpolarized light into linearly polarized light by transmitting only the incident beam that propagates in one direction and absorbing all others. To make the polarizer, the polyvinyl alcohol (PVA) in the polarizer film is stretched, which causes the polymers to align. Then, the film is dipped in a solution of iodine, and the iodine molecules attach themselves to the polymer. As PVA can dissolve in water and the iodine complex can break down in moisture, the polarizer is extremely sensitive to liquid condensation and moisture ingression. During LCD high-temperature high-humidity test, one common failure mode related to polarizers is its edge bleaching. Boric acid crosslink in PVA can be destroyed by moisture, therefore causing the polarizer to lose its birefringence. The damaged PVA will shrink from the edge and become thicker, which results in a FOS light bleaching issue. The whole process is illustrated in Fig. 64. The edge quality of the polarizer plays an important role in edge bleaching. The bleaching probability is higher alone in the polarizer stretch direction as compared to the non-stretch direction. Higher temperature would certainly enhance the moisture diffusion, therefore would cause more bleaching. The presence of contaminations such as copper ions from ITO tape oxidation will exacerbate the bleaching issue as Cu ions could form CuI_2 clusters so to act as catalyzers to enhance the polarizer bleaching resulting in a circular pattern.

The reliability modeling of polarizer edge bleaching is pretty straightforward. By running a DoE that consists of temperature and humidity variations, we should be able to determine bleaching thermal activation and humidity exponent. The bleaching width versus heat soak stress time can be modeled reasonably well by a square root

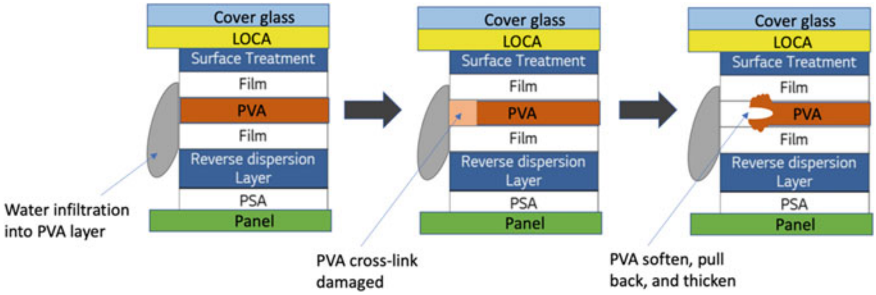


Fig. 64 Polarizer edge bleaching mechanism illustration

of time model as shown in Fig. 65, suggesting a diffusion-based failure physics for edge bleaching. The thermal activation energy was extrapolated to be 0.42 eV by using the Arrhenius relation as shown in Fig. 66.

Fig. 65 Polarizer edge bleaching modeling based on root-t diffusion physics

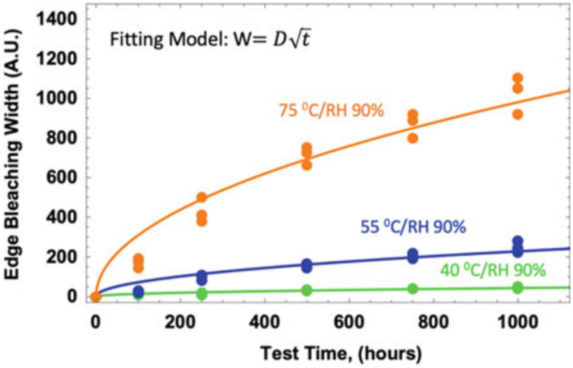
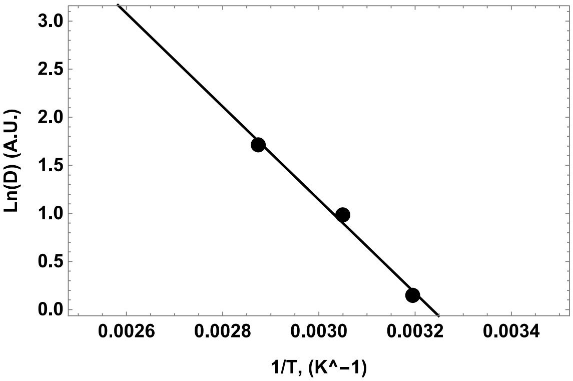


Fig. 66 Polarizer edge bleaching thermal kinetic study



5.6 Free-Fall Object Impact Test and LCD Glass Crack Failure Risk Assessment

The purpose of the free-fall object impact test is to ensure that the LCD under test can withstand a blunt impact when an object is dropped or bumped onto the LCD surfaces. For in-vehicle display, the most relevant damage modes from impact tests are LCD cover glass and TFT glass cracks. When conducting an object impact test, the amount of energy entering the UUT will influence the observed failure rate. In general, energy is dissipated by viscose-elastic deformation of object and UUT, internal UUT friction, possible plastic deformation of the metal enclosure, and when glass fracture occurs, energy to grow the cracks. Figure 67 illustrates a typical object impact test setup with the bottom supporting floor including a holding fixture, particleboard, and granite block. The coefficient of restitution (CoR) test usually is required to calibrate the object impact tester bottom supporting floor to assure a consistent energy dissipation during an impact test. The CoR is a measure of how much energy is dissipated during an impact event. A simple acoustic measurement technique can be used to determine the CoR based on Eq. 17. CoR value of about 0.5, in general, is recommended.

$$\mathrm{CoR} = \frac{\Delta t}{2}\sqrt{\frac{g}{2H}} \tag{18}$$

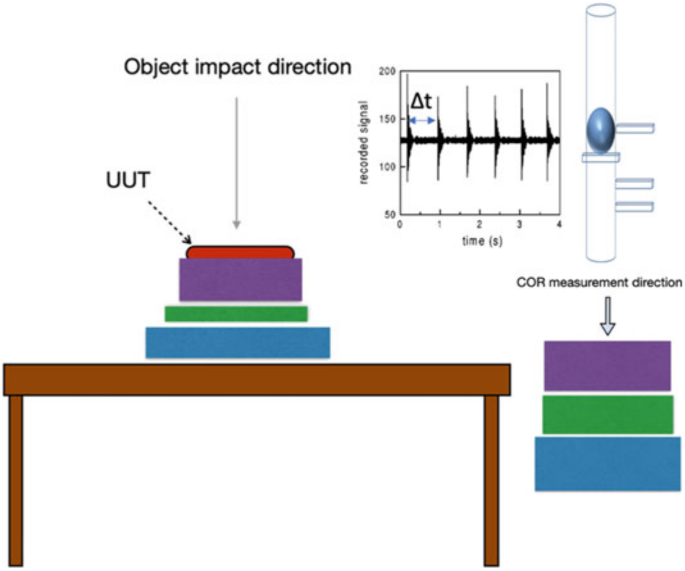


Fig. 67 Object impact test setup and calibration with CoR

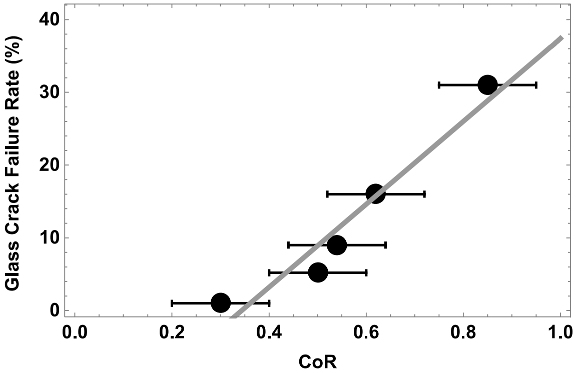


Fig. 68 LCD glass crack failure rate induced by object impact test versus CoR of impact system bottom supporting floor

where H is the drop height, g is gravitational acceleration, and Δt is the time delay between the first and second impact correlating to the bounciness of the testing surface. A CoR measurement setup is shown in Fig. 67. The CoR for a collision depends on the nature of the two bodies that collide, the ball, and the horizontal surface of the impact system bottom supporting floor in this experiment. As shown in Fig. 68, it was found that LCD glass crack failure rate has a good correlation with impact system bottom supporting floor CoR. Due to this reason, the object impact testing setup should not be moved around during any testing time to ensure consistency for data collection.

The impact objects can be steel and acrylic materials with varied shapes, sizes, and weights, although the sphere shape is the most common one. An object impact system consists of a guiding tube withstand, a stop pin for various drop heights, a bottom test fixture nest to hold UUT, and a bottom supporting floor system. An automated precision ball impact testing system with automatic ball release and collection, automated ball recovery with no secondary ball bounce impact, automated ball z-height set and drop, various initial ball drop velocities, automatic drop location XY movement, multiple temperature operations, and automated post-drop surface inspection is shown in Fig. 69 as a reference. The testing procedure usually requires dropping an object with pre-defined weight, size, and starting height to all critical surface locations of a UUT. To obtain a full spectrum of kinetic energy to fracture, increasing the drop height until a failure occurs is also required. Usually, the weakest location's reliability determines the entire panel's reliability based on the weakest link concept. The first phase object impact test is to identify the weakest location on the panel. Then, the kinematics at the weakest crack site can be further conducted by recording the height-to-failure at that specific location for a deep dive reliability analysis.

Figure 70 shows the object impact height-to-fracture distribution and corresponding fracture kinetic energy distribution Weibull plots. The equations to convert free-fall height-to-fracture to kinetic energy are shown in Eqs. 19–21. By knowing the fracture kinetic energy and field use kinetic energy distribution from a user study,

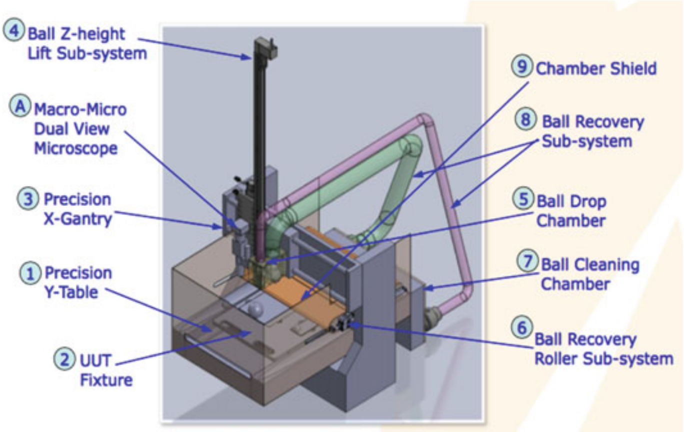

④ Ball Z-height Lift Sub-system
Ⓐ Macro-Micro Dual View Microscope
③ Precision X-Gantry
① Precision Y-Table
② UUT Fixture
⑨ Chamber Shield
⑧ Ball Recovery Sub-system
⑤ Ball Drop Chamber
⑦ Ball Cleaning Chamber
⑥ Ball Recovery Roller Sub-system

Fig. 69 Automated precision ball impact testing system example

we can conduct a stress-strength analysis as shown in Fig. 71, and obtain an accumulative failure rate for a specific cover glass embedded in a specific LCD system design. It is noted that in Fig. 71, the integrand does not fall under both stress and strength probability density function (PDF). In this case, above 200 J the stress cumulative density function (CDF) is close to 1, so that portion of the strength CDF will be nearly preserved.

$$h = \frac{1}{2} g t^2 \tag{19}$$

$$v = g t \tag{20}$$

$$K = \frac{1}{2} m v^2 \tag{21}$$

5.7 LED Lumen Degradation Reliability Modeling

The reliability of LED products may be affected by both lumen drop and color shift. LED lifetime usually is measured by lumen maintenance, which is how the intensity of emitted light tends to diminish over time. The Alliance for Solid-State Illumination Systems and Technologies (ASSIST) defines LED lifetime based on the time to 50% light output degradation (L50: for the display industry approach) or 70% (L70: for the lighting industry approach) light output degradation at room temperature. However, display panel manufacturers currently all require L70 as the spec for LEDs used in their usage conditions. Since the luminous flux degradation and color shift

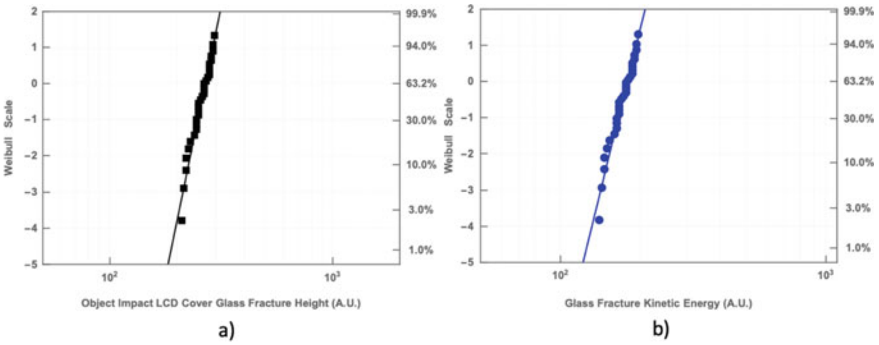


Fig. 70 Cover glass fracture height (**a**) and kinetic energy (**b**) from object impact testing

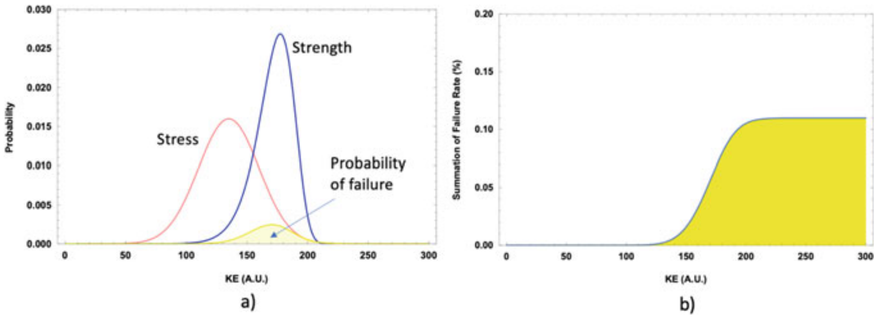


Fig. 71 Cover glass fracture reliability estimate by stress-strength analysis in (**a**) and the cumulative failure rate in (**b**)

usually take a very long time to observe under normal operating conditions, the accelerated temperature life test is always used as a substitute for the LED operating life test to quickly predict LED lifetime. In this chapter, a methodology to predict the lifetime of LEDs based on environmental accelerated stress testing with multiple stress factors is discussed. The process involves (1) measuring the light output of samples at each test condition; (2) estimating LED life under the accelerated test conditions using parametric curve fitting of transient Lumen under the test conditions based on L70 as lifetime target; (3) calculating a master acceleration factor covering temperature, humidity and current; (4) determine lifetime statistical behavior based on Weibull statistics and Poisson area scaling law; (5) predicting final product lifetime per automotive mission profiles by 3D stress-strength modeling.

A knowledge-based LED reliability validation testing plan is shown in Table 7, which covers the temperature, moisture, and current acceleration impacts. In the table, Tx refers to stress temperature conditions, RHx refers to relative humidity conditions, and Ix refers to stress current conditions. For the 1st step, as shown in Fig. 72, the luminance of each individual LED was recorded at multiple readout points. A prompt temporary increase in luminance at an early stage of the operation was observed in many tests. This initial burn-in behavior could be the result of several effects including annealing of defects in the LED epitaxial layer, bisbenzocyclobutene (BCB) photo-catalyzed disassociation, reduction in contact resistances, re-distribution of phosphors layer particles, and changes in the refractive index of materials in the light path. Because of this unique "prompt" positive shift signature, for the 2nd step, we proposed a hybrid power-law (PL) + exponential (Exp) decay model to describe our observed competing shift behavior. Our PL + Exp model could fit the entire time-dependent lumen shift very well as shown in Fig. 73. By comparing with several other LED lumen decay models which all can fit 3000-h data reasonably well (Fig. 73), we confirmed that PL + Exp model was the best choice because it not only could offer the best predictability based on our measured long-term 5500 h data point but also was the most conservative model as illustrated in Fig. 73. If a LED lifetime was passing based on PL + Exp model, it would be guaranteed to pass by using the other two models.

Based on the L70 lifetime target, time-to-failure from all the LEDs could be extrapolated using PL + Exp model. Then, the TTF distribution from, i.e., RH2/I2 cell could be constructed based on Weibull statistics as shown in Fig. 74 as an outcome from step 2. For the 3rd and 4th steps, a comprehensive LED lumen decay acceleration model including temperature, moisture, and current factors to scale TTF from stress level to field condition was established based on the following Eq. 22:

Table 7 LED reliability validation testing plan

Cells	RH1/I1	RH2/I1	RH2/I2	RH2/I3	RH3/I1
T1			22x		
T2	22x	22x	22x	22x	22x
T3			22x		
Duration	1 k-7 k hr	1 k-7 k hr	1 k-7 k hr	1 k-7 k hr	1 k-7 k hr

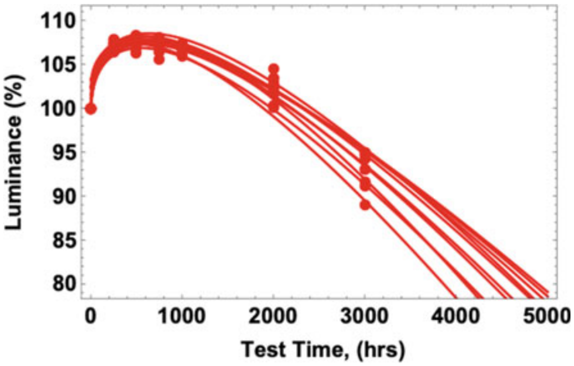


Fig. 72 LED lumen decay under an HTHH stress condition. The fitting lines were based on the hybrid power-law + exponential decay model

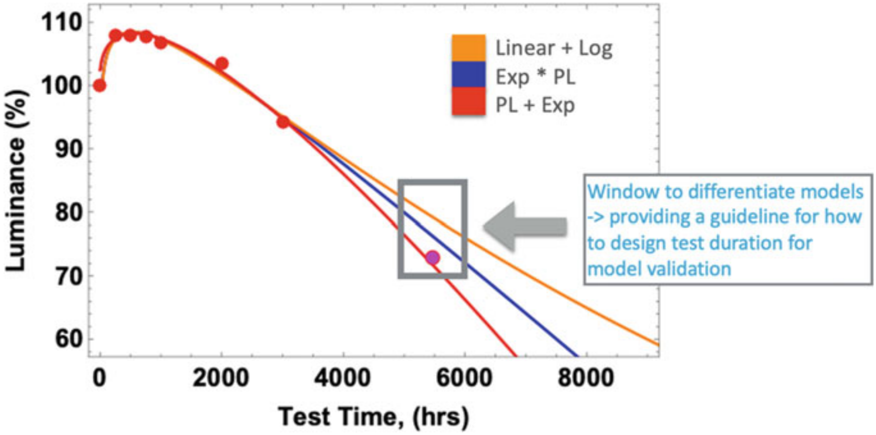


Fig. 73 LED lumen decay model comparison. PL + Exp model exhibits the best predictability and the most conservativeness

$$\mathrm{AF} = e^{\frac{\mathrm{Ea}}{k_B(T_u - T_s)}} \times \left(\frac{\mathrm{RH}_s}{\mathrm{RH}_u}\right)^n \times \left(\frac{I_s}{I_u}\right)^m \tag{22}$$

where Ea, n, and m were constants that could be experimentally determined by grouping all the testing cell data together with a common Weibull slope fitting according to Eq. 23:

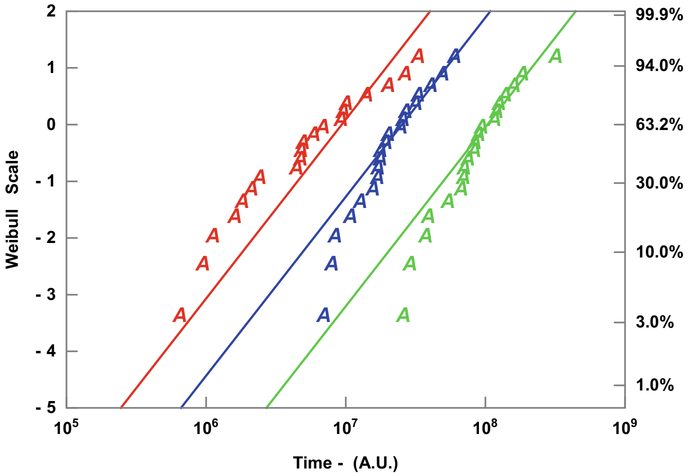
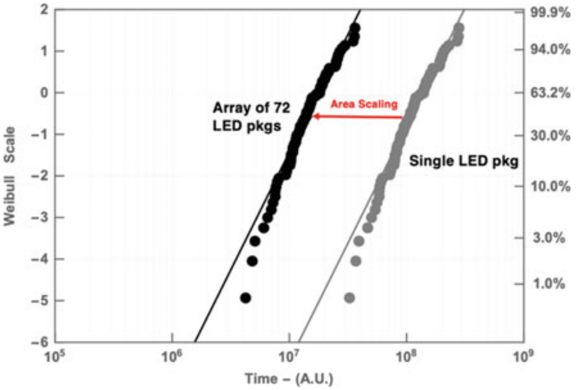


Fig. 74 Extrapolated LED lumen decay TTF plotted in Weibull distributions

Fig. 75 Reliability of an array of 72-LED packages transformed from a single LED package reliability data



$$f(t) = 1 - \text{Exp}\left[-\left[\frac{\text{EOL}}{\frac{A}{I^n}(\text{RH})^m \frac{1}{r^s}\text{Exp}\left[\frac{\text{Ea}}{k_B T}\right]}\right]^{\beta}\right] \tag{23}$$

Another factor that needed to be considered for step 4 was area scaling. Our data were measured from a single LED package. For an edge-lit LED LCD panel, a lightbar usually consisted of some LEDs. To model the reliability of an array of LEDs based on a single LED reliability, the Poisson area scaling law shown in Eq. 24 should be used. As shown in Fig. 75, a transformed TTF distribution representing a 72-LED array lifetime was calculated based on the Poisson formula of Eq. 24. It should be noted that the Poisson area scaling law can be used from one size of the LED array to another size of the LED array, which could be especially useful for mini-/micro-LED array performance and reliability modeling.

$$t_{\text{array}} = t_{\text{single}} \times \text{AR}^{\frac{-1}{\beta}} \tag{24}$$

where AR is the LED number ratio. For the Fig. 75 case, it was 72.

The last step of the LED lifetime modeling methodology is to estimate the LED lifetime or failure rate. As shown in Fig. 76, the field temperature and the in-vehicle display field usage time mission profiles could be described by two skewed normal PDFs according to Eq. 25. In this modeling process, for simplicity, we assume LED operating current and in-vehicle humidity levels during vehicle operation is constant. To calculate a realistic LED failure rate per our established temperature and operating time mission profiles, a 3D stress-strength failure rate equation shown in Eq. 26 was developed, which incorporates one Weibull failure rate CDF, and one skewed normal time PDF, and one skewed normal temperature PDF. In Eq. 26, T is temperature and t is time.

$$\text{PDF} = \frac{1}{\sqrt{2\pi}\sigma}\text{Exp}\left[\frac{-(x-\mu)^2}{2\sigma^2}\right] \times \text{Erfc}\left[-\frac{\alpha(x-\mu)}{\sqrt{2\pi}\sigma}\right] \tag{25}$$

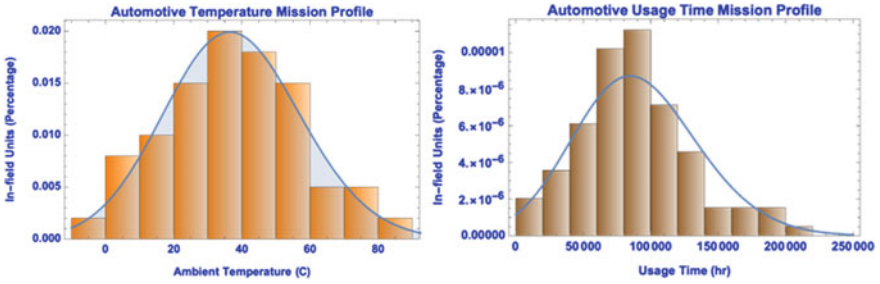


Fig. 76 Vehicle temperature and time usage mission profiles

$$
\begin{aligned}
\mathrm{FR} = \iint \Bigg\{ \Bigg(1 - \mathrm{Exp}\Bigg[-\Bigg(\frac{t}{\eta\mathrm{ref} \times \mathrm{Exp}\left(\frac{\mathrm{Ea}}{k}\left(\frac{1}{273+T} - \frac{1}{273+\mathrm{Tref}}\right)\right)} \Bigg)^{\beta ref} \Bigg] \Bigg) \\
\times \left(\frac{1}{\sqrt{2\pi}\sigma} \mathrm{Exp}\left[\frac{-(t-\mu)^2}{2\sigma^2} \right] \times \mathrm{Erfc}\left[-\frac{\alpha(t-\mu)}{\sqrt{2\pi}\sigma} \right] \right) \\
\times \left(\frac{1}{\sqrt{2\pi}\sigma} \mathrm{Exp}\left[\frac{-(T-\mu)^2}{2\sigma^2} \right] \times \mathrm{Erfc}\left[-\frac{\alpha(T-\mu)}{\sqrt{2\pi}\sigma} \right] \right) \mathrm{d}T\,\mathrm{d}t
\end{aligned}
\tag{26}
$$

The final product LED array failure rate PDF and CDF were plotted in Fig. 77. From its PDF plot, it was found a failure rate peak occurred at around 60 °C-53000 h points. This condition shall rarely happen for all in-vehicle display applications. The double integrand calculated failure rate from Eq. 26 was meeting the 99% reliability target at end of vehicle life.

Lastly, let us discuss the mini-LED luminance degradation uniformity issue and its impact on long-term mini-LED display reliability. Human eyes in general are more sensitive to cross-panel non-uniformity than absolute luminance decay. In other words, if every LED lumen degrades uniformly from 100 to 70% or even 50%, users usually still can tolerate such uniform brightness degradation. However, if each LED or a group of LEDs has substantially different lumen degradation rates, which result

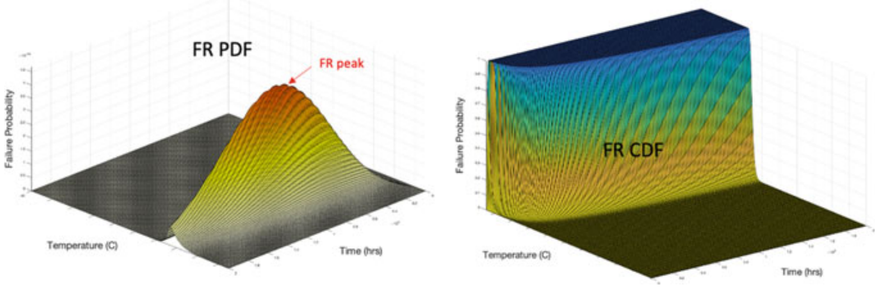


Fig. 77 LEF failure rate PDF and CDF plots based on 3D stress-strength failure rate analysis

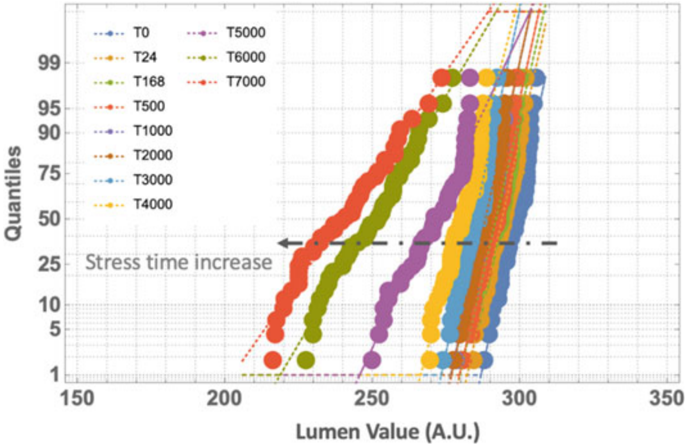


Fig. 78 Mini-LED lumen distribution and degradation

in a cross-panel uniformity to be less than 75%, users likely will pick this issue up immediately. As shown in Fig. 78, mini-LED single dies had a minor die-to-die luminance variation at T0. But this die-to-die variation showed regression under a normal life test condition. The longer the stress time was, the larger the die-to-die luminance variation exhibited. A clear bimodal distribution of lumen was observed at the end of 7000 h testing, suggesting potentially two-lifetime populations of tested mini-LEDs existed. One population showed a faster decay process than the other population did. Based on the obtained luminance decay characteristics, a Monte Carol (MC) simulation was used to predict the luminance decay behavior of an array of 10 k mini-LED over the course of 1000 h life usage as illustrated in Fig. 79. In the simulation, a randomly selected array case and a cluster array case were both simulated. Our MC simulations indicated that after 1000 h operating life usage, mini-LED display luminance uniformity could be degraded to below 70%. If “bad” mini-LEDs were cluster positioned, a cluster non-uniformity effect could be much worse. Therefore, an effective T0 screening and picking, plus an effective LED degradation control are critical to minimize brightness Mura for 2D mini-LED array display for its long-term usage.

6 Summary

From the previous chapters, we can see the differences between in-vehicle display and consumer display products. There are more strict requirements for in-vehicle displays in all aspects such as optical visual effect, reliability, quality requirements, safety, environmental mission profiles, warranty, and so on. The much longer product life cycle of the in-vehicle display (10–15 years for automotive versus 5 years for the

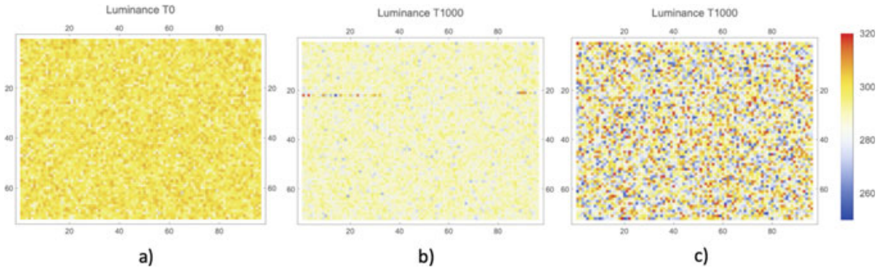


Fig. 79 MC simulation of on-panel lumen uniformity: **a** T0 case, **b** after 1000 h operating showing a row of cluster Mura effect, **c** after 1000 h operating showing a degraded uniformity with $\Delta L < 75\%$

consumer), together with the safety requirements involved with occupants and even pedestrians, makes the automotive OEMs define a high-standard specification for in-vehicle displays. Regarding the development trend of in-vehicle display technology, the following results can be observed:

a. Large size: In future, vehicle displays will no longer mainly display vehicle information, but also will serve as a major display role in mobile offices or mobile entertainment halls. Large-size displays can integrate more information platforms.
b. High resolution: To meet more diverse display applications, the resolution of the car display will increase as the size is enlarged to provide better visual effects and user experience.
c. Low-energy consumption: In the foreseeable future, automakers will need to achieve zero CO_2 emissions so that electric vehicles (EV) will be the mainstream vehicle of this evolution. To improve the energy/mile range conversion efficiency of electric vehicles, the power consumption of on-board displays needs to be minimized. Therefore, low-power consumption is one of the key design points for in-vehicle displays used in electric vehicles. In some locations with direct exposed sunlight, reflective type or transflective type may be considered to replace the general transmissive type display. Self-luminous mini-/micro-LED has the characteristics of low-power consumption and high reliability. Therefore, they may have the opportunity to become the mainstream technology of the in-vehicle display used in EVs.
d. Flexibility: The evolution of automobiles continues to increase its impact on the design of electronics. Future EV will gradually eliminate buttons and switches and replace them with a streamlined 3D curved surface design. The flexibility of the car display is the future of the car body. The key technology of 3D surface design plus OLED and mini-/micro-LED arrays can enable this design requirement.

In conclusion, the displays are the central interface between the car and human beings. Future displays will be more customized regarding size, shape, and aspect ratio to allow competition differentiation. The mature TFT LCD technology is still the mainstay for in-vehicle in the short term. OLED and mini-/micro-LED array technology have a relatively large advantage in the medium and long term. Due to the unique requirements and opportunities such as aggressive adoptions of EVs, digitalization, and autonomous vehicle, the in-vehicle display technology development will be moved from a “follower” to a “front runner”. Digital mirror, exterior display, and transparent car window display will have great potential to drive new applications in the modular and interaction vehicle field. The automotive display technology will continue to evolve as there are still many emerging technologies as well as the refinements of existing technologies are still under development. This continued growth will overwhelm the driver and occupants with maximum comfort and convenience on the road.

Acknowledgements The authors want to thank Brian Strom from Apple, and Ke Liu from Meta for their reviews and useful suggestions for this chapter.

References

1. G.H. Heilmeier, L.A. Zanoni, and L.A. Barton, Dynamic scattering: A new electro-optic effect in Certain Classes of Nematic Liquid Crystals. Proc. IEEE, 56, pp 1162–1170 (1968)
2. T.P. Brody, J.A. Asars, and G.D. Dixon, A 6x6 Inch 20 Lines-per-Inch Liquid-crystal display Matrix. U.S. patent No 3,654,606
3. Bernanose, A.; Comte, M.; Vouaux, P. (1953). A new method of light emission by certain organic compounds. *J. Chim. Phys.* **50**: 64
4. Tang, C. W.; Vanslyke, S. A. (1987). Organic electroluminescent diodes. *Applied Physics Letters*. **51** (12): 913
5. Sanyo, Kodak ramp OLED production line. *EETimes*. 6 December 2001
6. OLEDs in Display - NOVALED | Creating the OLED Revolution. *novaled.com*. Retrieved 27 November 2019
7. https://www.theglobeandmail.com/drive/culture/article-shift-from-huge-in-car-screens-may-be-under-way-but-first-theyll/#:~:text=In%201986%2C%20the%20Buick%20Riviera,to%20have%20a%20touch%20screen
8. http://www.xintech.com.hk/zh/others/glass-lcd-tnhtnstnfstnva/
9. https://www.merckgroup.com/tw-zh/expertise/displays/solutions/liquid-crystals/lcd-technologies/ips.html
10. https://www.materialsnet.com.tw/DocView.aspx?id=8068
11. https://driving.ca/cadillac/features/feature-story/cadillacs-new-oled-screens
12. https://www.cnet.com/tech/home-entertainment/what-is-mini-led-tv-and-how-does-it-improve-samsung-sony-and-tcl-tvs-2022/
13. https://www.digitimes.com.tw/tech/dt/n/shwnws.asp?cnlid=1&id=0000567318_6I14AUAO0WTC1F9IJO6LF
14. https://www.nature.com/articles/s41377-020-0268-1
15. David S Hermann, Automotive displays – trends, opportunities and challenges. AMFPD symposium, 2018, SS1–1
16. Liu Jinquan, Yang Shengjie and Ye Zhou, Automotive display trend and Tianma’s directions. AMFPD symposium, 2019, SS4–1
17. Valeriano Ferreras Paz and Efstathios Persidis, Automotive Display Requirements and Developments, AM-FPD symposium, 2019, SS3–1
18. http://www.isininc.com/en/home/products/20

19. X. Li, J. Wang, J. Liu, and F. Qin, International Conference on Display Technology, 2020, volume 51, issue S1, pg. 194
20. https://www.google.com/url?sa=i&url=https%3A%2F%2Fir.nctu.edu.tw%2Fbitstream%2F11536%2F64312%2F5%2F950305.pdf&psig=AOvVaw17yPd2qIU46eSaJGuP1Ss3&ust=1616662746923000&source=images&cd=vfe&ved=2ahUKEwjuzKSXyMjvAhXWzIsBHSbLB8UQr4kDegQIARBb
21. Khaled Layouni, T. M. Gross, M. Black, J. T. Harris, JS Park, Y. K. Qaroush, Retained Strength for AutoGrade^{TM} Cover Glass, 2019 SID, 8.2
22. J Yang, C. Liao, J An, Z. Peng, J. Liang, S. Zhang, International Conference on Display Technology, 2020, v52, Issue S1, pg 200
23. J. E. Pickett and D. J. Coyle, "Hydrolysis Kinetics of Condensation Polymers Under Humidity Aging Conditions," Polymer Degradation and Stability, vol. 98, pp. 1311-1320, 2013
24. D. Kang, H. Lim, C. Kim, et al., Amorphous gallium indium zinc oxide thin-film transistors: sensitive to oxygen molecules. J. Appl. Phys. Lett. 90(19):488 (2007)
25. M. D. H. Chowdhury, J. G. Um, J. Jin, "Remarkable changes in interface O vacancy and metal-oxide bonds in amorphous indium-gallium-zinc-oxide thin-film transistors by long time annealing at 250 °C", J. Appl. Phys. Lett. 105(23):2945 (2014)

Disk Drive for Data Center Storage

Zhen Wei and Xi Qian

Abstract In the modern era of the Internet of Things (IoT), data generation speed is exploding. Autonomous vehicles (AV), one type of IoT, are generating enormous amounts of data during operation. Most of the data will be saved in the data center. As the major data storage in data centers, hard disk drives (HDD) have a history of more than 60 years. HDD is the art of combining magnetic sensors, electromechanical components, and electronics. To meet the increasing demand of data storage, next generation HDD is under development with tremendous advantage in data areal density for high-capacity data storage, but meanwhile with challenge in reliability.

1 Introduction

Most of the data generated by AV will be saved in cloud storage. The cloud storage stores data on the Internet through a cloud provider who manages and operates data storage as a service. HDD forms the central element of cloud storage because of high capacity and low cost. Recent years, cloud storage demand far outpaces the considerable rate of innovation in HDDs. This chapter starts with describing the HDD application in the data center. Then, the overall HDD system design is illustrated. The last section of the chapter describes the exciting progress in the development of next generation HDD, including microwave-assisted magnetic recording (MAMR) and heat-assisted magnetic recording (HAMR). HAMR has more potential than MAMR in terms of achieving higher areal density but is facing more challenges in reliability due to high laser heating on the recording elements.

Z. Wei (✉) · X. Qian
Seagate Technology, Bloomington, MN, USA
e-mail: utexaszhen@gmail.com

Y. Li and H. Shi (eds.), *Advanced Driver Assistance Systems and Autonomous Vehicles*,
https://doi.org/10.1007/978-981-19-5053-7_12

2 Hard Disk Drive Application in Data Center

2.1 Data Storage for Autonomous Vehicle

Autonomous vehicle (AV) is considered as a moving computer with door control units (DCUs), electronic control units (ECUs), LIDARs, multiple long and short-range radars, many ultrasonic, and cameras [1]. For consumer AV, all these sensors could generate up to 15 terabytes (TB) of data per day depending on driving time. For robot taxi like AV, the expected operation time is nearly 24 h per day with 6–21 high resolution cameras. They can generate up to 450 TB data per day. Then the question is, with this big amount of data generated by AV, where is the data going? The answer is that it will be a mix. The AV must be functional independent of any connectivity. A small portion of the data will be processed within the vehicle focusing on the self-driving activities. Meanwhile, majority of the data need to be uploaded and stored in cloud. The cloud can utilize the powerful computing resource which is not available in the vehicles themselves [2]. The location of data storage is also depending on the phase of self-driving capability. There is a training phase and inference phase. During the training phase, the vehicle will collect local data as much as possible. The local data, including the surroundings and roads, is used to train the AI network algorithms to accurately detect the objects. When AV is operating in inference mode, as the self-driving capability has been successfully trained, the algorithm will continuously monitor the system performance and identify the new situation or object. This type of data will typically be sent to the cloud for future algorithm improvements.

There is continuing debate about how much data should be kept in the vehicle versus pushed up into the cloud, but one thing for sure is that the data center and its data storage are playing an important role in the final AV system.

In cloud storage, there are several classes. The standard storage is good for frequently accessed data. The nearline storage has a lower cost and is typically used for data that can be stored for at least a month, such as multimedia content and backup files. The cold-line or archive storage, having the lowest cost, is for long time data storage. In the standard storage, where high performance is required, the solid-state drives (SSDs) are mostly used in the full-flash array. In the nearline and cold storage, hard disk drives (HDDs) provide better combination of true cost to own (TCO) and performance. Even though the cost of SSDs is becoming cheaper, they are still more expensive on a dollar/gigabyte ($/GB) basis when compared with HDD. A hybrid architecture combining SSDs and HDDs will always provide flexibility and cost efficiency for different requirements. Therefore, in the foreseeable future, in data center secondary workload, or price sensitive application environments, HDD will continue to play an important role [3].

2.2 Data Storage Configurations in Data Center

Data center is a dedicated space within buildings to house computer systems and associated components, such as telecommunications and storage systems [4]. The first data center started in the 1940s. Due to their complexity, early computer systems required a special environment for operating and maintaining. During the boom of the microcomputer industry and the Internet, from the 1980s to 2000, companies desired structured network equipment, fast Internet connectivity, and non-stop operation through the Internet. Many companies started building very large facilities, called Internet data centers. In recent 10 years, with the growth of cloud data storage, the data center has changed to be very different from traditional one, which has on-premises physical servers. The cloud has virtual networks that support applications and workloads across pools of physical infrastructure and into a multi-cloud environment. In a cloud system, the data center must be able to communicate across multiple sites and connect the data across many types of clouds including edge, public, and private clouds [5].

Data center is composed of multiple elements including power supplies, communication and storage equipment, fire suppression equipment, heating, ventilation, and air-conditioning (HVAC) equipment, and monitoring system. They together provide safe and secure locations for data and equipment [6]. Here, the design of the communication and storage system is critical in a data center because it provides the network infrastructure to connect servers, the storage infrastructure to hold the data, and the computing resources to provide the processing and memory.

As the name suggests, data is the fuel for data centers. There are several configurations of data storage in the data center. The three most used configurations are: direct-attached storage (DAS), network-attached storage (NAS), and the storage area network (SAN) [7].

The storage configuration with storage device directly attached to a computer is called DAS. The storage device in DAS can be a single drive or part of an array or redundant array of independent disks (RAID) configuration. There is typically a host computer to control the DAS devices. Meanwhile, the host computers are connected in the network and the DAS drives can be shared across the network. DAS typically provides better performance than networked storage solution, which contends with network bottlenecks. Because of simple structure with directly attached devices, DAS is easier and cheaper to implement and maintain. But DAS is lacking scalability with limited number of expansion slots on host computer.

NAS is a file-level storage device, which has own IP address inside the local area network (LAN). Through the network, users and applications can access the data in NAS from a centralized system. To transfer the data, NAS also has a file transfer protocol. NAS devices are relatively inexpensive because they are easy to deploy and operate. NAS is the best choice for small to medium-sized business because of its flexibility in integration with cloud services. The downside of NAS is the contention when the network bandwidth is limited and cannot compete with other traffic on the network.

SAN is a block-level storage solution, which utilizes a dedicated high-speed network to connect multiple storage blocks. Each storage blocks can be made up of HDDS or SSDs or a hybrid configuration. SAN can be highly available and scalable with the right network topology and internal configuration. However, SAN is a complex environment that can be difficult to deploy and maintain. Many large enterprises invest in the SAN configuration when dealing with numerous or massive datasets. SANs can benefit use cases such as email, media libraries, or database management.

For many organizations, DAS, NAS, and SAN solutions are properly sized and configured to handle their workloads. Each approach offers both advantages and disadvantages, and they will likely play a vital role in the modern data center [7].

2.3 Hard Disk Drive Versus Solid-State Drive in Data Center

In any storage configuration, the physical components to store the data are mostly HDDs and SSDs. There are numerous comparisons between HDD and SSD, and prediction of when SSD will totally replace HDD. The fact is that more than 80% of the data in the data center are still stored in HDD every year. With high operational speeds and greater robustness, the SSDs have supplanted HDDs in the consumer market. But when it comes to data center implementations, the consideration is different.

In data centers with large-scale high-capacity storage installations, HDD is providing superiority in one of the key parameters: price/gigabyte ratio. In the past decade, there has been an order of magnitude difference between the price/gigabyte of enterprise HDD and SSD. SSD pricing has been reduced to catch up a little now, but the difference is still around 8x. In the foreseeable future, the pricing curves are likely to remain parallel between SSD and HDD. The innovation in HDD technology will keep increasing the HDD capacities and lowering the price/gigabyte ratio, which continuously makes HDD a highly attractive option in data center [8].

The production capacity of SSDs is another concern when replacing all the HDD resources with solid-state equivalents. In 2019, analyst firm Gartner reports HDD shipments equating to 890 Exabyte, while total SSD capacity coming to 153 Exabyte, which is only 16% of what HDD achieved. For SSD to completely replace HDD, SDD production output would need to increase six-fold. This will require thousands of billions to bring new fabs. This is just to get us to a point of data storage capacity in 2019. The amount of data that our society is generating is rising at an exponential rate. International data corporation (IDC) predicts that data generation will have surpassed 175 zettabyte annually by 2025. It is clear that reliance solely on a flash memory-based approach is not practical. Since the COVID-19 outbreak, the cloud service activity is increasing dramatically, which further underlines the credentials of HDDs in data center [8].

The clear advantage of an SSD is the performance. A single SSD with 2500 Gbps of bandwidth and 100K IOPS operational performance is much better than the fastest HDD, which has 250 Mbps and 300 IOPS. But things might be different when we are talking about large data storage systems. With smart architectures, a price-comparable system including many HDDs can match the performance of a system comprising smaller number of SSDs. In an architecture with 24 to 60 HDDs in RAID 10, the performance of 5 Gbps and 10K IOPS was achieved. This solution featuring many HDDS will result in a better price per capacity than fewer SSDs.

In conclusion, based on current price points and future price projection, production capacity, performance trade-off, and the HDDs are still the most commercially viable way to store data on a large scale in a data center.

3 Hard Disk Drive Design

3.1 Hard Disk Drive System

Hard disk drive (HDD) is a system which can convert the electrical signal into magnetic signals and vice versa. HDD is a non-volatile storage because the data can be stored in magnetic media without the support of electrical power. There are typically four components in an HDD system: recording magnetic components, mechanical components, electromechanical components, and electronics. Figure 1 shows the important components in the HDD system.

The first and most important component is the recording magnetic components. It has two parts, recording sliders and disk. Disk is made of aluminum substrate and coated with several layers of material. First layer is a soft underlayer (SUL). In

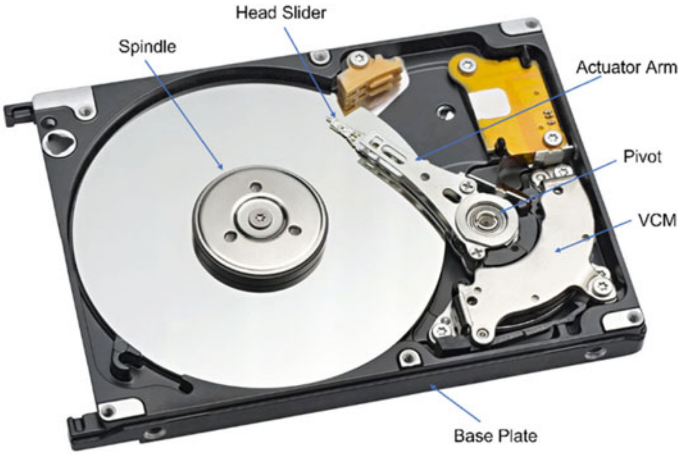


Fig. 1 Typical components in HDD [9]. Drawn by author

perpendicular recording, the writing magnetic field is penetrating the disk perpendicularly, and the SUL will help guide the magnetic flux and produce stronger magnetic field gradient. On top of the SUL, there are magnetic layers. The magnetic material with higher magnetic coercivity is selected to provide better thermal stability. Thus, the magnetic cell/grain can be smaller for higher areal density. The very top layer on the disk is a lube overcoat. It provides the wear durability and corrosion inhibition. During HDD operation, the recording head is flying on top of the spinning disk with the fly height < 1 nm. There are a lot of head-disk-interactions that could damage the disk and head surface. The lube overcoat on the disk can protect the underneath magnetic layer, and typically any minor scratches on the lube can be covered within several days due to lube migration. The recording sliders contain writing, reading element, thermal wire for contact detect, air bearing surface for fly control, heating block for height control. Details will be covered in the next section.

The second component is the electromechanical component. There are two different electromechanical components in the HDD system: (1) a spindle motor to spin the disk, (2) an actuator to re-position the read–write heads on the desired data track and to maintain its position precisely over the track while it is being read or written [9]. The spindle motor has been transferred from ball-bearing to fluid dynamic bearing (FDB) with the demand for higher areal density and faster spindle speed. At stationary, the FDB spindles have larger area of surface-to-surface contact and can withstand larger non-operational shock. At motion, FDB spindles have no contact between the rotating surface and the stationary surface. Moreover, the viscosity of the lubricant between the two surfaces is much higher than that for the lubricant used in ball bearings. As a result of these features, the FDB spindle motors generate less acoustic noise and lower non-repeatable runout (NRRO). Here, NRRO is a bottleneck in achieving higher track density. Another electromechanical component is the actuator. The recording slider is carried by suspension arms which are driven by a voice coil motor (VCM) actuator. VCM is a moving coil type actuator, in which, a coil is held suspended in the magnetic field produced by pairs of permanent magnets fixed to the casing of the HDD. When a current is passed through the coil, it moves.

The other two HDD system components are mechanical and electronic components. The mechanical parts include top cover and base plate. All other HDD system components will be mounted on the base plate, and the enclosure is then covered using the top cover. The electronic components include channel electronics for read/write, servo channel for head positioning, and disk controller for controlling various operations, such as read, write, and transfer data.

3.2 Components in Recording Head

In HDD, the recording head is the most complicated component. It determines the HDD performance and reliability, and it also drives the HDD technology innovations. What makes the recording head so essential in HDD? To answer this question, let us look at the recording head components in Fig. 2. The recording head needs to fly on top of a spinning disk, with typical rpm = 7200. The active clearance between recording head and disk media surface could be as low as 5A. To achieve such low clearance, the first step is mounting the recording head slider on head-gimbal assembly (HGA), where the suspension made of metal alloy will ensure the slider flying steadily at passive height (~20 nm). Here, the suspension is designed to have proper stiffness for sustainability and to avoid resonance frequency co-located with operational frequency, such as servo frequency. When looking into the recording head slider surface, there is an air bearing surface (ABS). The ABS defined the air flow between the slider and the disk surface. High-pressure air flow will make the slider flying at stiffer mode, improving the reliability. But it will bring more difficulties in characterizing the passive fly height for active clearance setting. if we zoom-in the trailing edge (TE) of the slider ABS, there is the core of the slider, which is usually called transducer. It is about 40um wide, with many functional structures at only 50–1000 nm wide.

Inside the transducer, there are complicated thin film layer structures. They are mainly for two functions: (1) write and read the magnetic signal; (2) thermal mechanical blocks to set active clearance during operation. The write element is a thin film coil structure that puts out a magnetic field when current is passed through the coil [9]. The generated magnetic flux can be guided through a magnetic soft underlayer in the media disk and is able to flip the vertically aligned magnetization bits. The

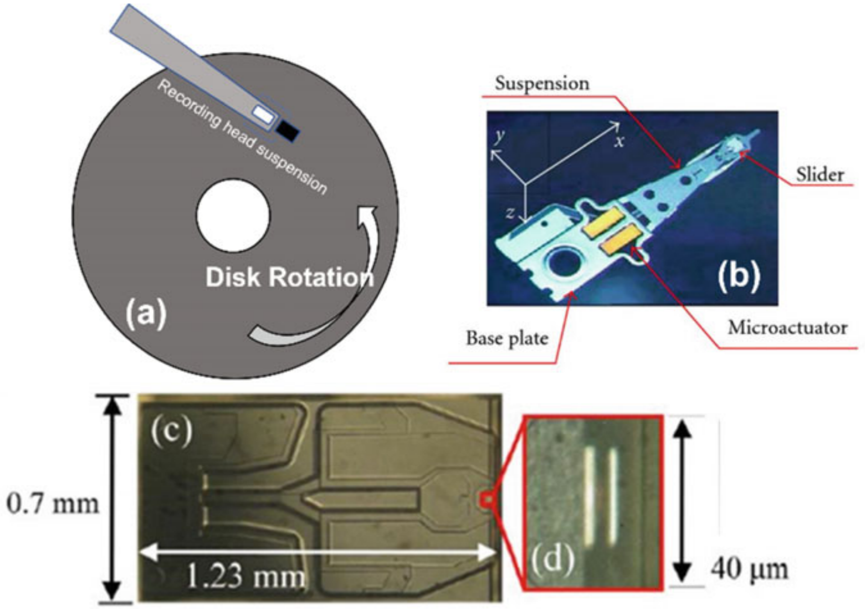


Fig. 2 Typical recording head components. **a** Schematic of recording head and media position; **b** typical head-gimbal-assembly; **c** slider air bearing surface (ABS); **d** slider trailing edge (TE) zoom-in for transducer [10, 11]. Here [10] author gave the permission, [11] is open resource

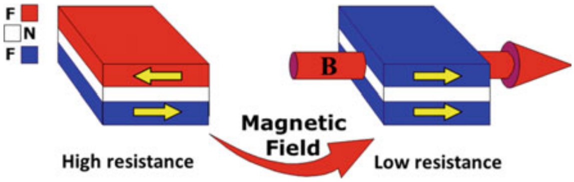


Fig. 3 Giant magnetoresistance (GMR) is defined as resistance change from ferromagnetic layers antiparallel orientation to parallel orientation. Drawn by author

read element is basically a multilayer spin valve structure with giant magnetoresistance (GMR), which generates up to 80% resistance change depending on the orientation of layer magnetization. Figure 3 shows the basic mechanics of GMR. In a simple spin valve structure, which contains two ferromagnetic layers separated by a nonmagnetic layer, when two ferromagnetic layers are oriented in same direction, the spin valve electrical resistance is low. When they are oriented in opposite direction, the spin valve resistance is high. When the reader goes through the magnetic bits, the magnetic field from the media disk can change the direction of magnetization in reader-free layers, hence the reader GMR. This small resistance variation will be detected as an electrical voltage change and then converted to computer binary.

The thermal mechanical element in the recording head is called heater block and thermal wire. The heater block is usually tungsten with low thermal coefficient resistance. When applying electrical current on the heater block, the metal layer will expand and push the key elements on the recording head toward the media disk until reaching target clearance. The thermal wire is used to detect the amount of electrical current needed to make recording head touching media. This can calibrate the amount of electrical current for any target active clearance. In later recording heads, to achieve extremely low head-media spacing during operation, there are two separate heater blocks dedicated for writer and reader. They are positioned to make writer or reader as close point during thermal protrusion. This can help bring the active fly height down to sub-nanometer territory.

3.3 Next Generation Hard Disk Drive

The hard disk drive needs to keep increasing the number of data bits on a given area (areal density) and lowering the cost of $/GB, to compete with other storage technology which has advantage in performance. Over the last 50 years, the areal density of hard disk drives increased by an average of 40% each year. However, recently, the increase rate has slowed down to be around 10% [12]. To achieve higher areal density in next generation hard disk drive, there are several technology innovations. One of innovations is the shingled magnetic recording (SMR), which utilizes the fact that written in profile is wider than read-back profile on a data track. In SMR, the data tracks are partially overlapped, which creates a pattern similar to the shingles [13]. SMR allows more data tracks to be written on the disk surface and increases the areal density. Western digital is working on microwave-assisted magnetic recording (MAMR). MAMR adds a spin torque oscillator, which generates high-frequency microwave field to help flipping the magnetic grains on the media disk. It allows smaller magnetic grain size on the media and increases the areal density. Seagate technology is working on heat-assisted magnetic recording (HAMR). For HAMR, there is a component called near-field transducer (NFT). The laser through the designed wave guide excites surface plasmons on NFT, which transfers energy to the media grains, and makes it easier to flip [12]. Same as MAMR, HAMR can write the media grain with much smaller size, hence higher areal density.

4 Challenges in the Performance and Reliability

4.1 The Need for Higher Areal Data Density

The storage density of hard disk drives (HDD) has doubled every three years since their introduction in 1955. HDD's areal density capacity (ADC) has increased by a million-fold in the past 60 years, with a compounded annual growth rate (CAGR) of about 30%. So far, HDD is still maintaining its ADC margin over flash storage, as shown in Fig. 4. This makes HDD the preferred choice of massive data storage. Several major technology advances, including the development of thin film media and recording heads, giant magneto-resistive readers, and perpendicular recording, have enabled the current areal density of 750 Gb/in^2, which is about half the theoretical limit for conventional magnetic recording materials [14].

As the magnetic grain size is reduced to increase the storage density, the grains may become superparamagnetic and render their magnetic state thermally unstable. The traditional longitudinal magnetic recording (LMR) hits the superparamagnetic limit at an areal density of around 100 Gb/in^2. In the 2000s, perpendicular magnetic recording (PMR) was commercialized and soon took over LMR to become the dominant magnetic recording method. By switching the magnetization direction from in-plane to out-of-plane versus the disk (Fig. 5), PMR can better maintain grain size while increasing areal density, thus allowing higher ADC before hitting the superparamagnetic limit. Moreover, by selecting materials with a large magnetic anisotropy, such as L1$_0$ FePt, the grain size can be as small as 2–3 nm in diameter, which in PMR can be translated into the storage densities up to 1.5 Tb/in^2. PMR is still the most dominant magnetic recording method as of today.

However, besides the superparamagnetic limit, another challenge for high ADC is that the coercivity of high-anisotropy materials will eventually be greater than the magnetic field generated by a recording head. That is to say, even if the magnetic grains can store the data, it is difficult to write the data into such a small area. Writing at high ADC is the major challenge for the HDD industry. Currently, there are two major approaches in increasing the ADC, namely microwave-assisted magnetic

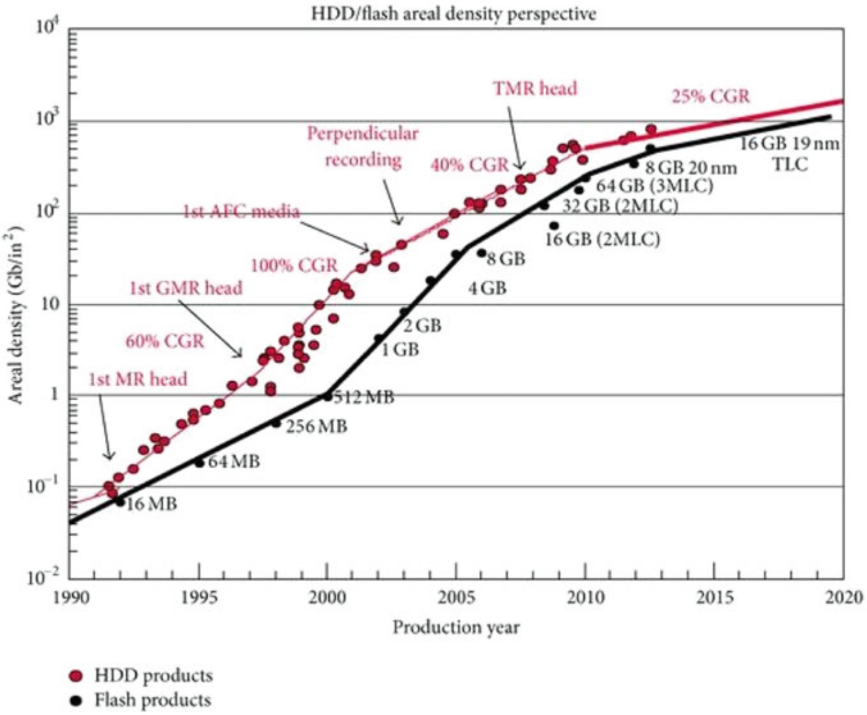


Fig. 4 Evolution of the areal density of HDD(red) and flash memory (black) since 1990 [15]. Copyright © 2013 Bruno Marchon et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited

recording (MAMR) and heat-assisted magnetic recording (HAMR), both technologies have undergone years of research, and both have developed near-commercialized prototypes.

4.2 *Microwave-Assisted Magnetic Recording (MAMR)*

The basic idea of MAMR is to use ferromagnetic resonance (FMR) to reduce the localized switching field, and thus enable writing at higher ADC. It is known that magnetization switching can be proceeded by magnetization precession that follows the Landau-Lifshitz-Gilbert (LLG) equation [17]. One example of controlling magnetization precession is spin transfer switching (STS) in non-volatile magnetic random access memory devices (MRAM), which transfers a spin polarized current to a localized magnetic moment via an angular momentum [18]. Similar to STS, microwave-assisted switching (MAS) is another way to control the magnetization

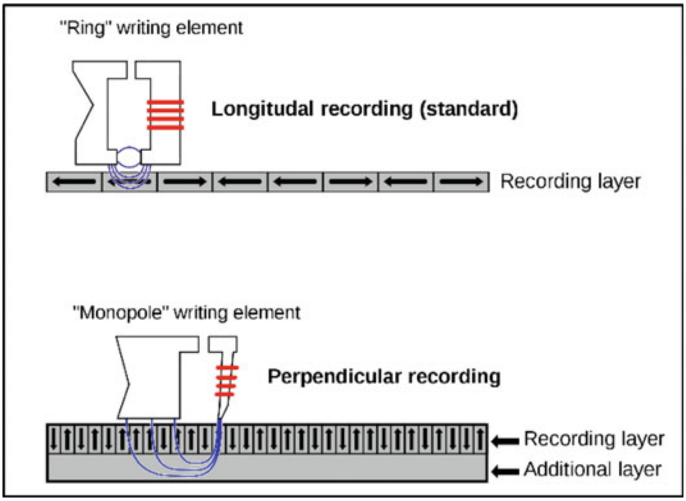


Fig. 5 Schematic of longitudinal recording (top) and perpendicular recording (bottom) [16]. Public Domain image by Luca Cassioli 2005

precession. MAS applies radio frequency (rf) magnetic field to drive the processional motion in a ferromagnetic element and thereby induces FMR. When large angle magnetization precession is excited, the resonance dynamics are in the nonlinear response regime. Previous study showed that the magnetic grain's switching field can be significantly reduced by applying a large amplitude of the pulse field. To better utilize the MAS effect for HDD purpose, spin torque oscillator (STO) has been proposed as a RF field generator. Since STO is very similar to the structure of the magneto-resistive reader head of the HDD, the drive integration would be relatively simple. As a result, MAS plus STO provides a solution for microwave-assisted magnetic recording.

Figure 6a depicts the expected recording scheme of MAMR. The STO is placed between the main pole and the trailing shield of the single pole type write head. The magnetization can be expressed by LLG equation:

$$\frac{\mathrm{d}m}{\mathrm{d}t} = -m \times h + \alpha m \times \frac{\mathrm{d}m}{\mathrm{d}t} \tag{1}$$

where m is the unit vector of magnetization, $\boldsymbol{\alpha}$ is the dimensionless Gilbert damping constant, t is the time in unit of $\left(|\gamma| H_k^{\text{eff}}\right)^{-1}$, in which, γ is the gyromagnetic ratio, and H_k^{eff} is the effective anisotropy field inducting the demagnetization field, which can be expanded as:

$$H_k^{\text{eff}} = 2\frac{K_u}{M_s} - 4\pi N M_s \tag{2}$$

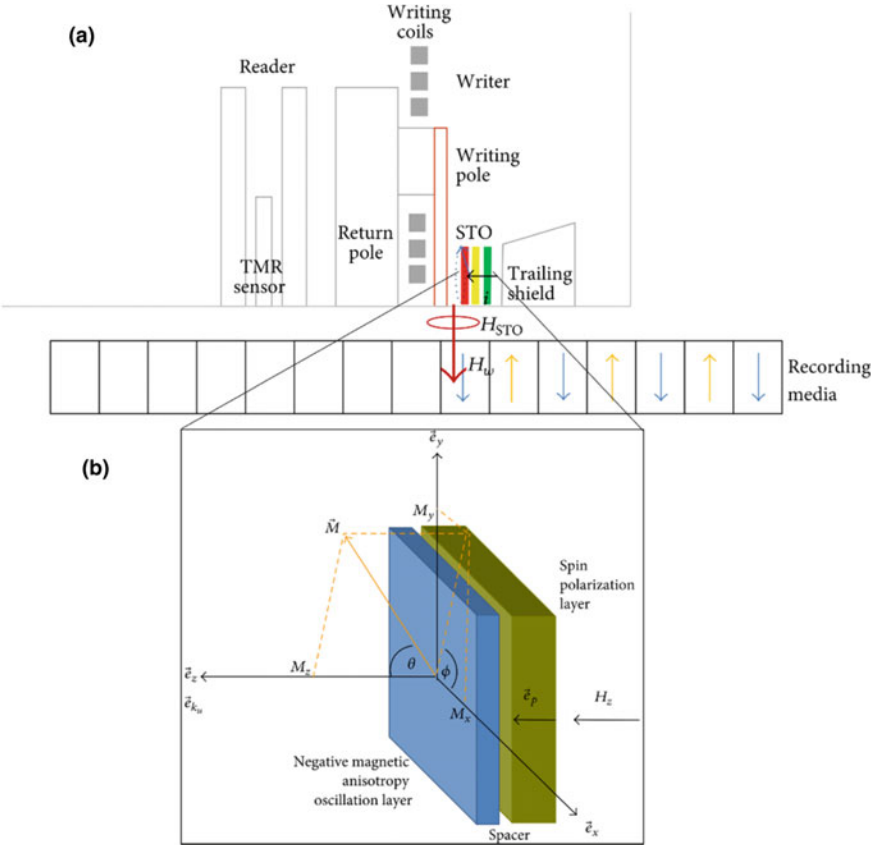


Fig. 6 **a** Schematic illustration of MAMR recording head, **b** illustration of STO layer [19]. Figure 6a and b are subjected to Copyright © 2015 Mingsheng Zhang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited

where K_u is the magnetic anisotropy, M_s is the saturation magnetization, and N is the demanganization factor the magnetic grain.

Moreover, h denotes the normalized effective field:

$$h = h_{\text{dc}} + h_k^{\text{eff}} + h_{\text{rf}} \tag{3}$$

where $h_{\text{dc}} = H_{\text{dc}}/H_k^{\text{eff}}$ and $h_{\text{rf}} = H_{\text{rf}}/H_k^{\text{eff}}$ are normalized dc (main) and rf (assistant) fields, respectively.

When the z-axis is set to be the direction of magnetic easy axis in head and media, the normalized anisotropy field H_k^{eff} is expressed as $m_z e_z$, where m_z is the z-component of m, and e_z is the unit vector along the z-axis (vice versa to e_x and e_y). The RF field is applied in the x–y plane, transverse to magnetic easy axis, and can be described as:

$$h_{\mathrm{rf}} = h_{\mathrm{rf}}\left(\cos\cos\omega t e_x + p\sin\sin\omega t e_y\right) \tag{4}$$

where ω is the angular frequency in unit of $\left(|\gamma| H_k^{\mathrm{eff}}\right)^{-1}$, p is the polarization of the RF field, for example, 1 for counterclockwise, 0 for linear and -1 for clockwise circular, respectively.

As shown in Fig. 6b [19], the applied oscillation field is in x–y plane, the effective field h is precessing with respect to z-axis, at an angle of θ_H. The magnetization precession is excited only when the chirality of the Rf field is in accord with that of the precessional motion. The angular frequency of materials precessional motion, $\omega_{0,}$ can be determined by:

$$\omega_0 = \gamma H_k^{\mathrm{eff}} \tag{5}$$

When $\omega \neq \omega_0$, the Gilbert damping effect (determined by α) will eventually pull h back to z direction. But if $\omega = \omega_0$, the ac oscillation can keep providing energy for the precession, and the precession can thus reduce the magnetic field required to flip the magnetic domain.

A more detailed implementation of MAMR is shown in Fig. 7. This is an early proposed design. The key is the perpendicular spin torque-driven oscillator, which generates microwave frequency ac fields. The oscillator has a perpendicular, permanent magnetic layer for spin polarization of the injected current, next to a metallic interlayer, then a high saturation moment field generating layer (FGL), and capped with a layer with perpendicular anisotropy. The last two layers form the oscillating stack via ferromagnetic exchange [20].

The spin torque mechanism is shown in Fig. 8. The magnetization of the FGL undergoes an effective magnetic field along the perpendicular axis, which is sum of the interlayer exchange field and the planar self-demagnetization field. Here, the effective anisotropy field (Eq. 2) can be re-written as:

$$H_k^{\mathrm{eff}} = \left(\frac{\sigma_{\mathrm{int}}}{M_s\delta} - 4\pi M_s\cos\theta\right)e_z \tag{6}$$

where σ_{int} is the interlayer exchange coupling surface energy density, M_s is the saturation magnetization, and δ is the thickness.

If the interlayer exchange field is greater than $4\pi M_s$, and without a spin polarized field, the magnetization will be damped eventually. However, with an AC polarized current, the generated spin torque can mitigate damping torque. At sufficient current density, θ can stay at a nonzero value, and the magnetization will precess at the angular frequency of ω_0. In uniformly magnetized FGL, when the spin torque is large enough, the spins at the interfaces between FGL and perpendicular layer (PL) are completely in-plane. As moving from the interface into the interior, the spins gradually align perpendicularly to the PL and form a half domain wall magnetization.

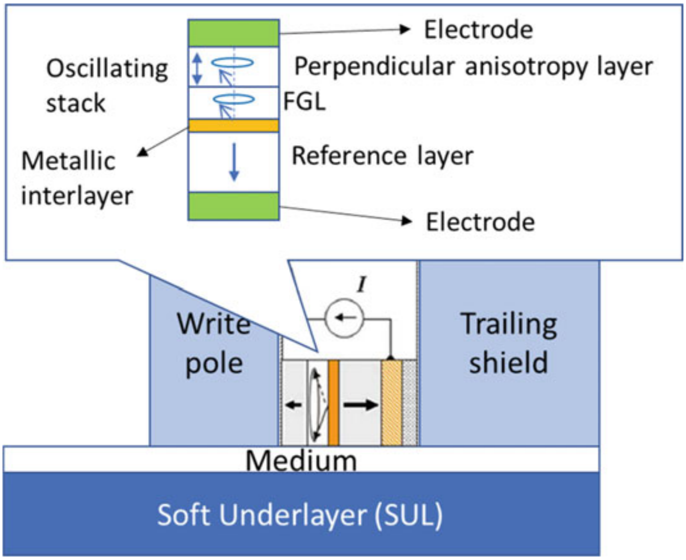


Fig. 7 Schematic illustration of the ac field-assisted perpendicular head design. The ac field generator drawing at the top has rotated 90° for better view. Drawn by author

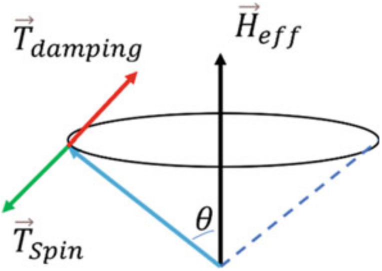


Fig. 8 Schematic of how a spin torque generates magnetic field by magnetization precession. Drawn by author

This formation of the magnetization in the PL will increase the effective recording head field [20].

More recent studies have explored the optimization of PMR writer and media design for MAMR. Basically, the MAMR-optimized media needs to have three higher damping than traditional PMR media to facilitate FMR, very low intergranular exchange coupling, and stronger layer-to-layer exchange coupling. Also, to writer at higher TPIC, the pole tip geometry also needs to be modified to improve lateral field gradient [21].

4.3 Heat-Assisted Magnetic Recording (HAMR)

Although MAMR has the advantage of relatively easier integration, its theoretical growth potential is limited compared to heat-assisted magnetic record (HAMR). Comparing to traditional technologies, HAMR is a revolutionary approach for aerial density increase: Instead of walking at the fine line between higher yet more localized magnetic field, HAMR, taking advantage of localized surface plasmon effect, uses a nanoscale "antenna" to focus laser into a hotspot much smaller than laser wavelength. The energy density is so high that the media within the hotspot is heated above its curie temperature, and thus allows writing with relatively lower magnetic power. The size of the written bit, which can be as small as nanometer size, is defined by the laser hotspot, rather than focused magnetic field. This will theoretically allow smaller bits to be written. Recent study showed that the HDD CAGR has been at the rate of 15% for a decade, but with the help of HAMR, the CAGR can return to 30% historical level at 2020s [22].

The basic idea of HAMR is depicted in Fig. 9. Basically, the media temperature is temporarily heated up above its curie temperature, and thus the media coercivity can be lowered below the applied magnetic field, allowing higher media anisotropy and smaller, meanwhile thermally stable grains against superparamagnetism. The heated area is then quickly cooled after the magnetic orientation is written into the media grains [23].

It is understood that the key of HAMR is the highly focused laser beam, which is a novel element in the HDD system, and requires several new components to be integrated. They are the light delivery system that can focus laser onto the bit, the thermomagnetic writer that can work at elevated temperature, a robust head-disk interface under such harsh working condition, and rapid cooling media that can lock the data, such as FePt [22]. This is different from MAMR, which can be built upon the current PMR system. An early example of the HAMR system is depicted in Fig. 10. It shows that a free-space laser beam is coupled into a waveguide on the trailing edge of the slider with a grating coupler [23]. Other than the write, the slider, gimbal assembly, air bearing, and the magneto-resistive reader can be borrowed from today's PMR drives.

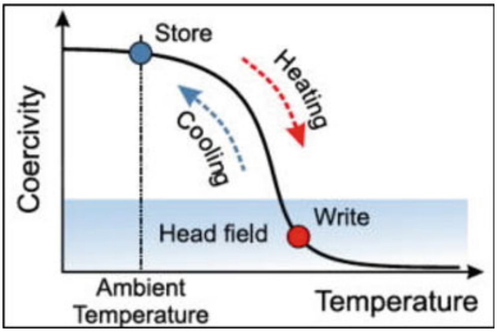


Fig. 9 Schematic diagram of the HAMR recording process [24]. Figure reprint permission obtained from publisher (open access article)

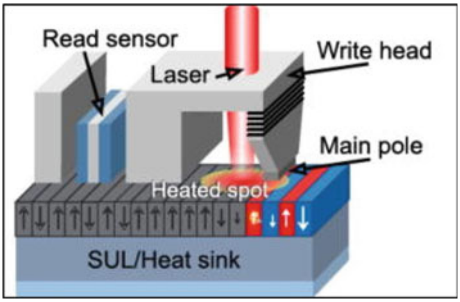


Fig. 10 Illustration of an early HAMR system example [24]. Figure reprint permission obtained from publisher (open access article)

The ability to record on very high-anisotropy materials can potentially increase the areal density by order of magnitudes over PMR, but several technical challenges must be addressed first. It is understood that optimizing thermal gradients and applied fields are the key to recording quality transitions. The light delivery system must produce spot sizes that are far below the optical diffraction limit to confine the thermal heating, and hence surface plasmon resonance (SPR), which is the resonant oscillations of surface charge at the dielectric-metal interface, is employed on top of optical focusing. SPR can be magnitudes stronger than the incident field, and can be much smaller than the incident wavelength. Such an SPR light focus apparatus is called near-field transducer (NFT), hence is the key to the success of HAMR.

As shown in Fig. 10, the field delivery structure has a magnetic core with a coil, where the write pole is patterned down to 100–300 nm cross-track width at the leading edge of the trailing pole. To write bits into the perpendicular medium, the edge of the waveguide (WG) core needs to be near the edge of the write pole (within 50 nm). On the other hand, since the magnetic materials interfere with propagation in the waveguide, it needs to be kept as far as possible from the core to prevent significant optical efficiency loss. Thus, the larger parts of the pole and the yoke are stepped away from the core.

A gold NFT with a sharp tip can couple light even more efficiently into a nearby medium, due to the combination of SPR and lightning-rod effect. An earlier NFT design ("Lollipop") and media are shown in Fig. 11. The NFT is located at the focal point of the laser. As resonance takes place, the surface charge oscillates along the length of the lollipop peg to generate an electric field at the tip of the peg, which couples energy into the media. A plasmonic metal beneath the recording layer acts as both a heat sink and an image plane for the electric field. The recording layer is effectively within the gap of two nanoparticles, the NFT and its image, resulting in a substantial enhancement in the coupling efficiency and further confinement of the electric field [14]. The incident light is polarized along the vertical axis of the NFT to excite the proper SPR.

The highly localized heat in HAMR has brought unprecedented challenges to drive integration. The most straightforward impact to drive mechanics is to address the protrusion induced by localized heat while maintaining an operable head-media spacing (HMS). For SPR to work, the HMS is believed to be no more than 5 nm [14].

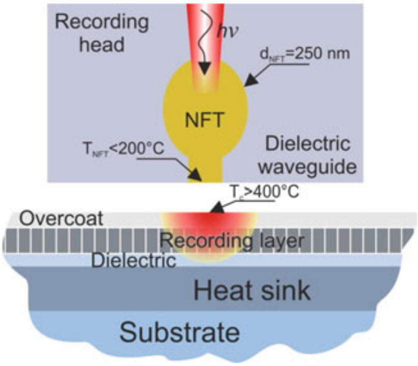


Fig. 11 Illustration of HAMR recording head (NFT), media, and head-media interface illustration of NFT temperature during writing process [24]. Figure reprint permission obtained from publisher (open access article)

Even in PMR, attaining such small spacing while the disk is spinning at 7200 RPM (or 20 m/s) is not a trivial challenge, whereas in the HAMR, the HMS control is further complicated by laser-induced protrusion, which has multiple scales in terms of both size and time. As the schematic of thermal protrusions at different times are shown in Fig. 12 [25], only a few tens of microseconds after laser turn on, a very localized protrusion occurs around NFT. This fast protrusion is typically referred to as NFT protrusion (Fig. 12a). The thermal energy diffuses from the NFT area to its surroundings over time, leading to a relatively larger lateral bulge in addition to the localized NFT protrusion. After a few 100 microseconds, the reader that is micrometers away also moves to the media surface (Fig. 12b). Meanwhile, the broaden NFT protrusion can eventually increase the air bearing pressure and renders extra lift force, which changes the baseline of the fly height (Fig. 12c). With milliseconds of laser-on time, the whole slider heats up and resulting in a crown and camber change that further changes the overall fly height. The overall fly height change can be optimized by air bearing design, and the laser diode-induced protrusion, due to its relatively longer time scale, can be compensated with the conventional thermal fly height control as in PMR. The localized NFT protrusion, however, is the most critical and most difficult to compensate for among all protrusions. The traditional contact detect-based protrusion measurement may be insufficient to fully characterize and compensate for the NFT protrusion. Therefore, more challenges have been brought up to properly set the writing clearance during drive operation.

Other than clearance, the NFT reliability is also challenged by the high operating heat, which can be increased over 300 °C during writing [22]. Previous study [18] indicated that NFT reliability is more a function of NFT temperature instead of media temperature. This, on one hand, demonstrates the feasibility of HAMR system that a NFT can be maintained functional even as the media must go over curie temperature, and also highlights the challenge of a carefully designed NFT that with minimal temperature rising and optimized thermal stability. Some good plasmonic materials,

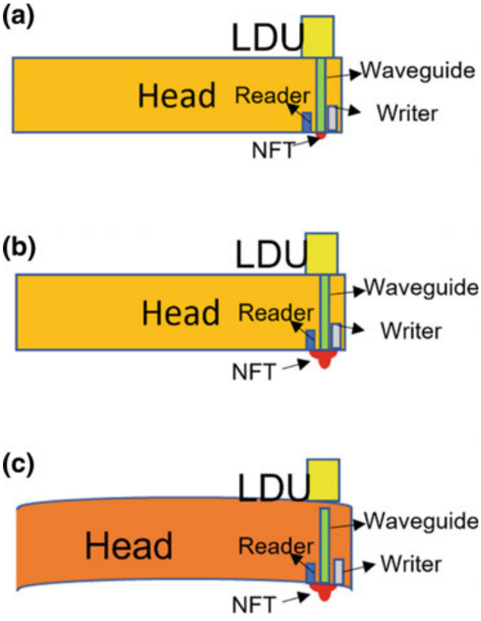


Fig. 12 Schematic of the laser-induced thermal protrusions at different time scales. **a** Localized NFT protrusion. **b** NFT broader protrusion. **c** Slider compensation. Drawn by author

such as Ag or even Au, can face significant thermal and mechanical instabilities in the HAMR system. The composition and crystal structure of NFT must be carefully engineered via thin film process and lithography. Some materials with less SPR coupling efficiency but better thermal stability may be included in NFT. Inevitably, there will be a trade-off between thermal power, or ADC, and reliability.

Besides NFT, other components in the recording head will be affected by high temperature too. For example, there are diamond-like carbon as tribological coatings on both head and media, and they are at risk of thermal decomposition too. At higher temperatures, carbon films are susceptible to micro-structural changes, turning from diamond-like to graphite-like, and thereby subjected to a degradation of film properties, such as hardness, wear resistance, etc. Among the solutions for improving the thermal robustness of head overcoats, the addition of species that form more thermally stable sp3 bonds shows some promise, as it can help prevent the graphitization of heat coatings.

The protective coatings at writer may be at greater risk, since it remains at high temperature throughout the writing process, whereas any media surface is heated only when data is written in that particular location [14]. For example, the high thermal gradient in the head induces strain gradients that challenge the mechanical integrity of the film. Surface defects and recording head materials with mismatched elastic moduli contribute to these locally high strains. As the recording heads are going through the rapid heating–cooling cycle for every bit recorded, which means terascale cycles during lifetime, leading to the potential for cyclic damage accumulation, or fatigue. Addition to thermal stability, the new coating material also should have a low optical extinction coefficient that can minimize optical loss at coating layers. Considering the difficulties in developing one coating that fits all the requirements, a patterned overcoating, in which materials and thicknesses are tailored for different functional regions of recording head, may provide the most optimized HAMR coating performances, though with increased manufacturing complexity [14]. For the media, an additional potential failure mode is the desorption of the media lubricant, which is a thin layer of perfluorinated polyether fluid that improves tribological performance. Localized heat may induce unwanted chemical reactions of interface contaminants, which makes it more difficult to maintain a clean interface that is required for a stable head-media interface.

With the excessive challenge in HAMR technology, the failure mode and effect analysis (FMEA) is a must in both design and process development. Inside HAMR recording head, the FMEA is focusing on key components, such as NFT/writer, reader, and resistive thermal wire. The function of each components has been illustrated in Sect. 3.2. As a standard FMEA process, the first step is the definition of failure mode. In HAMR, the two major failure modes are thermal-related failure and wear-related failure. The thermal stress is high in HAMR because of laser operation. The laser coupling and emission near NFT could bring up the local temperature to 300~400 °C and will also increase the temperature significantly in the components micrometers away from laser spot. As designed, many recording head components are made of metal thin film layers with complex structures. The high temperature will cause component failure via oxidation, plastic deformation, material separation, and voiding. Another failure mode is related to the mechanical wear stress. The laser heating introduces physical protrusion on the head surface, which is very difficult to be characterized. It generates a risk that the components on the recording head is in contact with media during operation. The contact between the head and the rotating media fails the component by mechanical wear or shocking stress. After identifying the failure mode, the next step is the reliability test to analyze the effect. There are several reliability tests routinely used on HDD HAMR recording head. One is the laser thermal life test. It is either a constant or cycled thermal stress by applying laser through NFT. The median lifetime and the cumulative failure rate at targeted life are the two specifications used for product qualification. Another reliability test for mechanical wear is the burnish test. It is facilitated by intentionally applying near zero or even into the media clearance to burnish the component and verify the component life over distance of burnish. The wear stress in the test could be constant or stepped stress to explore the onset of the degradation. With more maturity in the HAMR product development, the component lifetime is getting very long, and the accelerated life test (ALT) is often used. The thermal stress test can be accelerated by applying high laser current, which gives about 2.0 acceleration factor. The mechanical wear stress can be magnified by increasing the friction force between recording head and media, as well as environment temperature or relative humidity. The combined acceleration factor can usually reach 100. The reliability test will define the priority of each effect by considering the three factors together: severity, occurrence, and detection. Here, severity measures the hazard of effect, occurrence means how often the effect will happen, and detection indicates how easy the effect can be detected.

For example, some effects can be detected by small sample size in short verification test, but some effects will appear after long-term use in field. The last step is the action responding to the effects. There are two types of actions in FMEA process. One is preventive action, which is to minimize the likelihood of root cause. For example, if the new HAMR design fails the laser thermal life test, the design can be optimized by including more thermally robust material to prevent thermal degradation. Another type of action is adaptive. It is used to mitigate the hazard when effect or failure happens. For example, when characterizing laser-induced physical protrusion in clearance setting process, the test system is very difficult to maintain the good gage. As an adaptive action, there are a lot of parameters to be measured and working as safeguard in the test to mitigate the hazard from wear.

4.4 The Future of High-Volume Hard Disk Drive

Both MAMR and HAMR have shown potential in increasing recording density, and both adapted by the industry to develop the next generation hard disk drive. MAMR has the advantage of relatively easier integration, since it can be built upon the PMR system, as well as lower reliability risk due to little added heat to the system. But MAMR also faces the challenge of fabricating a consistent STO that can generate stable FMR. Of course, HAMR faces more challenges in reliability and integration difficulty, which could both increase the cost of HDD and make it less attractive over flash storage. In the long term, however, HAMR showed more promising potential for higher areal density, which is believed to be as high as 8 Tb/in^2, making a standard 3.5in HDD can store up to over 100 TB [26]. The potential of HAMR will be further unleashed by next generation HAMR specific media. For example, ordered granular media and bit-patterned media can further support narrower tracks and smaller bits.

Other than ADC and reliability, the overall performance of HDD will also need improvement to accommodate the increased storage data volume. For example, the IOPS-per-TB performance drops as the capacity of hard drives increases and using more actuators (the motor that controls the swinging of recording head gimbal over disk) instead of one can multiply the throughputs as well as improve IOPS-per-TB [26].

In conclusion, the HDD industry is harvesting the recent technology developments in, not only magnetic recording physics, but also optical, materials, mechanical engineering, and combined with improved manufacturing capability, to fulfill the future demand for high-volume data storage. It is believed that in the next several decades, HDD will remain as the preferred choice for massive data storage, such as data centers for autonomous vehicles, due to its relatively lower cost, better data safety, and more stable availability.

5 Summary

In the future AV ecosystem, the data center and its data storage are one of the most important elements. Considering the cost, production capacity, and performance trade-off, the HDDs are the best solution for large-scale data storage in data centers. To meet the exponential growth in data storage demand and to keep the relatively low $/GB, HDDs need significant improvement of areal density. Among the technology approaches in HDD industry, the most promising next generation technology is the HAMR. The success of HAMR will keep increasing the HDD areal density at the rate of 30% per year and lowering the total cost of ownership. This is making HDD a competitive data storage solution for the next decades to come.

Acknowledgements The editors would like to thank Mayue Xie from Microsoft and Weishen Chu from Western Digital for the critical reviews.

References

1. https://blocksandfiles.com/2020/02/03/autonomous-vehicle-data-storage-is-a-game-of-guesses/
2. https://www.forbes.com/sites/lanceeliot/2021/03/25/the-autonomous-vehicular-cloud-is-steering-into-view/?sh=724cd6167a7f
3. https://blocksandfiles.com/2021/02/26/hpe-sees-no-general-storage-need-to-switch-from-ssd-to-disk/
4. https://en.wikipedia.org/wiki/Data_center
5. https://www.cisco.com/c/en_in/solutions/data-center-virtualization/what-is-a-data-center.html
6. https://www.cablestogo.com/learning/library/data-center/data-center-basics
7. https://www.red-gate.com/simple-talk/sql/database-administration/storage-101-data-center-storage-configurations/
8. https://www.datacenterdynamics.com/en/opinions/continued-value-hdds-data-centers/
9. Mamun A.A., Guo G.X., Bi Ch., Hard Disk Drive_Mechatronics and Control
10. Antonis V., Andreas P., J. Phys. D: Appl. Phys. 43 (2010) 225301
11. Minoru S., Yoshihiro I., and Hiroyuki Y., Active Vibration Control of a Microactuator for the Hard Disk Drive Using Self-Sensing Actuation, Smart Materials Research, Vol 2012, Article ID 920747. https://doi.org/10.1155/2012/920747
12. A. Nordrum, The fight for the future of the disk drive, in IEEE Spectrum, vol. 56, no. 1, pp. 44-47, Jan. 2019
13. https://buffalotech.com/blog/cmr-vs-smr-hard-drives-in-network-attached-storage-nas
14. Kiely, J., Jones, P., & Hoehn, J. (2018). Materials challenges for the heat-assisted magnetic recording head–disk interface. MRS Bulletin, 43(2), 119–124
15. Marchon, T. Pitchford, Y. T. Hsia, and S. Gangopadhyay, The head-disk interface roadmap to an areal density of 4 Tbit/in^2, Adv. Tribol., vol. 2013
16. https://en.wikipedia.org/wiki/Magnetic_storage
17. J. C. Slonczewski, J. Magn. Magn. Mater. 159, L1 (1996)
18. T. D. Trinh, S. Rajauria, R. Smith, E. Schreck, Q. Dai and F. E. Talke, Temperature-Induced Near-Field Transducer Failure in Heat-Assisted Magnetic Recording. in IEEE Transactions on Magnetics, vol. 56, no. 6, pp. 1–4, June 2020

19. Zhang, M., Zhou, T., Yuan, Z., Analysis of Switchable Spin Torque Oscillator for Microwave Assisted Magnetic Recording, Advances in Condensed Matter Physics, vol. 2015, Article ID 457456, 6 pages, 2015
20. J. Zhu, X. Zhu and Y. Tang, Microwave Assisted Magnetic Recording, in IEEE Transactions on Magnetics, vol. 44, no. 1, pp. 125-131, Jan. 2008
21. M. Mallary et al., Head and Media Challenges for 3Tb/in^2 Microwave-Assisted Magnetic Recording, in IEEE Transactions on Magnetics, vol. 50, no. 7, 2014
22. Kief, M.T., Victora, R.H. Materials for heat-assisted magnetic recording. MRS Bulletin 43, 87–92 (2018)
23. M. H. Kryder et al., Heat Assisted Magnetic Recording, in Proceedings of the IEEE, vol. 96, no. 11, pp. 1810–1835, Nov. 2008
24. Weller, D., Parker, G., Mosendz, O., Lyberatos, A., Mitin, D., Safonova, N. Y., & Albrecht, M. (2016). FePt heat assisted magnetic recording media. Journal of Vacuum Science & Technology B, Nanotechnology and Microelectronics: Materials, Processing, Measurement, and Phenomena, 34(6), 060801
25. S. Xiong et al., Setting Write Spacing in Heat Assisted Magnetic Recording, in IEEE Transactions on Magnetics, vol. 54, no. 8, pp. 1–7, 2018
26. https://www.tomshardware.com/news/seagate-technology-roadmap-2021

Role and Responsibility of Hardware Reliability Engineer

Hualiang Shi and Lixin Jia

Abstract The role and responsibility of reliability engineers change with project milestones. During the design phase, reliability engineers define reliability targets, lead teams to review design weakness, brainstorm potential failure mode and root cause, define test plan, customize stress profile, allocate samples, and prepare for test program and test equipment. During the development phase, reliability engineers execute the test, analyze data by fitting life distribution and doing hypothesis tests, and drive failure analysis and corrective action. Once a product is released to market, reliability engineers work on field return and warranty analysis. This chapter covers some of these topics, including risk assessment methodologies (failure mode and effect analysis, fault tree analysis and stress-strength analysis), accelerated life testing and highly accelerated life testing, reliability statistics (sample size calculation, life distribution analysis by Linear least square regression and Maximum likelihood estimation, confidence interval calculation, hypothesis tests for mean and variance), failure analysis and corrective/preventive actions, system reliability metrics, reliability block diagram methods, and repairable system. Various case studies are used to illustrate the ideas, including cameras, cold plates, dash mount audio device, LED display, Lidar bracket, magnetic sensor, network and multimedia PCB boards, power supplies, Radar, and waterblock.

1 Introduction

The role and responsibilities of hardware reliability engineers have been evolving. If "hardware reliability engineer" is used as keyword for job search, different types of positions will be found. Depending on employers, the job openings might be Module Reliability Engineer (like ASIC, sensor, power electronics, battery cell, structure,

H. Shi (✉)
Palo Alto, CA, USA
e-mail: hualiang.shi@gmail.com

L. Jia
Nuro, Mountain View, CA, USA

Y. Li and H. Shi (eds.), *Advanced Driver Assistance Systems and Autonomous Vehicles*,
https://doi.org/10.1007/978-981-19-5053-7_13

Ecosystem), System Reliability Engineer (like Mac, iPhone), Operation Reliability Engineer, Sustaining Reliability Engineer, Design for Reliability, Supplier Reliability Engineer, Field Reliability Engineer, Reliability Scientist, etc. Although job titles and product types vary, the candidates are usually expected to cooperate with cross functional teams from the beginning (concept) of a product lifecycle to the end (field returns) of a product lifecycle with the following common role and responsibilities

- participate in design reviews
- define reliability spec and test plans with sample size
- execute accelerated life test
- analyze data by using statistical method
- facilitate failure analysis and understand physics of failure
- mitigate risk.

This chapter is organized to cover these topics from both theory and application points of view.

- The section “Risk assessment methodologies” reviews three common methodologies, including FMEA, FTA, and stress-strength analysis.
- The section “Accelerated life testing (ALT) and highly accelerated life testing (HALT)” discusses acceleration models and stress profile customization.
- The section “Reliability statistics” discusses sample size calculation methodologies, illustrates life distribution analysis by linear least square regression method and maximum likelihood estimation method, introduces various confidence level calculation approaches, and provides several hypothesis test methods for mean comparison and variance comparison.
- The section “Failure analysis (FA) and corrective/preventive actions (CAPA)” covers general FACA process flow and some structured problem-solving methods.
- The section “System level reliability” talks about system reliability metrics and reliability block diagrams of various system configurations.
- The section “Repairable system” discusses MTBF, MTTR, and availability briefly.

2 Risk Assessment Methodologies

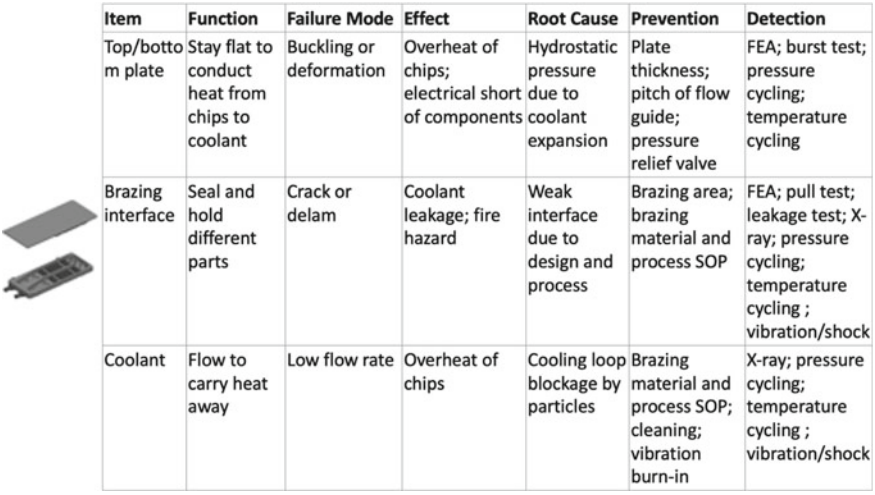
2.1 Failure Mode and Effect Analysis (FMEA)

As described in [1–4], FMEA is a bottom-up risk assessment method. When FMEA is performed on a design, it is called design failure mode and effect analysis (DFMEA); when FMEA is performed on a process, it is called process failure mode and effect analysis (PFMEA). AIAG VDA standard [3] proposes seven-step process flow, and SAE J1739 standard [4] proposes six-step process flow. The key steps include

- Review structure tree, block diagram, p-diagram, interface matrix, or process flow chart, which define the scope or boundary of FMEA, the components involved, and the interfaces between components.

- List primary functions of each component, including basic function, interface function, and safety function.
- Based on the function of each component, brainstorm potential failure modes, including intermittent failure, loss of function, degradation of function, and unintended function.
- Guess the effect of component failure, including local effect, next level effect, and end effect.
- Rank the Severity of failure effects. Depending on the company, the ranking scale can be 1–5 or 1–10. The ranking score increases with Severity. Generally, if the failure effect is related to customer safety or company brand name, the ranking score is highest.
- List the potential root causes for each failure mode. Methodologies like Fishbone Diagram and 5 Whys can be used to investigate the potential root causes.
- Review existing prevention control items, like design guideline, design standard, field lessons, Standard Operating Procedure (SOP), Incoming Quality Control (IQC), Outgoing Quality Control (OQC), In Process Quality Control (IPQC), and Preventive Maintenance (PM). These items are expected to reduce the occurrence of failures.
- Rank the Occurrence of failure based on the existing prevention control items. With the increase of effectiveness of prevention control items, the ranking score decreases. If the prevention control items are 100% effective, the ranking score is lowest, 1.
- Review existing detection control items, like Finite Element Analysis (FEA), tolerance analysis, design review, periodic maintenance inspection, inline inspection or test sampling, and reliability test plan and spec. These items are expected to improve the detection of failures.
- Rank the Detection of failure based on the existing detection control items. With the increase of effectiveness of detection control items, the ranking score decreases. If the detection control items are 100% effective, the ranking score is lowest, 1.
- Calculate Risk Priority Number (RPN) by using Severity (S) x Occurrence (O) x Detection (D). Usually, the item with the highest RPN has the highest priority. However, if an item has the highest Severity like customer safety, this item shall have the highest priority no matter what RPN looks like. Besides RPN, Action Priority (AP) can be obtained by combining S, O, and D.
- Recommend corrective actions with Direct Responsible Individual (DRI) and Estimated Completion Date (ECD) identified. By improving design, the severity and occurrence of failure can be reduced. By implementing new tests, detection capability can be enhanced.
- Once corrective actions have been approved and implemented, repeat the steps above to assess risk and prioritize risk again.

FMEA analysis can be carried out by using spreadsheet or commercial software like Relyence [5], ReliaSoft [6], Plato [7], and APIS IQ [8]. Figure 1 is a truncated DFMEA for the cold plate used inside a self-driving car [9]. Due to the massive



Item	Function	Failure Mode	Effect	Root Cause	Prevention	Detection
Top/bottom plate	Stay flat to conduct heat from chips to coolant	Buckling or deformation	Overheat of chips; electrical short of components	Hydrostatic pressure due to coolant expansion	Plate thickness; pitch of flow guide; pressure relief valve	FEA; burst test; pressure cycling; temperature cycling
Brazing interface	Seal and hold different parts	Crack or delam	Coolant leakage; fire hazard	Weak interface due to design and process	Brazing area; brazing material and process SOP	FEA; pull test; leakage test; X-ray; pressure cycling; temperature cycling ; vibration/shock
Coolant	Flow to carry heat away	Low flow rate	Overheat of chips	Cooling loop blockage by particles	Brazing material and process SOP; cleaning; vibration burn-in	X-ray; pressure cycling; temperature cycling ; vibration/shock

Fig. 1 DFMEA of cold plate

computing activities, thermal management is crucial for self-driving technology. Hybrid cooling with air cooling and liquid cooling is introduced. Cold plates are attached to CPUs, GPUs, and power systems. Liquid coolant circulates through cold plates and carries heat away from CPUs, GPUs, and power systems. The DFMEA revealed several failure modes. Design improvement was implemented to reduce the occurrence of failure. New testing methods were developed to enhance the detection of failure.

2.2 *Fault Tree Analysis (FTA)*

As described in [10], FTA is a top-down risk assessment method. It uses Boolean logic gates to combine various lower-level errors, failures, faults, and normal events which can trigger the top-level Undesired Event (UE). The key steps to do FTA are:

- Define the system by reviewing system design, manufacturing process, operation procedure, maintenance plan, functional diagrams, and reliability block diagrams.
- Define the top-level Undesired Event (UE) by reviewing hazard analysis, product requirement, and certification requirement.
- Construct a fault tree by using a cause-effect analysis. The cause at top-level failure is the effect at lower level. Boolean logic gates are used to combine various events. Different methods can be used for cause-effect analysis. For example, functional flow can be used backwards to identify the cause-effect.

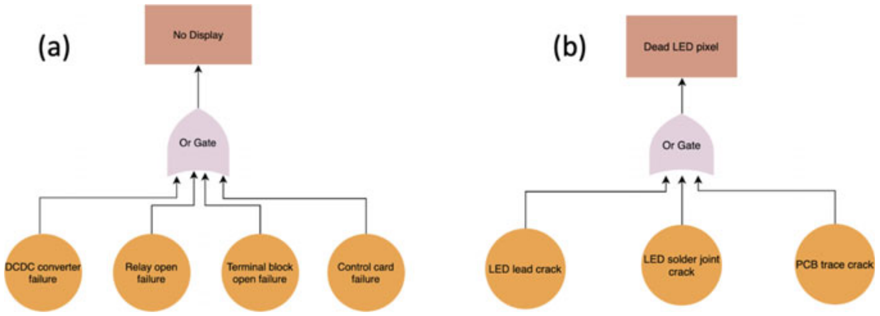


Fig. 2 FTA of Cartop advertising LED display

- Evaluate fault trees by doing qualitative analysis and quantitative analysis. In qualitative analysis, Cut Set (CS) is generated. In quantitative analysis, component failure rates are used to calculate the system failure probability.
- Control the hazard identified based on CS and probability.

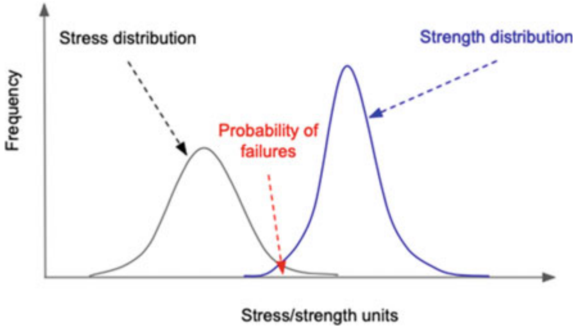
Cartop LED displays have been used to customize advertisements based on location and timing. In general, LED displays have multiple components in series, including DCDC converter, relay, terminal block, control card, LED panel with LED soldered to PCB board, etc. Figure 2a shows a FTA for failure modes “No Display,” and Fig. 2b shows a FTA for failure mode “Dead LED pixel.” These FTAs can help engineers deep dive into design details such as component spec to prevent failures and generate failure analysis debug procedures.

2.3 Stress-Strength Analysis

During product development and launch, risk analysis is often needed to quantify the reliability impact of different scenarios to facilitate decision-making. During the earlier phases of product development, there are usually multiple design options with different implications on cost, performance, reliability, and schedule. Reliability risks need to be quantified for each option in order to help the team select the best one. During or after product launch, quality excursions almost always happen that will negatively impact reliability. Such excursions can include material deviations, assembly line errors, supplier quality control problems, etc. Understanding the field reliability impact for the units built with these defective materials is critical in deciding whether to initiate field recalls, purge inventory, or accept the cost associated with higher field return rate.

To quantify the reliability impact, reliability engineers typically use the stress-strength interference model [11, 12]. The basic idea of the model is illustrated in Fig. 3. In the field, the stress experienced, e.g., temperature, pressure, mechanical shock, vibration, etc., is different from unit to unit and expressed as a distribution.

Fig. 3 Stress-strength interference model



This distribution typically comes from field data collection of the usage environment. Similarly, the reliability strength also varies from unit to unit due to material and assembly/installation process variations and again follows a distribution. This distribution normally comes from reliability testing data collected in the laboratory or at the manufacturing site, or warranty data. Whenever the stress experienced by a unit exceeds its strength, it would be categorized as a failure. This failure area is illustrated as the overlapped region of the two distributions.

Mathematically, there are two approaches to calculate the field failure rate using the model. First one is rather obvious by using Monte-Carlo simulation. In Monte-Carlo simulation, a random number (μ) between 0 and 1 is generated. Because Cumulative Distribution Function (CDF) is uniform between 0 and 1, Eq. (1) can be used to generate Weibull random sample (x) based on random number (μ) and distribution parameters (η, β). Similar approach can be used for other distributions.

$$F(x) = 1 - e^{-\left(\frac{x}{\eta}\right)^{\beta}} = \mu \tag{1}$$

One needs to generate a larger number of random samples following the distributions of both stress and strength. The values of stress and strength for each sample can then be compared to determine the failures. As long as the sample size is large enough (>100,000), the failure rate calculated should be sufficiently accurate.

The second approach involves analytical calculation using the following formula:

$$F = P\left[\text{Stress} \geq \text{Strength}\right] = \int_{0}^{\infty} f_{\text{Strength}}(x) \cdot R_{\text{Stress}}(x)\mathrm{d}x \tag{2}$$

where

F: failure rate
$P\left[\text{Stress} \geq \text{Strength}\right]$: probability of stress higher than strength
$f_{\text{Strength}}(x)$: PDF of strength distribution
$R_{\text{Stress}}(x)$: 1-CDF of stress distribution.

Once the field failure rate is quantified, it is always a good idea to review within the reliability team first, and then with the product team before sharing with management. This affords other team members the opportunity to challenge underlying data used and assumptions made to ensure the validity of the projection.

A mechanical vibration sensor with a design similar to Fig. 4 in a consumer electronic product is used as a case study for stress-strength analysis. It has two springs attached to the frame and a moving mass in the middle, which vibrates during field usage. The unit is designed to survive tens of millions of vibration cycles in the field. Given the long-term life requirement, it probably comes as no surprise that material fatigue-induced spring breakage is the dominant failure mode. During the development phase, the reliability engineer worked together with the mechanical design team to optimize the design geometry, material, and processes to address the fatigue failure and produced a final design that was capable of meeting the life requirements. However, just prior to launch, a material composition control issue was reported by one of the suppliers. As a result, millions of units already built were found to have a significantly higher fatigue failure rate. In order to determine if these units still have an acceptable field return rate, the reliability team was asked to perform a risk assessment.

First, based on past product information collected from the field, the number of cycles as a function of usage percentile is shown in Table 1.

By using this data, a reliability engineer can generate a statistical distribution that best fits this data. In this case, the Weibull distribution was found to fit the data well as shown in Fig. 5. This distribution has a shape parameter β of 2.16 and a scale parameter η of 45. This is the stress distribution with the pdf shown in Fig. 6.

Fig. 4 Mechanical sensor design

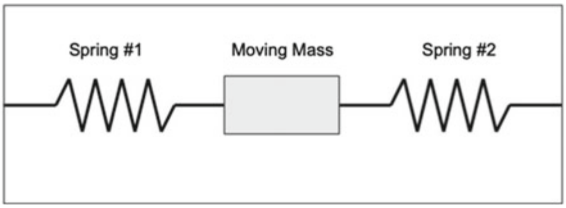


Table 1 Field usage (# of cycles) versus percentile

Percentile	# of cycles (in millions) for 1-year field usage
50	38
75	52
90	67
95	76
97.5	83
99	88

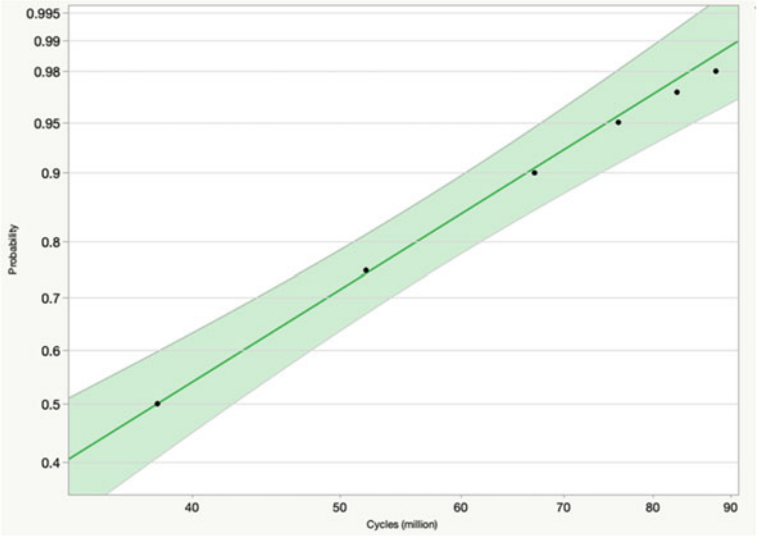


Fig. 5 Weibull distribution fit of field 1-year usage data with shape parameter $\beta = 2.16$ and scale parameter $\eta = 45$

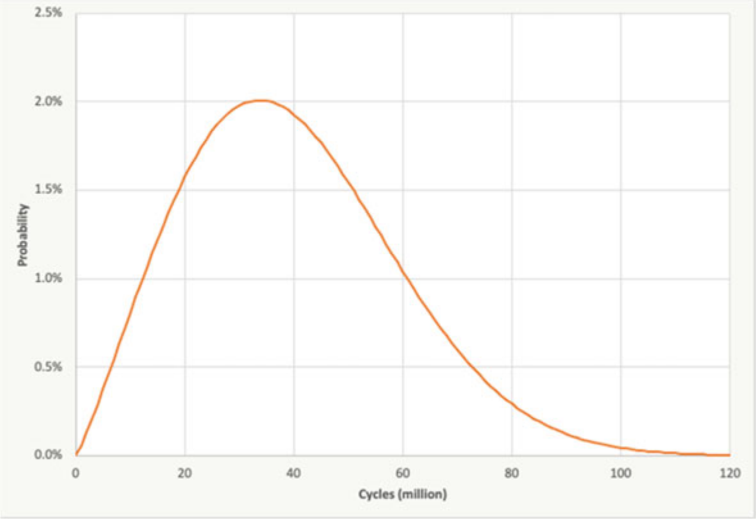


Fig. 6 Pdf of Weibull distribution with shape parameter $\beta = 2.16$ and scale parameter $\eta = 45$

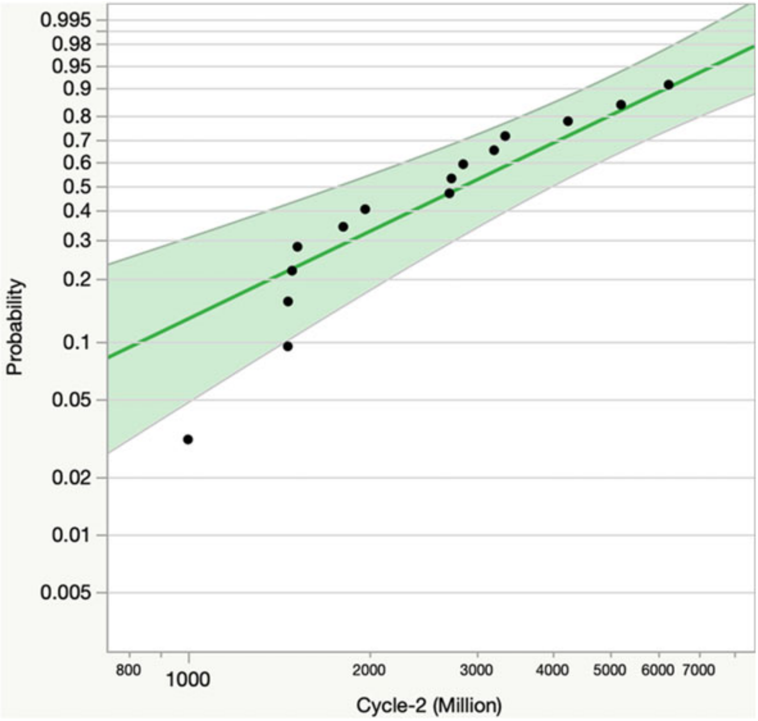


Fig. 7 Weibull distribution fit of life test failure data with shape parameter $\beta = 1.54$ and scale parameter $\eta = 3635$

Next, we fit the reliability life testing failure data to a Weibull distribution. As can be seen from Fig. 7, the fitting has a shape parameter β of 1.54 and a scale parameter η of 3635. This is the strength distribution.

Now, we can use Monte-Carlo simulation to calculate the field DPPM (defective parts per million) by creating a random stress profile using Weibull distribution with $\beta = 2.16$ and $\eta = 45$ and a random strength profile using Weibull distribution with $\beta = 1.54$ and $\eta = 3635$. Obviously, the sample size can be as large as we want. But in our case, we used a sample size of 100,000. The 1-year DPPM was estimated to be ~1100. One can use the closed form to do the calculation and get the exact DPPM of 1051. Similarly, one can get the 3-year DPPM to be 5650 and 5691, respectively, using Monte-Carlo simulation and closed-form calculation.

With this failure rate information, the management was able to decide that 1100 1-year DPPM is an acceptable risk and the product continued to ship with the defective material until a fix was implemented.

3 Accelerated Life Testing (ALT) and Highly Accelerated Life Testing (HALT)

3.1 Introduction

Accelerated Life Testing is one of the most important tools for reliability engineers to identify and resolve reliability issues earlier [11, 12]. During the product development phase, given the scheduling pressure faced by all technology companies, shortening the time required to perform the reliability testing and identify all the failure modes can be a rather significant competitive advantage. With well-designed Accelerated Life Testing, reliability growth can be achieved more efficiently, enabling faster product time-to-market, and at lower overall cost.

Taking corrosion as an example, the corrosion rate increases exponentially as temperature increases or follows power law as humidity increases. One can therefore use elevated temperature and humidity to identify corrosion-related issues quickly. But care must be taken to ensure the tests are properly designed to achieve the purpose of accelerated learning. For example, if a plastic with a T_g (glass transition temperature) of 80 °C is used as the enclosure material for a unit because the design team is confident that the design will not be subjected to any temperature close to 80 °C, one should not want to perform Accelerated Life Testing at 80 °C or higher.

The general process flow of designing an Accelerated Life Testing is shown in Fig. 8. Each of the steps will be covered in detail in the following sections, but a quick overview is shown below:

1. Identify field stress factors. This step is to determine what stress factors, e.g., temperature, humidity, mechanical vibration, pressure, etc., are the key drivers of the failure modes.
2. Determine stress levels. Appropriate levels of stress for each factor are critical for the success of the test. Under-stress would result in a low number of failures or no failures at all within reasonable test duration. Over-stress has the danger of creating artificial failure modes that do not exist in the actual user environment.
3. Select acceleration models. Depending on the stress factors, different acceleration models should be used to relate the results to product life in the field.

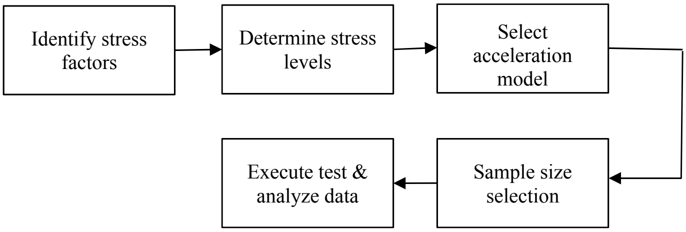


Fig. 8 Process of designing an accelerated life testing

4. Sample size selection. Sufficient number of units are needed at each stress level to produce enough failures for high confidence statistical analysis.
5. Execute the test and analyze the data. Stresses should be controlled and monitored to avoid surprises and erroneous data.

Highly Accelerated Life Testing (HALT) is another important reliability tool. By exposing samples to various stresses (i.e., thermal step, thermal shock, vibration step) sequentially or parallelly, the design operating limits and destruction limits can be revealed [13]. Due to the small sample size and non-constant stress, HALT is qualitative. This is different from Accelerated Life Testing which is quantitative. Team should do failure analysis to understand the root causes behind the failure modes revealed by the HALT test, and improve the design and process to eliminate the failure modes. By running HALT at different development milestones or phases, the team can monitor the type of failure modes to make sure existing failure modes are mitigated by new design or process and there is no new failure mode.

3.2 Identify Field Stress Factors

In order to achieve the goal of using accelerated testing to generate the failure modes in the field, one needs to first identify the main stressors. Typically, each critical failure mode is driven by just a small number of stressors. For example, corrosion is typically controlled by humidity, temperature, and chemical environment; solder joint fatigue is typically controlled by mechanical vibration and temperature variation.

Learnings from similar products in the past are a good source to help determine the main contributors. This is especially important for products with only incremental improvement. Even for a completely new product, a lot of times, similar technologies may have been used in other areas. Therefore, learning can be used to narrow down the list.

For truly revolutionary products with new technologies being used for the first time, FMEA (failure mode and effect analysis) can be very useful here. By brainstorming together with the key technical team members, one should be able to get a reasonably good idea about the key failure modes and their corresponding stressors. One word of caution for performing effective FMEA is to keep the audience to only the core technical experts. Otherwise, the whole effort can become a paper exercise of documenting a long list of everything imaginable, rendering the whole process useless.

Modeling work can also provide significant insight. By better understanding the failure modes and failure mechanism of each design, one should be able to determine the main stressors.

3.3 Determine Stress Levels

Once the stress factors are identified, the next step is to determine the level of stresses under field usage and testing.

Stress under field usage is dynamic and non-constant. Histogram is widely used to summarize the stress distribution, including temperature, temperature change, humidity, voltage, etc. Based on linear damage accumulation and histogram, dynamic mission profiles can be converted to effective constant stress profiles. As indicated in Fig. 9, over the same time interval, effective constant stress generates an equivalent amount of degradation as original dynamic stress [14]. As discussed in Sect. 3.4, lifetime and stress might follow power law, Arrhenius, and exponential. When lifetime and stress follow power law relationships (such as temperature cycle, relative humidity, voltage), Eq. (3) can be used. Subscript i means i_{th} stress bin in histogram, m is the total number of stress bin, and n depends on design and material. When lifetime and stress follow the Arrhenius equation (such as temperature), Eq. (4) can be used. E_{a} is activation energy, and K_{B} is Boltzmann constant. When lifetime and stress follow exponential lifetime relationships, Eq. (5) can be used.

$$\xi_{\text{effective}} = \left[\sum_{i=1}^{m} (\text{duty cycle})_i \times (\xi_i)^n\right]^{\frac{1}{n}} \tag{3}$$

$$\xi_{\text{effective}} = \frac{-\frac{E_{\mathrm{a}}}{K_{\mathrm{B}}}}{\ln\left[\sum_{i=1}^{m} (\text{duty cycle})_i \times e^{-\frac{E_{\mathrm{a}}}{K_{\mathrm{B}}\xi_i}}\right]} \tag{4}$$

$$\xi_{\text{effective}} = \frac{\ln\left[\sum_{i=1}^{m} (\text{duty cycle})_i \times e^{\gamma \xi_i}\right]}{\gamma} \tag{5}$$

The stress level under testing should be appropriate. The word “appropriate” here has several meanings:

1. It goes without saying that one cannot under-stress since it may not generate enough failures or have no failures at all. This would defeat the whole purpose of the test. While failures are hated by the design and process engineers, they

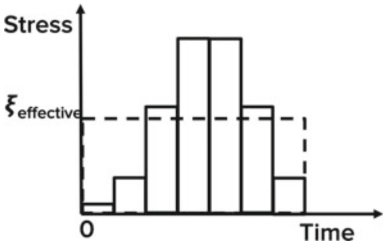


Fig. 9 Conversion of dynamic stress to effective constant stress

are immensely valuable for reliability engineers since that is the source of all the learnings.

2. Do not overstress since it may create artificial failure modes that will never happen in the field. For example, let us assume the compute module in a self-driving car uses CPUs with maximum junction temperature of 100 °C at which point they start to throttle. Let us further assume a temperature delta to the external ambient of 30 °C at maximum loading. It is obvious that during normal field operation, CPUs processing power should not be compromised. If the reliability engineer, however, uses 85 °C chamber temperature at maximum load and therefore increases junction temperature to 115 °C, it would generate an unrealistic failure mode of CPU throttling or even shutdown. The example is a rather easy situation to detect and avoid. Unfortunately, in the real world, most cases are quite hard to recognize. It really pays for the reliability engineer to check with the design and process teams first before starting any accelerated test.
3. Cost and product schedule are other factors that need to be considered when designing accelerated tests. In general, one should strive to accelerate as much as possible to shorten the test time. But equipment capability and other resource constraints would sometimes limit the options available to reliability engineers. Under such a scenario, careful consideration needs to be given to the test design so that easy execution can be maintained while under reasonable cost.

Several approaches can be used to determine the appropriate stress levels. Learnings from past products are obviously one source of knowledge that can be leveraged. This is especially true if there is enough information for one to estimate acceleration factors. Even if this is the first time the product is being developed, information from products using similar technology can also be valuable in providing a starting point. Another useful source is published literature on the same failure mode and material used. Doing a literature search before starting any accelerated test will usually pay off. One final method is to perform a pre-test DOE (design of experiment). One especially useful test technique is to perform a step-stress DOE where the stress levels are successively increased at predetermined stages in order to produce failure quickly. The data from this DOE would give a rough estimate of the acceleration factors at different stress levels which can then be used to design the constant stress-accelerated testing [15].

3.4 Acceleration Models and Acceleration Factor

Once the appropriate stress levels are determined, it is important to establish beforehand the proper acceleration models to relate the test results to end use conditions. This is critical to help define the test duration and interpret the test results. With the help of the model, one can then predict how the units will perform in the field, the levels of field failure expected, and the projected warranty cost.

At a high level, there are only two types of acceleration models: exponential and inverse power. Different types of stress require different acceleration models. For example, temperature always uses exponential acceleration. Humidity, mechanical vibration, and thermal cycling, etc., generally use an inverse power model. Voltage stress can be either depending on the underlying failure mechanisms.

Acceleration factor (AF) relates the lifetime of a design at two different stress levels. It is defined as:

$$\mathrm{AF} = \frac{L(\mathbf{ref})}{L(\mathbf{accel})} \tag{6}$$

where

- L(ref): the lifetime of the design at reference (e.g., end user) stress
- L(accel): the lifetime of the design at accelerated stress.

Once the acceleration model has been identified, it is fairly straightforward to calculate the AF. In the following sections, typical examples will be given to illustrate how each acceleration model is used and AF calculated.

3.4.1 Arrhenius Model

Temperature is one of the most common stress factors that affects the reliability of all products. To analyze the temperature effect, one would use the Arrhenius model. It was developed based on Arrhenius equation [16] that relates the rate of chemical reaction to temperature. For reliability purposes, the model is expressed mathematically as:

$$L(T) = Ae^{\frac{E_\mathrm{a}}{\mathrm{KT}}} \tag{7}$$

where

- $L(T)$: the lifetime of the design as a function of temperature T
- A: constant that is design specific
- K: Boltzmann constant (8.617×10^{-5} electron volts/degree Kelvin)
- T: absolute temperature in Kelvin
- E_a: activation energy in electron volts.

AF is then calculated to be:

$$\mathrm{AF} = e^{\frac{E_\mathrm{a}}{\mathrm{K}}\left(\frac{1}{T_\mathbf{use}} - \frac{1}{T_\mathbf{test}}\right)} \tag{8}$$

where

- T_test: the test temperature in Kelvin
- T_use: the field temperature in Kelvin.

As one can see, higher activation energy and higher temperature difference will produce larger acceleration factors. Activation energy depends on the nature of the underlying failure mechanism and thus is different for different designs, materials, and processes. It is a parameter that has to be either determined experimentally or estimated based on published literature.

3.4.2 Coffin-Manson Model

Thermal cycling is a standard reliability test that is adopted by a wide spectrum of industries to assess thermomechanical stress impact under changing temperature conditions. The stress is induced by CTE mismatch between mechanically connected parts. One of the typical failure types is solder joint cracks between BGA ICs and the PCB (printed circuit board).

To analyze thermal cycling acceleration results, one would use the Coffin-Manson model [17] described mathematically below:

$$N(\Delta T) = \frac{C}{(\Delta T)^m} \tag{9}$$

where

- $N(\Delta T)$: the number of cycles as a function of temperature delta ΔT
- C: constant that is design specific
- ΔT: delta between maximum and minimum temperature
- m: Coffin-Manson exponent.

AF is then calculated to be:

$$\mathrm{AF} = \left(\frac{\Delta T_{\mathbf{test}}}{\Delta T_{\mathbf{use}}}\right)^m \tag{10}$$

where

- ΔT_{test}: temperature delta during test
- ΔT_{use}: temperature delta during field use.

Higher temperature delta difference between test and usage condition as well as larger exponent value m will generate larger acceleration factors. The exponent value, typically measured experimentally, depends on design and material. For ductile materials like solder, m ~ 1–3 [18].

3.4.3 Basquin Model

Basquin model, another one of the inverse power types, is typically used to predict mechanical fatigue-induced failures. Fatigue can happen whenever a structure is subjected to cyclic stress or strain conditions. This failure mode should be considered in the reliability testing for most of the automotive components due to constant mechanical vibrations experienced although the risk level varies significantly from component to component due to mounting location, method, and design.

Basquin's equation describes how to calculate the number of cycles a component can survive under constant stress S:

$$N(S) = CS^{\frac{1}{b}} \tag{11}$$

where

- $N(S)$: the number of fatigue cycles as a function of stress S
- C: constant that is design specific
- S: reversing stress magnitude
- b: fatigue strength exponent. It is always negative.

AF is then calculated to be:

$$\mathrm{AF} = \left(\frac{S_{\mathbf{use}}}{S_{\mathbf{test}}}\right)^{\frac{1}{b}} \tag{12}$$

where

- S_{test}: stress amplitude during test
- S_{use}: stress amplitude during field use.

Higher stress delta between test and field condition as well as larger absolute exponent value $\frac{1}{b}$ will generate larger acceleration factors. The exponent value, typically measured experimentally, differs for different designs and materials [19].

3.4.4 Multi-stress Models

Some of the failure modes will be driven by multiple stress factors. In such cases, the test needs to be designed with all the stresses accelerated. The mathematical formula for life and acceleration factor can be combined for individual stress to predict field life performance.

Let us take corrosion as an example which can be accelerated by both temperature and humidity. Temperature acceleration follows Arrhenius model, and humidity acceleration follows inverse power law. The formula for life calculation thus becomes:

$$L(T, \mathrm{RH}) = Ce^{\frac{E_a}{\mathrm{KT}}}\left(\frac{1}{\mathrm{RH}}\right)^m \tag{13}$$

where

- $L(T, \mathrm{RH})$: the lifetime of the design as a function of temperature T and RH (relative humidity in percent)
- C: constant that is design specific
- K: Boltzmann constant (8.617×10^{-5} electron volts/degree Kelvin)
- T: absolute temperature in Kelvin
- E_a: activation energy in electron volts
- m: humidity exponent.

AF is then calculated to be:

$$\mathrm{AF} = e^{\frac{E_\mathrm{a}}{K}\left(\frac{1}{T_\mathbf{use}} - \frac{1}{T_\mathbf{test}}\right)} \left(\frac{\mathbf{RH_{test}}}{\mathbf{RH_{use}}}\right)^m \tag{14}$$

where

- T_test: the test temperature in Kelvin
- T_use: the field temperature in Kelvin
- $\mathrm{RH}_\mathrm{test}$: the test relative humidity in percent
- RH_use: the use relative humidity in percent.

Same logic applies if there are more than two stressors, but it is quite rare to combine three or more factors together.

3.5 Case Study

3.5.1 Vibration Profile Customization

As shown in Table 2, different industry standards [20, 21] have different vibration profiles; within the same industry standard, vibration profiles depend on locations of modules inside vehicles. The best approach to define vibration profiles is to do road testing with accelerometers attached to locations right next to the modules of interest. When instrumentation data is not available, industry standards can be leveraged and scaled to unblock engineering development.

Table 2 Random vibration profiles for passenger car

Location	Standard	
	ISO 16750-3	GMW3172
Sprung mass	Grms: 27.8 m/s^2 Duration: 8 h per axis	Grms: 19.6 m/s^2 Duration: 16 h per axis
Unsprung mass	Grms: 107.3 m/s^2 Duration: 8 h per axis	Grms: 107.3 m/s^2 Duration: 16 h per axis

Figure 10 is an example of road testing in the bay area of north California. Testing conditions included some urban streets at San Francisco (SF) at speed of 35 MPH, the Highway 280 between Palo Alto and San Francisco at speed of 65 MPH, and a cobblestone street at speed of 15 MPH. A Tri-axis accelerometer was attached to the roof of a vehicle. The three graphs in the first row showed time-domain acceleration data of three orthogonal directions (X, Y, Z). The three graphs in the second row showed frequency-domain Power Spectral Density (PSD) data. T. Irvine provided a series of free tutorials and scripts about converting time-domain acceleration data to frequency-domain PSD data [22]. At low frequency region, rough roads (cobblestone street, SF urban street) generated higher PSD; at high frequency region, smooth road (highway) generated higher PSD. Because PSD curves are related to road conditions and vehicle speeds, it is recommended to customize a random vibration profile based on Operational Design Domains (ODD).

Figure 11 showed the procedures of customizing a random vibration profile for accelerated life testing at laboratory [9]. In step 1, power spectral density (PSD) curves in the field were obtained by road testing at GoMentum Station, which is a

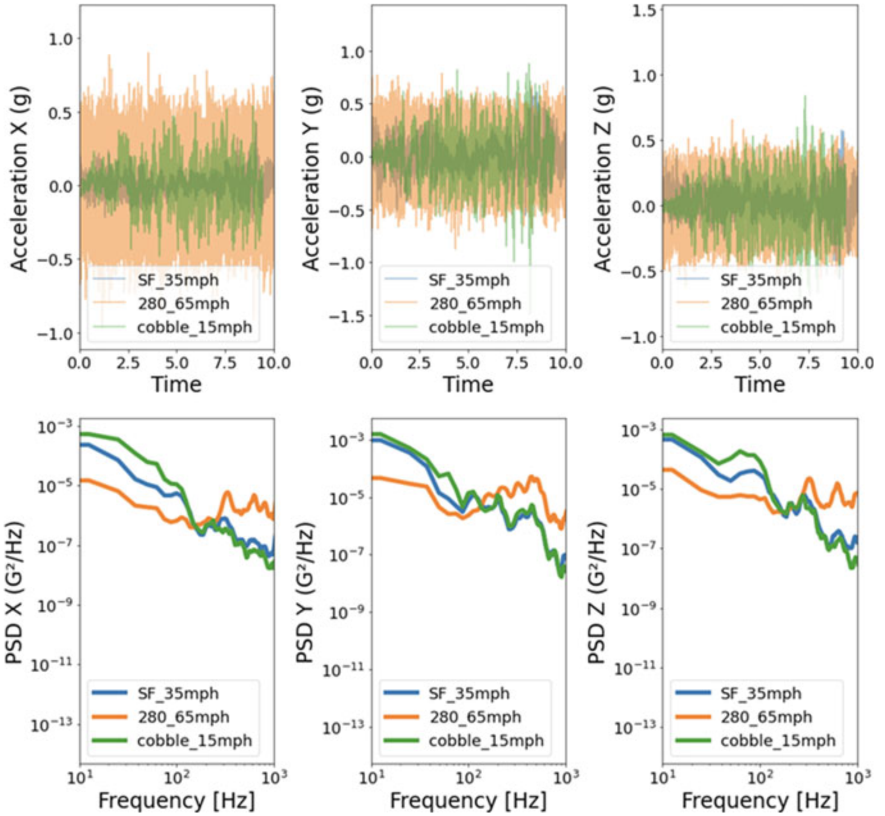


Fig. 10 Road instrumentation for vibration profile customization

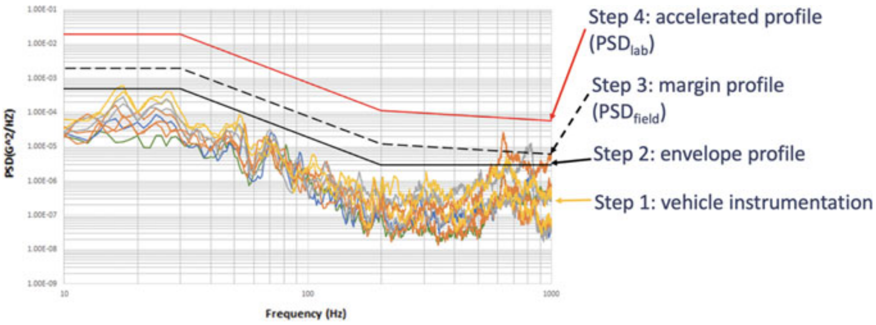


Fig. 11 Procedure of customizing random vibration profile

testing ground located in Concord, California, United States. In step 2, an envelope profile was drawn to envelope all of the field PSD curves. In step 3, a margin profile ($\mathrm{PSD}_{\mathrm{field}}$) was obtained by adding 3 dB to the envelope profile. This 3 dB margin was used to account for variation among vehicles, modules, and roads. Some reasons for this include (1) wear down of suspension components over time can cause an increase in input, (2) variation in mounting of module may cause a shift in input, (3) variation in mass of the module, and (4) occasional roads that may be worse than what were captured by instrumentation test data. In step 4, an accelerated profile ($\mathrm{PSD}_{\mathrm{lab}}$) was obtained by using a certain acceleration factor (AF) and MIL-STD-810G, 514.6A-5 method.

Assuming

(1) the device is ON for 11 h a day, 5 days a week, 52 weeks per year, and 2 years. The total operation hour in the field (t_{field}) is 5720 h.
(2) the random vibration test time at the laboratory (t_{lab}) is 2 h per axis.
(3) fatigue exponent (m) is 6.

Equation (15) was used to convert the margin profile ($\mathrm{PSD}_{\mathrm{field}}$) to accelerated profile ($\mathrm{PSD}_{\mathrm{lab}}$). The accelerated profile can be used for random vibration tests at the laboratory. Grms is the root-mean-square acceleration, which is the square root of the area under the PSD versus frequency curve.

$$\mathrm{AF} = \frac{t_{\mathrm{field}}}{t_{\mathrm{lab}}} = \left[\frac{\mathrm{PSD}_{\mathrm{lab}}}{\mathrm{PSD}_{\mathrm{field}}}\right]^{\frac{m}{2}} = \left[\frac{\mathrm{Grms}_{\mathrm{lab}}}{\mathrm{Grms}_{\mathrm{field}}}\right]^{m} \tag{15}$$

3.5.2 Temperature Cycling Profile Customization

As indicated in Table 3, different industry standards [21, 23] have different temperature profiles. Instead of using industry standards, the temperature profile can be customized based on Operational Design Domains (ODD) and vehicle design. When

Table 3 Temperature cycling profiles for passenger car

Type	Standard	
	ISO 16750-4	GMW3172
Temperature cycle	Temperature change: $T_{\text{min}} \langle - \rangle T_{\text{max}}$ Number of cycles: 30	Temperature change: $T_{\text{min}} \langle - \rangle T_{\text{max}}$ Number of cycles: 211, 248, 309
Thermal shock	Temperature change: $T_{\text{min}} \langle - \rangle T_{\text{max}}$ Number of cycles: 100, 300	Temperature change: $T_{\text{min}} \langle - \rangle T_{\text{max}}$ Number of cycles: 632, 927, 100

the device is OFF, the temperature change of the device is due to ambient temperature change and solar loading; when the device is ON, the temperature change of the device is due to ambient temperature change, solar loading, and the temperature rise induced by device self-heating. When a device is available, instrumentation with thermocouples and built-in temperature sensors inside chip and PCB board under different vehicle operation modes (like Idle and AC On) can be used to customize temperature mission profile. When a device is not available, thermal simulation and historical environmental data (i.e., the National Centers for Environmental Information (NCEI) at National Oceanic and Atmospheric Administration (NOAA) [24]) can be used to estimate temperature change.

Assuming

(1) the operational design domain of a device covers Concord, Livermore, Palo Alto, and San Francisco. Figure 12 shows the daily ambient temperature change distributions for these areas from 2010 to 2017–2019. 25 °C can be used to envelope the 99th percentile of daily ambient temperature change for this ODD.

(2) the device is ON for 5 days a week, OFF for 2 days a week, 52 weeks per year, and 2 years. The total number of temperature cycles experienced by devices in the field (N_{field}) was approximated as 728 cycles (one cycle per day). The temperature rise induced by self-heating when the device was ON was estimated to be around 20 °C. Table 4 summarizes the daily temperature change (ΔT), number of cycle, and duty cycle of the two states of device (“ON” and “OFF”).

(3) the weakest link is the solder joint ($n = 1.9$).

(4) the temperature change during temperature cycle test at laboratory (ΔT_{lab}) is 90 °C. The extreme low temperature (T_{min}) of the temperature cycle test at the laboratory is −10 °C. The extreme high temperature (T_{max}) of the temperature cycle test at the laboratory is 80 °C. These T_{min} and T_{max} were bound by the component capabilities and material properties inside the device. Accelerated stress should not introduce stress artifacts.

Equation (3) was used to convert the dynamic temperature cycle mission profile to an effective constant temperature cycle mission profile. Subscript “i” means different

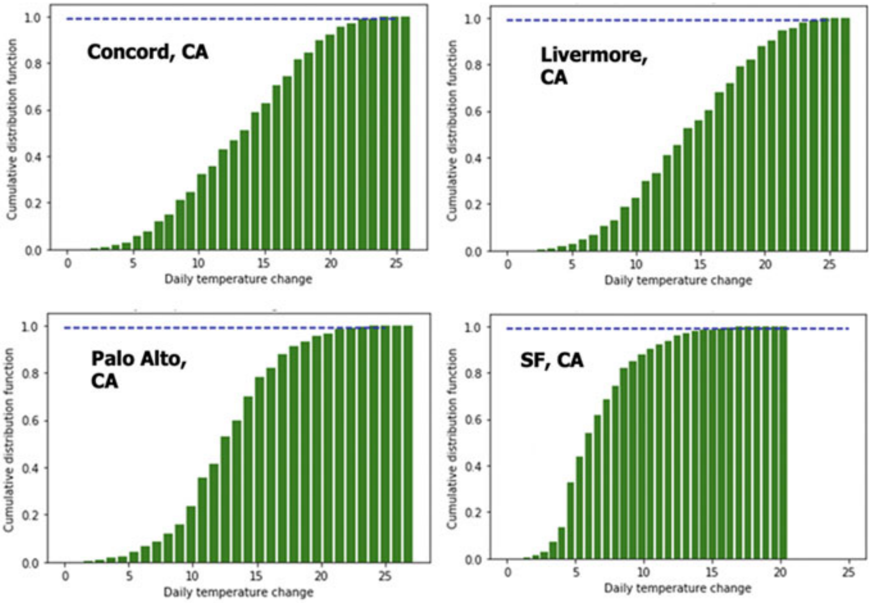


Fig. 12 Daily ambient temperature change

Table 4 Device temperature change and duty cycle in field

Device	ΔT (°C)	Number of cycle	Duty cycle (%)
OFF	25	208	28.57
ON	45	520	71.43

device states (e.g., "ON," "OFF" when $m = 2$). The effective constant daily temperature change of device in field (ΔT_{field}) was calculated to be 40.2 °C. After substituting ΔT_{field}, ΔT_{lab}, N_{field}, and n into Eq. (16), the number of temperature cycle at laboratory (N_{lab}) was calculated to be 158 cycles. 158 cycles can be used as a temperature cycle test spec for this system.

$$\text{AF} = \frac{N_{\text{field}}}{N_{\text{lab}}} = \left[\frac{\Delta T_{\text{lab}}}{\Delta T_{\text{field}}}\right]^n \tag{16}$$

Once $T_{\min}$, $T_{\max}$, and Number of Cycle have been defined, a reliability engineer needs to further define ramp rate and dwell time. Ramp rate should be proper to avoid water condensation during chamber ramping up from $T_{\min}$ to $T_{\max}$. Dwell time should be long enough such that the device can reach $T_{\min}$ and $T_{\max}$.

Assuming

(1) a device has mass (m) 5 kg, specific heat capacity (c_{p}) 890 J/Kg K, surface area (A) 0.8 m^2, and air convection coefficient (h) 10 W/m^2 K.

(2) the device has been soaked at $T_{\min}$ (−10 °C) for a long time.

(3) the chamber was ramped up and reached $T_{\max}$ (80 °C) within a short duration.
(4) due to thermal mass, the device will take some time to heat up. The acceptable minimum final device temperature (T_{final}) during ramp up is 75 °C.

Based on Newton's law of cooling, Eq. (17) can be derived. t is the time the device takes to ramp up from $T_{\min}$ to T_{final}. By using the values above, $t = 27$ min. So the minimum dwell time at $T_{\max}$ is recommended to be 27 min.

$$t = \frac{mC_{\text{p}}}{hA}\text{Ln}\left[\frac{T_{\max} - T_{\min}}{T_{\max} - T_{\text{final}}}\right] \tag{17}$$

3.5.3 High Temperature Operating Life (HTOL) Profile Customization

In Sect. 3.5.2, 99th percentile of daily ambient temperature change of weather station data was used to customize temperature cycling profile. The pros behind this approach are adding safety margin into the design, and the cons are overstressing. If failure occurs and cannot be resolved by design or process mitigation, reliability spec should be revisited to enable product shipment. In this subsection, instead of using 99th percentile, the histogram and duty cycle of weather station data were used to customize HTOL profile.

Assuming

(1) a device is ON for 11 h a day (7:00 AM to 18:00 PM), 365 days per year, and 2 years. The total operation hour in the field (t_{field}) is 8030 h. The temperature rise induced by self-heating when the device was ON was estimated to be around 20 °C.
(2) the operational design domain covers Palo Alto. Table 5 shows the histogram summary of ambient temperature at Palo Alto based on NOAA weather station data 2012–2021 from 7:00 AM to 18:00 PM [24].
(3) the weakest link has activation energy $E_{\text{a}} = 0.7$ eV.
(4) HTOL test at the laboratory has $T_{\text{lab}} = 85$ °C.

Equation (4) was used to convert the dynamic temperature mission profile to an effective constant temperature mission profile $T_{\text{field}} = 39.3$ °C. After substituting T_{field}, T_{lab}, t_{field}, and E_{a} into Eq. (18), the test duration at laboratory (t_{lab}) was calculated to be 292 h. 292 h can be used as a HTOL spec for this system.

$$\text{AF} = \frac{t_{\text{field}}}{t_{\text{lab}}} = e^{\frac{E_{\text{a}}}{K_{\text{B}}}\left(\frac{1}{T_{\text{field}}} - \frac{1}{T_{\text{lab}}}\right)} \tag{18}$$

Table 5 Histogram of ambient temperature at Palo Alto based on NOAA weather station data 2012–2021 from 7:00 AM to 18:00 PM

T (°C)	Duty cycle (%)
2.4	0.04
5	0.41
7.5	1.35
10.1	6.39
12.7	15.44
15.2	11.94
17.8	23.09
20.4	14.43
23	17.12
25.5	5.1
28.1	3.22
30.7	0.82
33.2	0.49
35.8	0.11
38.4	0.04
40.9	0.02

3.5.4 UV Profile Customization

In-vehicle infotainment (IVI) systems are usually installed next to the vehicle dashboard. They are used to deliver entertainment and vehicle system information to driver and passenger through audio and video interfaces. UV radiation is one of the potential risks for devices next to the dashboard. During the summer season, devices surrounding dashboard location in certain cities can have temperatures higher than 80 °C. The ASTM G154 UV test condition is widely adopted in industry [25]. Condition “Cycle 4” uses lamp UVA-340 with irradiance 1.55 W/m^2/nm and approximate wavelength 340 nm, 8 h UV at 70 °C black panel temperature, and 4 h condensation at 50 °C black panel temperature. UV radiation test time at the laboratory ($t_{\text{UV-A}}$) can be calculated by using Eq. (19).

$$t_{\text{UV}-\text{A}} = \frac{P_{\text{Solar}} \times p_{\text{UV}-\text{A sunlight}} \times T_{\text{UV}-\text{A}}}{P_{\text{UV}-\text{A lamp}}} \tag{19}$$

where

(1) P_{Solar} is the solar insolation on the earth ground. The average daily insolation on the earth ground is approximately 6 kWh/m^2 = 21.6 MJ/m^2 [26]. Assume the lifetime is two years. The total solar insolation (P_{Solar}) is 21.6 MJ/m^2 × 365 × 2 = 15,768 MJ/m^2.

(2) $p_{\text{UV-A sunlight}}$ is the percentage of UV-A light within sunlight. At earth ground level, 3% of sunlight is ultraviolet (UV). Among the UV radiation reaching Earth's surface, more than 95% is UV-A [27]. So $p_{\text{UV-A sunlight}}$ ~3%.

(3) $T_{\text{UV-A}}$ is the UV-A transmittance through the vehicle windshield. Based on Wachler's study, car windshields blocked about 96% of UV-A rays on average [28]. So $T_{\text{UV-A}} = 4\%$.

(4) $P_{\text{UV-A lamp}}$ is the power of a UV-A lamp. Based on the UVA-340 irradiance profile [29] and ASTM G154 Cycle 4 condition, $P_{\text{UV-A lamp}}$ is estimated to be 77.5 W/m^2.

By substituting P_{Solar}, $p_{\text{UV-A sunlight}}$, $T_{\text{UV-A}}$, and $P_{\text{UV-A lamp}}$ into Eq. (19), the UV radiation test time at laboratory ($t_{\text{UV-A}}$) was estimated to be 68 h or 9 cycles under ASTM G154, Cycle 4.

3.5.5 Vibration Test of Network Switch

Steinberg did extensive studies about the impact of PCB board deflection on electronic components' fatigue lifetime [30]. As indicated in Eq. (20), Steinberg found that when the three-sigma displacement at the center of PCB board which is perimeter supported ($Z_{3\sigma}$) is less than a critical value ($Z_{\max}$), the lifetime of component is expected to be about 20×10^6 stress reversals under random vibration environments.

$$Z_{\max} = \frac{0.00022 \times B}{C \times h \times r \times \sqrt{L}} \tag{20}$$

where B is the length of PCB edge parallel to component (in "inch"), L is the length of electronic component (in "inch"), h is the thickness of PCB board (in "inch"), C is a constant depending on the type of electronic components, and r is a factor depending on the relative position of component on PCB board.

Shi et al. [9] used this approach to study the impact of different vibration profiles on the lifetime of off-the-shelf (OTS) network switch boards. Figure 13a showed the network switch board. Key concern is the big IC in the middle of the board in the red circle. Based on parameters ($B = 9.71''$, $L = 1.228''$, $h = 0.061''$, $r = 1$, $C = 1.75$), Steinberg $Z_{\max}$ is 0.018 in. (0.458 mm). Figure 13b presented three random vibration profiles, "Lyft," "GMW," and "ISO." The vertical axis is displacement power spectral density (DPSD). This Lyft profile was customized based on one operational design domain and vehicle model. GMW profile [21] and ISO profile [20] were sprung mass.

Table 6 summarized FEA 3σ displacements and Miner's cumulative fatigue damage ratio. When GM profile and ISO profile were used, FEA $Z_{3\sigma}$ were larger than Steinberg's $Z_{\max}$. Hence, this board might not survive 20 million stress reversals in random vibration under GM or ISO profile. The cumulative damage ratios (R) were all less than 1. ISO profile generated the largest amount of damage, and Lyft profile generated least amount of damage.

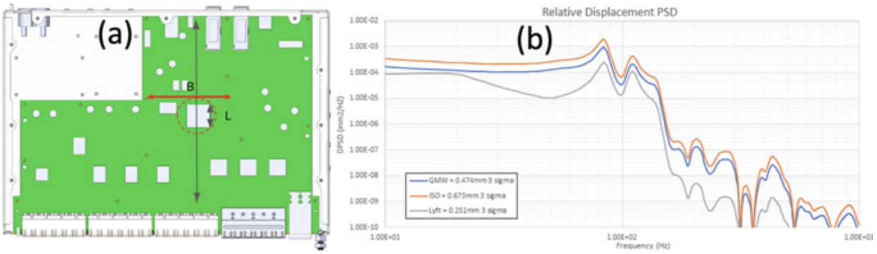


Fig. 13 **a** Network switch board for random vibration study and **b** random vibration profiles

Table 6 Random vibration results summary

Profile	Steinberg's $Z_{\max}$ (mm)	FEA $Z_{3\sigma}$ (mm)	Miner's damage index
ISO 16750	0.458	0.673	0.22
GMW3172	0.458	0.474	0.04
Lyft	0.458	0.251	0.00039

3.5.6 Shock Test of Network and Multimedia PCB Boards

When robo-taxi is running in the field, there are different shock events, including collision, door closure/slamming, hard brake, pothole, speed bump, vehicle over curb stone, etc. As indicated in Fig. 14, out-of-plane shock test was used to down select off-the-shelf (OTS) multimedia and network boards.

Table 7 summarizes three shock profiles. The first shock profile "Lyft" was based on road instrumentation with accelerometers from certain operational design domains and vehicle model. The second shock profile "GM" is the GMW3172 mechanical

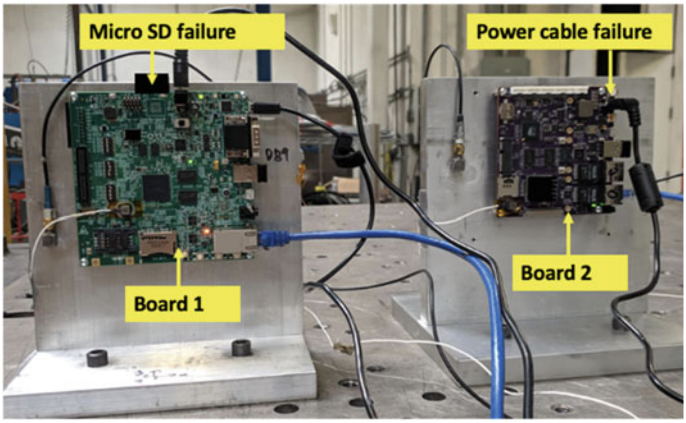


Fig. 14 Out-of-plane shock test of network and multimedia PCB boards

Table 7 Various shock profiles

Profile	Shape	Amplitude (G)	Duration (msec)
Lyft	Half sine	9	25
GM	Half sine	25	10
ISO	Half sine	50	6

Table 8 Failure rate post-shock test

Board type	Lyft profile	GM profile	ISO profile
Type 1	0F/20	0F/20	11F/40
Type 2	0F/20	1F/20	10F/20

shock—pothole profile at location “Sprung masses (all other areas, including Cradles and Frames”) [21]. The third shock profile “ISO” is the ISO 16750-3 mechanical shock at location “on the body and on the frame” [20]. During the shock tests, boards were operational.

Table 8 summarized the out-of-plane shock test results. “xF/y” means when boards went through “y” number of shock tests, “x” number of shock tests failed. Board type 1 had a micro-SD card I/O error. During out-of-plane shock tests, the PCB board has out-of-plane deflection. Because this micro-SD card is located in the middle of the board edge and away from mounting holes, it is expected to have large deflection at that location. Hence, the metal pins of the micro-SD card might have poor contact with the pins of the micro-SD connector or socket, inducing intermittent electrical open failure. Board type 2 had self-rebooting failure, which was suspected to be related to the barrel jack power cable connection issue. In general, the failure rates increased with pulse amplitude. For example, “Lyft” profile had zero failure and the “ISO” profile had the highest failure rate.

3.5.7 Mechanical Sensor Fatigue

Let us return to the same mechanical sensor in Sect. 2.3. Now, we want to increase the number of cycles in the field for new user applications. We are willing to reduce the vibration magnitude to compensate for the cycle count increase in order to use the same design. In this case, we need to understand the acceleration factor due to stress change because of vibration amplitude change.

To do that, we designed an experiment where 8 units each were tested at 100% and 65% vibration magnitude. In order to accelerate the failures, the environmental condition was set at high temperature and high humidity to complete the DOE in a reasonable amount of time (more on this in the next section). Table 9 showed the failure cycle counts for each magnitude. Note, for the column named F/P (fail/pass), a value of *P* means the unit never failed and thus the data is right censored.

To analyze the data, we need to determine two factors: life distribution and life–stress relationship. In this case, we use Weibull as the life distribution, and Basquin

Table 9 Mechanical sensor failure cycle counts for different vibration magnitude

Vibration magnitude	Cycle counts (million)	F/P	Vibration magnitude	Cycle counts (million)	F/P
65%	1139	P[a]	100%	43.7	F
	1111.7	P[a]		94.9	F
	1111.8	P[a]		87.6	F
	436.1	F		143	F
	283.4	F		122.6	F
	172	F		131.7	F
	70.4	F		119.4	F
	1116.5	P[a]		68	F

[a] Right censored data

model as the life–stress relationship since we know the failure mechanism is fatigue. Using the software to do analysis, we get a fatigue strength exponent value of −0.176 (Fig. 15). With this information, we can easily estimate the factor of expected life improvement at reduced vibration amplitude (see Table 10). This would then enable the trade-off management between the number of cycle counts and the reduction in vibration magnitude.

3.5.8 Mechanical Sensor Temperature and Humidity Stress

As mentioned in the previous section, the fatigue failure mode of the mechanical sensor in Sect. 2.3 is accelerated by high temperature and high humidity due to stress corrosion cracking (SCC). Understanding the acceleration factors for both stresses is critical in relating reliability test data at high temperature and high humidity to normal user conditions. To do that, we designed a three-cell DOE as listed in Table 11 to separate the two factors.

After the DOE is completed, in order to do the analysis, we need to determine the life–stress relationship first. We will use the Arrhenius model for thermal and inverse power for humidity. The AF (acceleration factor) is calculated based on Eq. (21):

$$\mathrm{AF} = e^{\frac{E_\mathrm{a}}{K}\left(\frac{1}{T_\mathrm{l}}-\frac{1}{T_\mathrm{u}}\right)}\left(\frac{\mathbf{RH}_\mathrm{u}}{\mathbf{RH}_\mathrm{l}}\right)^m \qquad (21)$$

where

- T_u: the upper temperature in Kelvin
- T_l: the lower temperature in Kelvin
- RH_u: the upper relative humidity in percent
- RH_l: the lower relative humidity in percent.

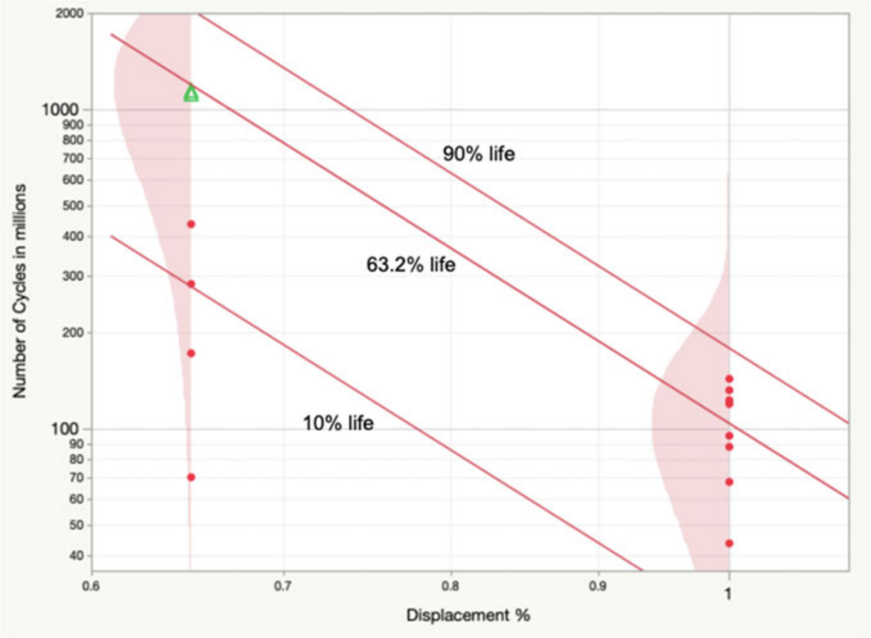


Fig. 15 Mechanical sensor life versus vibration amplitude analysis using Weibull-inverse power relationship. Fatigue exponent b is estimated to be −0.176

Table 10 Life improvement factor versus vibration magnitude

Vibration magnitude (%)	Factor of life improvement
65	11.5
70	7.6
75	5.1
80	3.5
85	2.5
90	1.8
95	1.3
100	1

Table 11 Temperature and humidity DOE

DOE cell	Temperature (°C)	Relative humidity (%)
#1	40	90
#2	55	90
#3	55	70

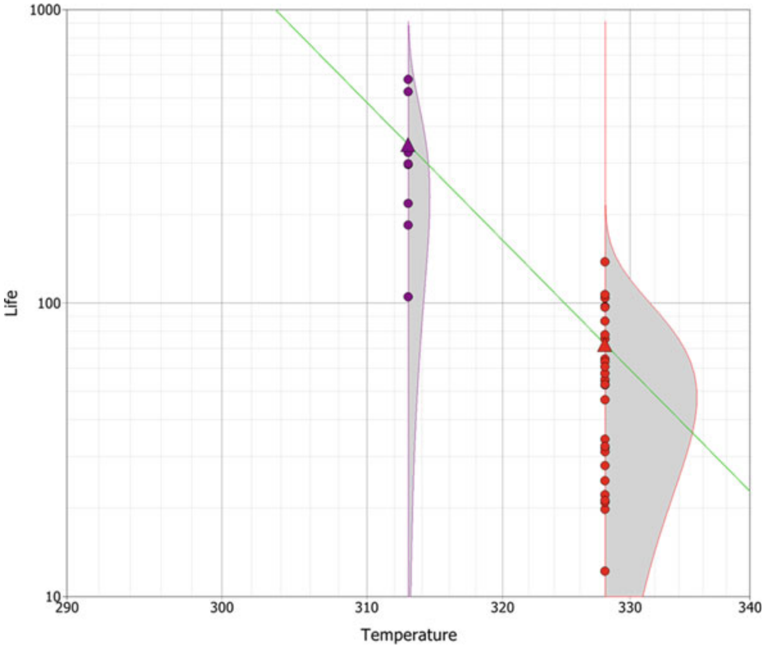


Fig. 16 Temperature acceleration with $E_a = 0.92$ eV

The only unknowns are now thermal activation energy E_a and relative humidity (RH) exponent value m. Using software to perform the analysis, we get $E_a = 0.92$ eV and $m = 4.84$. Figures 16 and 17 illustrate the temperature and humidity dependence graphically. Note, the temperature values in the calculation are all in absolute temperature (Kelvin).

With this information, we now can calculate the AF for any temperature and humidity combination. Table 12 shows the AF values for several different conditions.

3.5.9 Thermal Step Test of Radar

When sensors are deployed in the field, they might experience cold weather and hot weather. Thermal step test was applied to check the Radar performance under different ambient temperatures. As indicated in Fig. 18, Radar was loaded into a thermal oven and a reflector was placed at a certain distance away from oven/Radar. Once Radar reached the desired temperature inside the oven, the oven door was opened, and short-duration data collection was recorded. Number of detections was used as characteristics. This procedure was repeated under different temperatures. When temperature was above 45 °C, false detection occurred, which was ascribed to the Radar design issue.

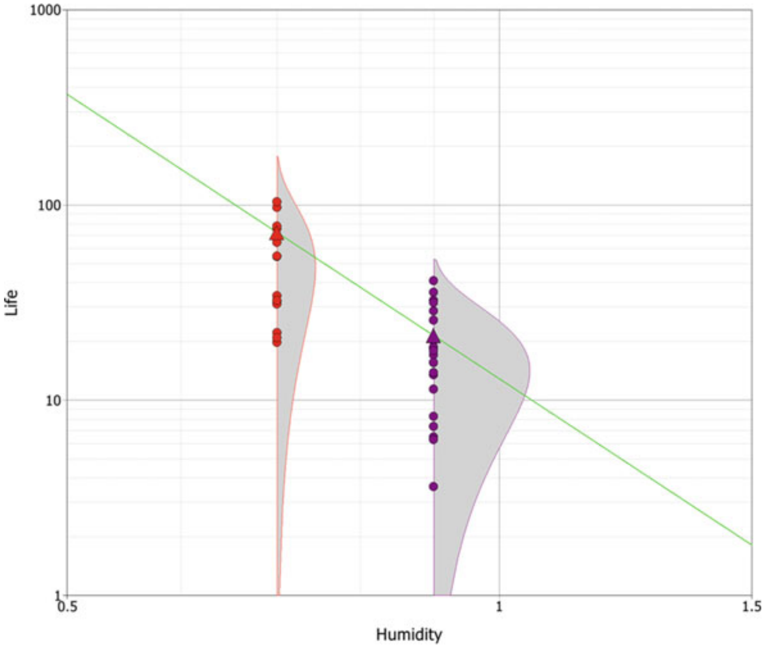


Fig. 17 Relative humidity acceleration with $m = 4.84$

Table 12 AF for different temperature/humidity combinations

Temperature (°C)	Relative humidity (%)	AF (acceleration factor)
40	90	1
40	70	3.4
25	90	5.6
25	70	18.8
25	50	95.8

3.5.10 Voltage Step Test of LED Panel

To understand the voltage operating limit of certain LED panel designs, a voltage step test was carried out. Each single LED module inside the LED panel has red pixel, green pixel, and blue pixel. Based on design, the nominal input voltage into the LED panel is 5 V. In Fig. 19, the LED panel on the left was operated at 5 V and used as reference; the LED panel on the right had input voltage spanned from 1.6 to 11.6 V. At low voltages (≤1.6 V), the LED panel on the right cannot be powered on. When input voltage was between 1.7 and 4.1 V, the LED panel on the right had abnormal output (i.e., discoloration). When input voltage was between 4.2 and 6.9 V, the LED panel on the right had normal output. When input voltage was between 7

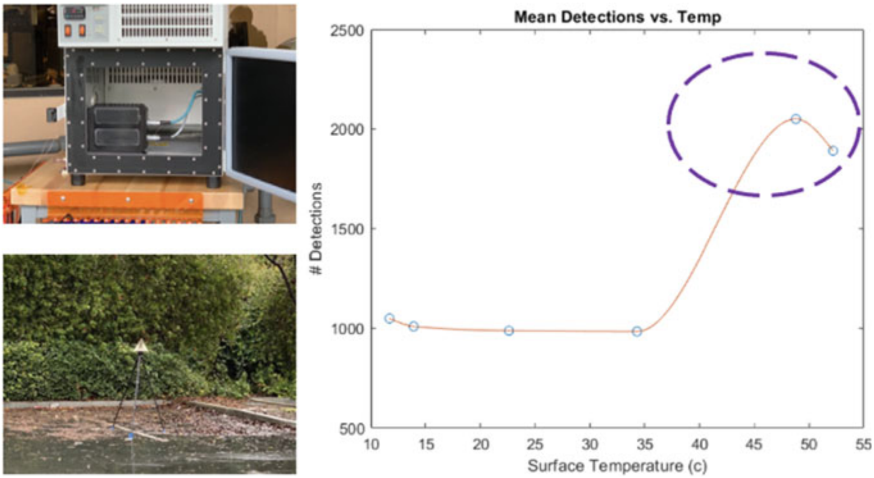


Fig. 18 Thermal step test of Radar

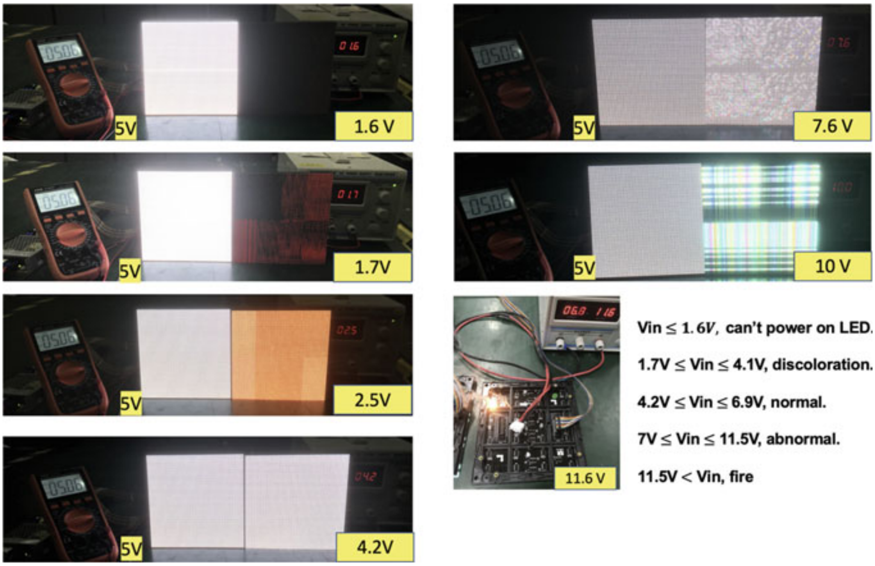

5V
1.6 V
5V
1.7V
5V
2.5V
5V
4.2V
5V
7.6 V
5V
10 V
11.6 V
Vin ≤ 1.6V, can't power on LED.
1.7V ≤ Vin ≤ 4.1V, discoloration.
4.2V ≤ Vin ≤ 6.9V, normal.
7V ≤ Vin ≤ 11.5V, abnormal.
11.5V < Vin, fire

Fig. 19 Voltage step test of LED panel

and 11.5 V, the LED panel on the right had abnormal output (i.e., defect line). When input voltage was above 11.5 V, the LED panel on the right had a fire and smoke. This DOE validated the LED panel operating limit.

4 Reliability Statistics

4.1 Sample Size

A test plan is composed of stress type, stress level, test duration, sample size, and pass/fail criteria. Previous subsections discussed stress type, stress level, and test duration. This subsection will focus on sample size calculation. Table 13 summarizes several classic methods for sample size calculation. Due to the page limitation, instead of deep diving into the math behind each method or equation, we will focus on the applications of some of these methods.

4.1.1 Non-parametric Binomial Sample Size

When the life distribution model parameters are unknown and test results are either pass or fail, Non-Parametric Binomial sample size can be used. When zero defect or failure is allowed, it is called Zero Defect Sampling (ZDS). As shown in Fig. 20, for the same reliability (R), sample size (n) increases with confidence level (CL); for the same confidence level, sample size increases with reliability; for the same sample size, reliability increases when confidence level decreases.

A chipmaker is going to run a temperature cycle test (−55 °C/125 °C/1000 cycle) to qualify a new epoxy underfill material. Functional test is either pass or fail. To demonstrate product reliability (R) 97% (or failure rate $F = 3\%$) with 50% confidence level (CL) (R97C50), what is the minimum sample size (n) assuming zero failure?

By substituting F and CL into the Non-Parametric Binomial ZDS sample size Eq. (22), n is calculated to be 23.

$$n = \frac{\ln(1 - \mathrm{CL})}{\ln(1 - F)} \tag{22}$$

Similarly, the sample size for 90% reliability with 60% confidence level (R90C60) is calculated to be 9 and the sample size for 99% reliability with 90% confidence level (R99C90) is calculated to be 230.

4.1.2 Parametric Binomial Sample Size

In Fig. 21, n_0 samples without failure are used to demonstrate reliability R_0 at lifetime t_0 with confidence level CL, n_1 samples without failure are used to demonstrate reliability R_1 at lifetime t_1 with confidence level CL, and n_2 samples without failure are used to demonstrate reliability R_2 at lifetime t_2 with confidence level CL.

If the life distribution follows Weibull distribution, Eqs. (1) and (22) can be combined to get

Table 13 Sample size calculation methods

Method	Formula	Note
Non-parametric binomial	$1 - \mathrm{CL} = \sum_{i=0}^{k} \binom{n}{i} F^{i} (1 - F)^{n-i}$ for $k = 0$, zero defect sampling (ZDS) $n = \frac{\ln(1-\mathrm{CL})}{\ln(1-F)}$	CL is confidence level n is sample size k is the number of failures F is the failure rate
Parametric binomial	$n_0 = \frac{\ln(1-\mathrm{CL})}{\ln(1-F)}$ $n_1 = n_0 \left(\frac{t_0}{t_1}\right)^{\beta}$	n_0 is ZDS sample size to demonstrate a failure rate F at time t_0 with confidence level CL n_1 is ZDS sample size at time t_1 β is known Weibull shape parameter, >0
Exponential Chi-square	$n = \frac{\mathrm{MTTF} \cdot \chi^{2}_{1-\mathrm{CL},2k+2}}{2T}$ or $n = \frac{\chi^{2}_{1-\mathrm{CL},2k+2}}{2F}$	n is sample size CL is confidence level. k is number of failure. MTTF is mean time to failure F is failure rate $\chi^{2}{}_{1\text{-CL},\, 2\,k+2}$ is the inverse of the right-tailed probability (1-CL) of the chi-square distribution with degree of freedom $2\,k + 2$ T is test duration
Poisson zero defect sampling (ZDS)	$n = \frac{1}{F} \ln\left[\frac{1}{1-\mathrm{CL}}\right]$	n is sample size CL is confidence level. F is failure rate
Cochran	$n_{\text{unlimited}} = \frac{z^{2} F(1-F)}{\varepsilon^{2}}$ $n_{\text{limited}} = \frac{n_{\text{unlimited}}}{1+\frac{n_{\text{unlimited}}-1}{N}}$	z is standard normal percentile for (1 + CL)/2 CL is confidence level. F is failure rate ε is margin of error. $n_{\text{unlimited}}$ is sample size when population size is unlimited n_{limited} is sample size when population size is limited

$$n_1 = n_0 \left(\frac{t_0}{t_1}\right)^{\beta} \tag{23}$$

$$n_2 = n_0 \left(\frac{t_0}{t_2}\right)^{\beta} \tag{24}$$

Based on Eq. (23), by increasing test time from t_0 to t_1 ($t_1 > t_0$), sample size is reduced from n_0 to n_1 ($n_1 < n_0$). Based on Eq. (24), by increasing sample size from n_0 to n_2 ($n_2 > n_0$), test time is reduced from t_0 to t_2 ($t_2 < t_0$).

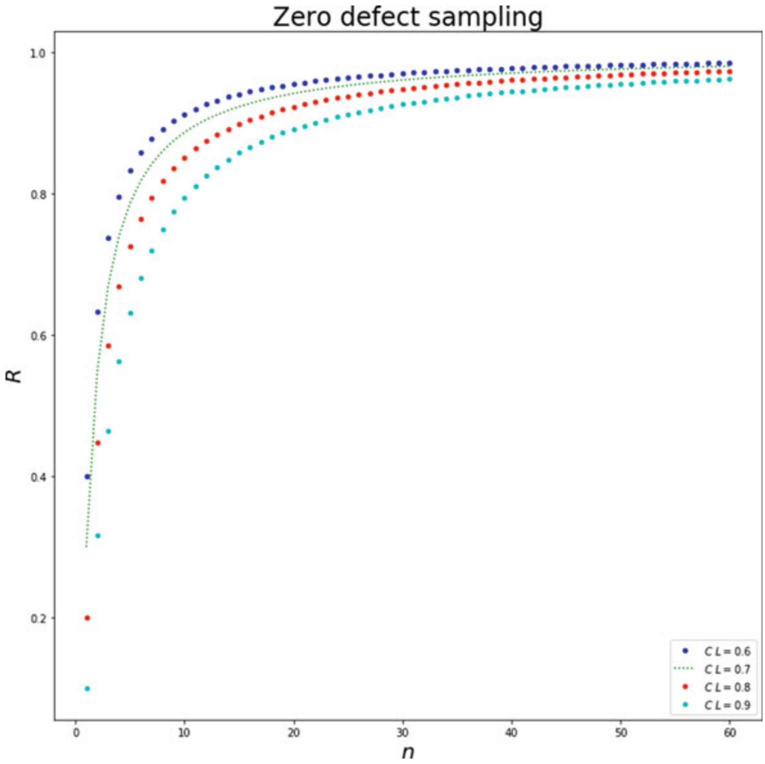
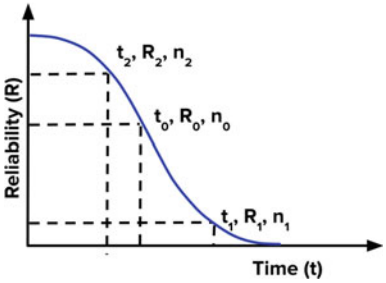


Fig. 20 Non-parametric binomial zero defect sampling

Fig. 21 Parametric binomial zero defect sampling



Let us revisit the example in Sect. 4.1.1. To demonstrate R97C50 at 1000 cycles, 23 samples without failures are needed. If the temperature cycle chamber can only load 15 samples, what is the new number of temperature cycles, assuming $\beta = 2$? By substituting $n_0 = 23$, $n_1 = 15$, $t_0 = 1000$, and $\beta = 2$ into Eq. (23), new test duration $t_1 = 1238$ cycles.

Table 14 Sample size to demonstrate 100 FIT

AF	1	10	100
n	9163	917	92
T_{lab} (hour)	1000	1000	1000

4.1.3 Exponential Chi-Square Sample Size

Frequently, reliability engineers are asked to provide sample size to demonstrate certain FIT or MTTF (or MTBF) for random hardware failure. Exponential Chi-square sample size can be used for this purpose.

To demonstrate a safety–critical IC meets ASIL-B requirements (or FIT < 100 for random failures in a field with 60% confidence level (CL = 0.6) and zero failure (f = 0)), what is the sample size (n) per stress type?

By using Eq. (25) and FIT value, MTTF can be calculated. By using Eq. (26), the product of n and t_{field} can be calculated. By using Eq. (27), test time at the laboratory can be converted to time in field. Sample sizes with various acceleration factors (AF) and fixed test time at the laboratory are provided by Table 14. When AF is increased, sample size is reduced.

$$\text{MTTF} = \frac{1000{,}000{,}000}{\text{FIT}} = 10{,}000{,}000\,\text{h} \tag{25}$$

$$n \times t_{\text{field}} = \frac{\text{MTTF} \times \chi 2_{1-\text{CL},2f+2}}{2} = 9{,}162{,}908 \tag{26}$$

$$t_{\text{field}} = t_{\text{lab}} \times \text{AF} \tag{27}$$

4.1.4 Approach to Mitigate Small Sample Size

To improve statistical confidence and cover variations from process and material, larger sample size for reliability tests is preferred. However, the BOM behind self-driving cars is expensive. Frequently, reliability engineers are asked to reduce sample size to save cost. Different methodologies have been reported [9].

First, DFMEA is a good tool to lead a team to review design and process weakness, brainstorm the potential failure mode and root cause, and architecture reliability test plan. Material allocation for reliability tests can be prioritized based on the RPN ranking or Action Priority.

Second, instead of using individual samples for different reliability stress tests in parallel, the same samples can go through a series of reliability tests in sequence. The advantage behind this sequential waterfall test is less sample size needed, while the disadvantage behind this sequential waterfall test is complicated data interpretation needed when failure occurs.

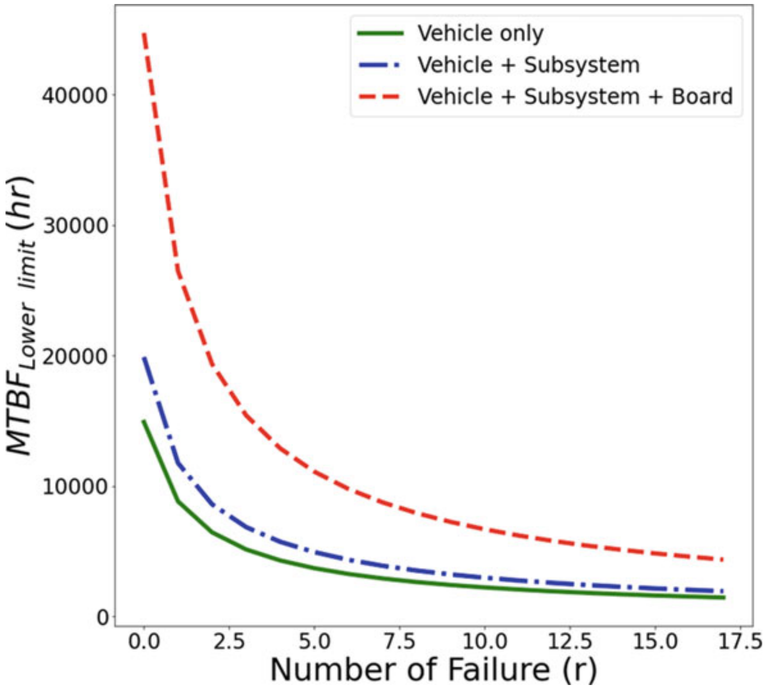


Fig. 22 Effect of mixed sample size on MTBF estimation

Third, reliability tests can be done at different levels. By lumping test data from system level (like vehicle), subsystem level (like sensing system on roof and computing system inside trunk), and module level (like camera, Lidar, Radar, PCB board, etc.), effective sample size is increased. In Fig. 22, by combining samples across various levels, the demonstrated MTBF is improved.

Fourth, reliability tests can be executed at different hardware development phases. By summing data from proto build, engineering validation test (EVT) build, design validation test (DVT) build, production validation test (PVT) build, and mass production (MP) build, effective sample size is increased.

Fifth, stress-to-fail methodology can be used. Once samples pass the customized reliability test, they can be further stressed under industry standards such as GM standard and ISO standard to reveal more potential failure modes.

Sixth, Finite Element Analysis (FEA) can be used for life analysis. During design, FEA can be executed to investigate cumulative damage and check design safety margin.

4.2 *Life Distribution Analysis*

Once a reliability test plan has been executed, a reliability engineer needs to deep dive into the testing result for life distribution analysis. Nowadays, there are many sophisticated commercial statistical softwares for life distribution analysis, including Relyence [5], ReliaSoft [6], JMP [31], etc. Although these softwares are very powerful, there is a need for reliability engineers to understand the fundamentals behind life distribution analysis. So reliability engineers can interpret the analysis result correctly and customize scripts based on special needs by using open-source programming languages like Python [32] and R [33] or commercial programming languages like Excel VBA [34] and JMP JSL [31]. In this subsection, life distribution analysis by Linear least square regression and Maximum Likelihood Estimation (MLE) are briefly discussed. For details, please refer to Nelson [35, 36], Yang [37], Bertsche [38], and O'Connor [39].

4.2.1 Linear Least Square Regression

Introduction of Linear Least Square Regression

Assuming

(1) there are N sets of experimental data or observations (x_{oi}, y_{oi}), $i = 1, \ldots, N$.
(2) these data are sampled from a population which follows a linear relationship as shown in Eq. (28),

$$y = a + b \cdot x \tag{28}$$

where a is intercept and b is slope.

The error or deviation (r_{i}) between observed value (y_{oi}) and predicted value ($a + b \cdot x_{\mathrm{oi}}$) is given by Eq. (29).

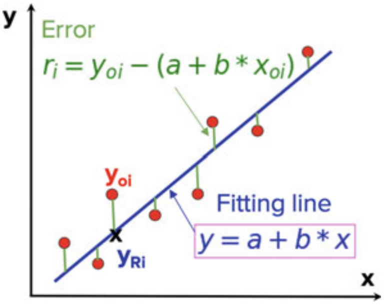
$$r_i = y_{\mathrm{oi}} - (a + b \cdot x_{\mathrm{oi}}) \tag{29}$$

As shown in Fig. 23, the idea behind linear least square regression is to find optimal parameters (a and b) which can minimize the sum of squared error, defined by Eq. (30).

$$S = \sum_{i=0}^{N} r_i^2 \tag{30}$$

As indicated in Eqs. (31) and (32), this optimization can be achieved by setting the differentials of sum of squared errors relative to a and b to zero.

Fig. 23 Minimization of the sum of squared errors or residues



$$\frac{\partial S}{\partial a} = 0 \tag{31}$$

$$\frac{\partial S}{\partial b} = 0 \tag{32}$$

The optimal parameters are given by

$$\hat{a} = \underline{y} - \hat{b} \cdot \underline{x} \tag{33}$$

$$\hat{b} = \frac{N \cdot \sum_{i=1}^{N} [x_i y_i] - \left[\sum_{i=1}^{N} x_i\right] \cdot \left[\sum_{i=1}^{N} y_i\right]}{N \cdot \left[\sum_{i=1}^{N} x_i^2\right] - \left[\sum_{i=1}^{N} x_i\right]^2} \tag{34}$$

$$\underline{x} = \frac{\sum_{i=1}^{N} x_i}{N} \tag{35}$$

$$\underline{y} = \frac{\sum_{i=1}^{N} y_i}{N} \tag{36}$$

One thing to note is that if all of the observations have the same x values, (34) is invalid and the slope b cannot be determined.

The degree of correlation is measured by Eq. (37). The closer to 1 the absolute value of ρ, the better the linear fitting. When ρ is closer to 0, data points have no correlation.

$$\rho = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \frac{\sum_{i=1}^{N} \frac{x_i \cdot y_i}{N} - \underline{x} \cdot \underline{y}}{\sigma_x \cdot \sigma_y} \tag{37}$$

Linear Transformation of Life Distribution Models

Several popular life distribution models are summarized in Table 15. In order to utilize linear least square regression method, we need to do linear transformation for either

Table 15 Examples of life distribution models

Model	Density function (pdf)	Failure probability (CDF)	Reliability
Exponential	$f(t) = \lambda \cdot e^{-\lambda t}$	$F(t) = 1 - e^{-\lambda t}$	$R(t) = e^{-\lambda t}$
Normal	$f(t) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$	$F(t) =$ $\frac{1}{\sqrt{2\pi}\sigma} \int_0^t e^{-\frac{(\tau-\mu)^2}{2\sigma^2}} \mathrm{d}\tau$	$R(t) =$ $\frac{1}{\sqrt{2\pi}\sigma} \int_t^\infty e^{-\frac{(\tau-\mu)^2}{2\sigma^2}} \mathrm{d}\tau$
Lognormal	$f(t) = \frac{1}{\sqrt{2\pi}\sigma t} e^{-\frac{[Ln(t)-\mu]^2}{2\sigma^2}}$	$F(t) =$ $\frac{1}{\sqrt{2\pi}\sigma} \int_0^t \frac{1}{\tau} e^{-\frac{[\mathrm{Ln}(\tau)-\mu]^2}{2\sigma^2}} \mathrm{d}\tau$	$R(t) =$ $\frac{1}{\sqrt{2\pi}\sigma} \int_t^\infty \frac{1}{\tau} e^{-\frac{[\mathrm{Ln}(t)-\mu]^2}{2\sigma^2}} \mathrm{d}\tau$
Weibull	$f(t) = \frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1} e^{-\left(\frac{t}{\alpha}\right)^\beta}$	$F(t) = 1 - e^{-\left(\frac{t}{\alpha}\right)^\beta}$	$R(t) = e^{-\left(\frac{t}{\alpha}\right)^\beta}$

Table 16 Linear transformation of cumulative distribution function

Model	Before transformation	After transformation ($y = a + b \cdot x$)
Exponential	$F(t) = 1 - e^{-\lambda \cdot t}$	$\mathrm{Ln}(1 - F) = -\lambda \cdot t$
Normal	$F(t) = \frac{1}{\sqrt{2\pi}\sigma} \int_0^t e^{-\frac{(\tau-\mu)^2}{2\sigma^2}} \mathrm{d}\tau$	$Z(F) = -\frac{\mu}{\sigma} + \frac{t}{\sigma}$
Lognormal	$F(t) = \frac{1}{\sqrt{2\pi}\sigma} \int_0^t \frac{1}{\tau} e^{-\frac{[\mathrm{Ln}(\tau)-\mu]^2}{2\sigma^2}} \mathrm{d}\tau$	$Z(F) = -\frac{\mu}{\sigma} + \frac{\mathrm{Ln}(t)}{\sigma}$
Weibull	$F(t) = 1 - e^{-\left(\frac{t}{\alpha}\right)^\beta}$	$\mathrm{Ln}[-\mathrm{Ln}(1 - F)] = -\beta \cdot \mathrm{Ln}(\alpha) + \beta \cdot \mathrm{Ln}(t)$

cumulative distribution function (CDF, $F(t)$) or cumulative hazard function ($H(t)$). Table 16 summarizes the linear transformation of cumulative distribution function. In Exponential distribution, $y = \mathrm{Ln}(1 - F)$ and $x = t$. In Normal distribution, $y = Z(F)$ and $x = t$. $Z(F)$ is the 100F$^{\text{th}}$ percentile of standard normal distribution. In Lognormal distribution, $y = Z(F)$ and $x = \mathrm{Ln}(t)$. In Weibull distribution, $y = \mathrm{Ln}[-\mathrm{Ln}(1 - F)]$ and $x = \mathrm{Ln}(t)$.

Assign Failure Probability to Failure Time by Order Statistics

As discussed in Sect. 4.2.1.2, after linear transformation, $x = t$ or $\mathrm{Ln}(t)$. t is the failure time from testing, like hours, cycles, and miles. As shown in Table 17, there are different types of reliability data. If the exact failure time of each unit is known, data type is complete data. If some units do not fail by the end of test time, data type is right censored data. If failure times are within interval, data type is interval-censored data. In linear least square regression, only failed items are plotted. Each failed item is assigned a failure order based on the increasing order sequence of failure times. Right censored items or suspended items are not plotted, but the existence of censored items might impact the overall ranking of failed items. About interval-censored data,

Table 17 Type of reliability data

Data type	Example			
Complete data	Time_to_failure			
	100			
	200			
	300			
	400			
	500			
	600			
Right censored data	Time_to_failure		Censored	
	0.45		No	
	0.46		Yes	
	1.15		No	
	1.15		No	
	1.56		Yes	
Interval censored data	Start	End	Number_failure	Number_entry
	0	6.12	5	167
	6.12	19.92	16	162
	19.92	29.64	12	146
	29.64	35.4	18	134
	35.4	39.72	18	116
	39.72	45.24	2	98
	45.24	52.32	6	96
	52.32	63.48	17	90

failure time is unknown and can be estimated by spreading items uniformly inside interval, or using midpoint or endpoint of interval. As discussed in Sect. 4.2.1.1, if all failed items have the same failure time, slope b cannot be determined. Hence, at least two different failure times are needed for life distribution.

As discussed in Sect. 4.2.1.2, y is a function of failure probability (F). Before linear least square regression, it is necessary to assign failure probability to failure time by order statistics. If the reliability tests with the same sample size are repeated many times, the failure time for i^{th} failure has certain distributions. This also means for failure time t_i, and the failure probability has certain distributions. Order statistics can be applied to estimate the failure probability. As indicated in Eq. (38), a density function ϕ can be expressed as the function of failure probability (F), failure order (i) and sample size (n). Basically, there are $i - 1$ failures before i^{th} order or failure and $n - i$ units after i^{th} order or failure. Equation (39) defines median rank or 50th confident limit of failure probability. If 0.5 on the right-hand side is replaced by 0.95 or 0.05, the 95th confidence limit of failure probability and 5th confident limit of failure probability can be determined. Benard's approximation for median rank is

Table 18 Some data ranking methods for reference

Data type	Method	
Complete data	Benard's approximation for median rank	$F_i = \frac{i-0.3}{N+0.4}$ i is failure order number N is sample size
Right censored data	Adjusted median rank	$j_i = j_{i-1} + \Delta_i$ $j_0 = 0$ $\Delta_i = \frac{N+1-j_{i-1}}{1+[N-\text{number of preceding items(fail,censor)}]}$ $F_i = \frac{j_i-0.3}{N+0.4}$ i is failure order number j_i is adjusted mean order number for each failed item i N is sample size
Interval data	Modified Kaplan–Meier method	$R_i^{'} = 1 - \frac{f_i}{n_i}$ $R_i = R_i^{'} \cdot R_{i-1}$ $R_0 = 1$ $F_i = 1 - R_i$ f_i is number of failure within ith interval n_i is number of units within ith interval

given by Eq. (40).

$$\phi = \frac{n!}{(i-1)! \cdot (n-i)!} F^{i-1} \cdot (1-F)^{n-i} \tag{38}$$

$$\int_{0}^{F_{\text{median}}} \phi(i, F)\mathrm{d}F = 0.5 \tag{39}$$

$$F_i = \frac{i-0.3}{N+0.4} \tag{40}$$

Different data ranking methods for different data types are summarized in Table 18. When data type is mixed and complex (i.e., interval data and right censored data), we can simplify the scenery by assigning failure time to each unit within interval and treating the whole set of data as approximate exact life (i.e., right censored data). For details, please refer to [6, 37–39].

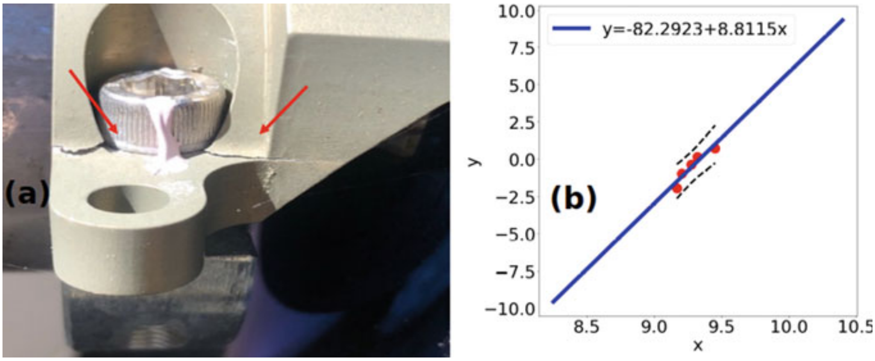


Fig. 24 **a** Lidar bracket crack, **b** two-parameter Weibull life distribution by linear least square regression

Table 19 Linear least square regression of Lidar bracket crack

Order	t (mile)	F	$x = \text{Ln}(t)$	$y = \text{Ln}[-\text{Ln}(1F)]$	x^2	xy
1	9586	0.130	9.168	−1.975	84.053	−18.102
2	9947	0.315	9.205	−0.973	84.733	−8.954
3	10,652	0.5	9.274	−0.367	85.998	−3.399
4	11,146	0.685	9.319	0.145	86.841	1.349
5	12,731	0.870	9.452	0.715	89.336	6.753
Sum			46.417	−2.454	430.961	−22.353
Slope (b)	8.812					
Intercept (a)	−82.292					
Weibull shape (β)	8.812					
Weibull scale (α)	11,374.775					

Case Study—Lidar Bracket Crack

Lidar plays a crucial role in self-driving cars. Before self-driving cars are released for operation, a series of vehicle and sensor calibrations are carried out. If the mechanical fixture like Lidar bracket fails in the field, calibration is violated, and Lidar performance is impacted. Figure 24a shows the crack of Lidar bracket on the front bumper of one vehicle model, and Fig. 24b shows the two-parameter Weibull life distribution of field return data by using linear least square regression. Table 19 shows the raw data and fitting results. Equation (40) was applied to assign failure probability for each failure time. Table 16 was used to transform the Weibull distribution into linear form. Equations (33)–(36) were used for linear least square regression. Once slope and intercept were obtained, Table 16 was used to estimate Weibull shape parameter and scale parameter.

Table 20 Percentile and acceleration factor (AF)

Model	Percentile	AF
Exponential	$t_{\mathrm{p}} = -\frac{\ln(1-P)}{\lambda}$	$\mathrm{AF} = \frac{\lambda_{\mathrm{test}}}{\lambda_{\mathrm{use}}}$
Lognormal	$t_{\mathrm{p}} = e^{\mu + Z(P)\cdot\sigma}$	$\mathrm{AF} = e^{\mu_{\mathrm{use}} - \mu_{\mathrm{test}}}$
Weibull	$t_{\mathrm{p}} = \alpha \cdot [-\ln(1-P)]^{\frac{1}{\beta}}$	$\mathrm{AF} = \frac{\alpha_{\mathrm{use}}}{\alpha_{\mathrm{test}}}$

Analysis of Accelerated Life Testing

When accelerated life testing is carried out under various stress levels, the method discussed in previous subsections can be applied to each stress level such that the life distribution model parameters under each stress level (i) can be obtained individually, including (a_i, β_i) for Weibull distribution, (μ_i, σ_i) for lognormal distribution, or λ_i for exponential distribution.

For the same failure mechanism, the σ of lognormal distribution or β of Weibull distribution is assumed to be the same across stress levels. The common σ or β can be estimated by using Eq. (41) [37]. r_i is the number of failed items under i^{th} stress level. There are m stress levels.

$$\hat{\sigma} = \frac{\sum_{i=1}^{m} r_i \cdot \hat{\sigma}_i}{\sum_{i=1}^{m} r_i} \tag{41}$$

The μ of lognormal distribution, a of Weibull distribution, and λ of exponential distribution depend on stress levels. Table 20 summarizes the percentile and acceleration factor for various life distribution models.

By combining AF in Table 20 and various acceleration models in Sect. 3, Eqs. (42)–(44) can be derived. "T" is temperature, "ΔT" is temperature change, "V" is voltage, "RH" is relative humidity, and "G_{rms}" is root-mean-square g value. γ_i is multiple linear coefficient. Once a_i for Weibull distribution, μ_i for lognormal distribution, or λ_i for exponential distribution and stress conditions (T_i, ΔT_i, V_i, RH_i, $\mathrm{G}_{\mathrm{rms}\,i}$) are entered into Eqs. (42)–(44), γ_i can be obtained by multiple linear regression fitting.

$$\begin{aligned}\mathrm{Ln}(\alpha) = \gamma_0 + \gamma_1 \cdot \left[\frac{1}{T}\right] + \gamma_2 \cdot [\ln(\Delta T)] + \gamma_3 \cdot [\ln V] \\ + \gamma_4 \cdot [\ln(\mathrm{RH})] + \gamma_5 \cdot [G_{\mathrm{rms}}]\end{aligned} \tag{42}$$

$$\begin{aligned}\mu = \gamma_0 + \gamma_1 \cdot \left[\frac{1}{T}\right] + \gamma_2 \cdot [\ln(\Delta T)] + \gamma_3 \cdot [\ln V] \\ + \gamma_4 \cdot [\ln(\mathrm{RH})] + \gamma_5 \cdot [G_{\mathrm{rms}}]\end{aligned} \tag{43}$$

$$\mathrm{Ln}\left(\frac{1}{\lambda}\right) = \gamma_0 + \gamma_1 \cdot \left[\frac{1}{T}\right] + \gamma_2 \cdot [\ln(\Delta T)] + \gamma_3 \cdot [\ln V] + \gamma_4 \cdot [\ln(\mathrm{RH})] + \gamma_5 \cdot [G_{\mathrm{rms}}] \tag{44}$$

The method covered in previous subsections is based on a single independent predictor variable (x). Same concept can be extended to multiple linear regression. Assuming

(1) there are N sets of experimental data or observations (y_i, x_{i1}, …, x_{ik}), i = 1, …, N.
(2) these data are sampled from a population which follows a multiple linear relationship as shown in Eq. (45),

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} \tag{45}$$

where β_0 is intercept and β_1 … β_k are slopes.

The error or deviation between observed value and predicted value is given by Eq. (46).

$$S = \sum_{i=1}^{N} r_i^2 = \sum_{i=1}^{N} \left[y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right]^2 \tag{46}$$

The key idea behind multiple linear regression is to find coefficients β_0, …, β_k which can minimize the error. As indicated in Eq. (47), this optimization can be achieved by setting the differentials of sum of squared errors relative to β_0, …, β_k to zero.

$$\begin{aligned} &N\beta_0 + \beta_1 \sum_{i=1}^{N} x_{i1} + \beta_2 \sum_{i=1}^{N} x_{i2} + \cdots + \beta_k \sum_{i=1}^{N} x_{ik} = \sum_{i=1}^{N} y_i \\ &\beta_0 \sum_{i=1}^{N} x_{i1} + \beta_1 \sum_{i=1}^{N} x_{i1}^2 + \beta_2 \sum_{i=1}^{N} x_{i1}x_{i2} + \cdots + \beta_k \sum_{i=1}^{N} x_{i1}x_{ik} = \sum_{i=1}^{N} x_{i1} y_i \\ &\ldots \\ &\beta_0 \sum_{i=1}^{N} x_{ik} + \beta_1 \sum_{i=1}^{N} x_{ik}x_{i1} + \beta_2 \sum_{i=1}^{N} x_{ik}x_{i2} + \cdots + \beta_k \sum_{i=1}^{N} x_{ik}^2 = \sum_{i=1}^{N} x_{ik} y_i \end{aligned} \tag{47}$$

4.2.2 Maximum Likelihood Estimation

Introduction of Maximum Likelihood Estimation

Assuming

(1) there are N samples. Each sample has lifetime t_i, $i = 1, \ldots, N$.
(2) these samples are from a population which follows a certain life distribution $f(t, \alpha_1, \ldots, \alpha_k)$, where t is lifetime, α_j ($j = 1, \ldots, k$) are unknown model parameters.
(3) for complete data, the probability of failing at a small interval (Δt) containing t_i is characterized by probability density function, as shown in Eq. (48). Because Δt is arbitrary and independent of model parameters, it can be omitted for simplicity. For right censored data, the probability of surviving at t_i is characterized by reliability function, as shown in Eq. (49). For interval-censored data, the probability of failing inside i^{th} interval or between $t_{i\text{L}}$ and $t_{i\text{U}}$ ($t_{i\text{L}}$ is lower end of i^{th} interval, and $t_{i\text{U}}$ is upper end of i^{th} interval) is characterized by the subtraction of two cumulative distribution functions, as shown in Eq. (50).

$$p_i = f(t_i, \alpha_1, \ldots, \alpha_k) \cdot \Delta t \tag{48}$$

$$p_i = R(t_i, \alpha_1, \ldots, \alpha_k) \tag{49}$$

$$p_i = F(t_{i\text{U}}, \alpha_1, \ldots, \alpha_k) - F(t_{i\text{L}}, \alpha_1, \ldots, \alpha_k) \tag{50}$$

By multiplying the probabilities of N samples, the likelihood for the occurrence of these N observations is defined by Eq. (51) and log-likelihood is defined by Eq. (52).

$$L = \prod_{i=1}^{N} p_i \tag{51}$$

$$\text{Ln}(L) = \sum_{i=1}^{N} \text{Ln}(p_i) \tag{52}$$

As indicated in Fig. 25, the key idea behind MLE is to find the model parameters (a_i) which can maximize the log-likelihood.

As indicated in Eq. (53), this optimization can be achieved by setting the differentials of log-likelihood relative to α_i to zero.

$$\frac{\partial \text{Ln}(L)}{\partial \alpha_i} = 0, i = 1, \ldots, k \tag{53}$$

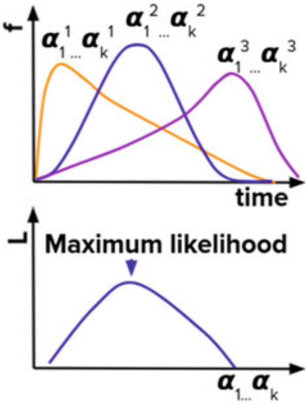


Fig. 25 Maximum likelihood estimation

Once model parameter estimations $(\widehat{\alpha_i})$ are obtained through an iterative numerical method, next step is to calculate the confidence bound for these model parameters. Various methods have been proposed for confidence bound calculation [6, 37–39], including Fisher matrix bound, likelihood-ratio bound, Bayesian bound, Monte-Carlo bound, etc. Fisher matrix bound is based on a normality assumption and straightforward from a scripting point of view. In the following, Fisher matrix bound will be introduced briefly.

Equation (54) is the Fisher information matrix, which is the negative second partial derivative of log-likelihood relative to model parameters. By introducing model parameter estimations $(\widehat{\alpha_i})$, Eq. (54) provides us with local estimate of Fisher information matrix $\widehat{I}(\alpha_i)$.

$$I(\alpha_i) = \begin{bmatrix} -\frac{\partial^2 Ln(L)}{\partial \alpha_1^2} & -\frac{\partial^2 Ln(L)}{\partial \alpha_1 \partial \alpha_2} & \cdots & -\frac{\partial^2 Ln(L)}{\partial \alpha_1 \partial \alpha_k} \\ -\frac{\partial^2 Ln(L)}{\partial \alpha_2 \partial \alpha_1} & -\frac{\partial^2 Ln(L)}{\partial \alpha_2^2} & \cdots & -\frac{\partial^2 Ln(L)}{\partial \alpha_2 \partial \alpha_k} \\ \cdots & \cdots & \cdots & \cdots \\ -\frac{\partial^2 Ln(L)}{\partial \alpha_k \partial \alpha_1} & -\frac{\partial^2 Ln(L)}{\partial \alpha_k \partial \alpha_2} & \cdots & -\frac{\partial^2 Ln(L)}{\partial \alpha_k^2} \end{bmatrix} \tag{54}$$

Once the local estimate of Fisher information matrix is available, Eq. (55) can be used to calculate the covariance matrix, which is the inverse of the local estimate of Fisher information matrix.

$$\widehat{\sum} = \begin{bmatrix} \widehat{\mathrm{Var}}(\alpha_1) & \widehat{\mathrm{Cov}}(\alpha_1, \alpha_2) & \ldots & \widehat{\mathrm{Cov}}(\alpha_1, \alpha_k) \\ \widehat{\mathrm{Cov}}(\alpha_2, \alpha_1) & \widehat{\mathrm{Var}}(\alpha_2) & \ldots & \widehat{\mathrm{Cov}}(\alpha_2, \alpha_k) \\ \ldots & \ldots & \ldots & \ldots \\ \widehat{\mathrm{Cov}}(\alpha_k, \alpha_1) & \widehat{\mathrm{Cov}}(\alpha_k, \alpha_2) & \ldots & \widehat{\mathrm{Var}}(\alpha_k) \end{bmatrix} = I(\alpha_i)^{-1} \tag{55}$$

Once the covariance matrix is available, the two-sided 100(1 − γ)% confidence interval is given by Eqs. (56) or (57). $z_{1-\gamma/2}$ is 100(1 − γ/2)th standard normal percentile. The one-sided 100(1 − γ)% confidence bound can be obtained by replacing $z_{1-\gamma/2}$ with $z_{1-\gamma}$ and using proper sign in Eqs. (56) or (57).

$$[\alpha_{i,\mathrm{L}}, \alpha_{i,\mathrm{U}}] = \hat{\alpha}_i \pm z_{1-\gamma/2}\sqrt{\widehat{\mathrm{Var}}(\hat{\alpha}_i)} \tag{56}$$

$$[\alpha_{i,\mathrm{L}}, \alpha_{i,\mathrm{U}}] = \hat{\alpha}_i e^{\pm\frac{z_{1-\gamma/2}\sqrt{\widehat{\mathrm{Var}}(\hat{\alpha}_i)}}{\hat{\alpha}_i}} \tag{57}$$

As indicated in Tables 15 and 20, reliability metrics (CDF, reliability, and percentile) are function of model parameters, $g(\alpha_1, \ldots, \alpha_k)$. By using model parameter estimations ($\hat{\alpha}_i$), we can get the estimation for g, $\hat{g}(\hat{\alpha}_1, \ldots, \hat{\alpha}_k)$. By using Eq. (58), we can get the variance of g.

$$\widehat{\mathrm{Var}}(\hat{g}) \approx \sum_{i=1}^{k}\left(\frac{\partial g}{\partial \alpha_i}\right)^2 \widehat{\mathrm{Var}}(\hat{\alpha}_i) + \sum_{i=1}^{k}\sum_{j=1, i\neq j}^{k}\left(\frac{\partial g}{\partial \alpha_i}\right)\left(\frac{\partial g}{\partial \alpha_j}\right)\widehat{\mathrm{Cov}}(\hat{\alpha}_i, \hat{\alpha}_j) \tag{58}$$

The two-sided 100(1 − γ)% confidence interval of g is given by Eqs. (59) or (60). $z_{1-\gamma/2}$ is 100(1 − γ/2)th standard normal percentile. The one-sided 100(1 − γ)% confidence bound can be obtained by replacing $z_{1-\gamma/2}$ with $z_{1-\gamma}$ and using proper sign in (59) or (60).

$$[g_{\mathrm{L}}, g_{\mathrm{U}}] = \hat{g} \pm z_{1-\gamma/2}\sqrt{\widehat{\mathrm{Var}}(\hat{g})} \tag{59}$$

$$[g_{\mathrm{L}}, g_{\mathrm{U}}] = \hat{g} e^{\pm\frac{z_{1-\gamma/2}\sqrt{\widehat{\mathrm{Var}}(\hat{g})}}{\hat{g}}} \tag{60}$$

Two-Parameter Weibull Distribution

By substituting the probability density function, cumulative distribution function, and reliability function of Weibull distribution into Eqs. (48–50, 52), the log-likelihood of Weibull distribution can be derived.

$$\mathrm{Ln}(L) = \sum_{i=1}^{R}\mathrm{Ln}\left[\frac{\beta}{\alpha}\left(\frac{t_i}{\alpha}\right)^{\beta-1} e^{-\left(\frac{t_i}{\alpha}\right)^{\beta}}\right] - \sum_{j=1}^{M}\left(\frac{t_j}{\alpha}\right)^{\beta} + \sum_{l=1}^{P}\mathrm{Ln}\left[e^{-\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}} - e^{-\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}}\right] \tag{61}$$

The first term on the right-hand side of Eq. (61) is the log-likelihood of complete data which has R samples. The second term on the right-hand side is the log-likelihood of right censored data which has M samples. The third term on the right-hand side is the log-likelihood of interval-censored data, which has P samples. $t_{l\mathrm{L}}$ is the lower end of l^{th} interval, and $t_{l\mathrm{U}}$ is the upper end of l^{th} interval.

By combining Eqs. (53) and (61), Eqs. (62) and (63) are derived, which can be used to solve unknown model parameters (α and β) by iterative numerical methods. If β is known based on historical data or literature, Eq. (62) itself will be used to solve α and maximize log-likelihood.

$$\begin{aligned}\frac{\partial \mathrm{Ln}(L)}{\partial\alpha} &= \sum_{i=1}^{R}\left[-\frac{\beta}{\alpha}+\frac{\beta}{\alpha}\left(\frac{t_i}{\alpha}\right)^{\beta}\right]+\sum_{j=1}^{M}\frac{\beta}{\alpha}\left(\frac{t_j}{\alpha}\right)^{\beta}\\ &+\sum_{l=1}^{P}\frac{\frac{\beta}{\alpha}\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}e^{-\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}}-\frac{\beta}{\alpha}\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}e^{-\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}}}{e^{-\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}}-e^{-\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}}}=0\end{aligned} \tag{62}$$

$$\begin{aligned}\frac{\partial \mathrm{Ln}(L)}{\partial\beta} &= \sum_{i=1}^{R}\left[\frac{1}{\beta}+\ln\left(\frac{t_i}{\alpha}\right)-\left(\frac{t_i}{\alpha}\right)^{\beta}\ln\left(\frac{t_i}{\alpha}\right)\right]-\sum_{j=1}^{M}\left(\frac{t_j}{\alpha}\right)^{\beta}\ln\left(\frac{t_j}{\alpha}\right)\\ &+\sum_{l=1}^{P}\frac{-\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}\ln\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)e^{-\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}}+\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}\ln\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)e^{-\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}}}{e^{-\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}}-e^{-\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}}}=0\end{aligned} \tag{63}$$

When the data type has complete data only ($M = 0$, $P = 0$), Eqs. (62) and (63) can be simplified as Eqs. (64) and (65). By using Eq. (64), β can be solved first. By using Eq. (65), α can be solved second.

$$\frac{1}{\beta}+\sum_{i=1}^{R}\frac{\ln(t_i)}{R}-\frac{\sum_{i=1}^{R}t_i^{\beta}\ln(t_i)}{\sum_{i=1}^{R}t_i^{\beta}}=0 \tag{64}$$

$$\alpha=\left[\frac{\sum_{i=1}^{R}t_i^{\beta}}{R}\right]^{\frac{1}{\beta}} \tag{65}$$

When the data type has both complete data and right censored data ($P = 0$), Eqs. (62) and (63) can be simplified as Eqs. (66) and (67). t_i^{fail} is the lifetime of complete data or failed items.

$$\frac{1}{\beta}+\sum_{i=1}^{R}\frac{ln\left(t_i^{\text{fail}}\right)}{R}-\frac{\sum_{i=1}^{R+M}t_i^{\beta}\ln(t_i)}{\sum_{i=1}^{R+M}t_i^{\beta}}=0 \tag{66}$$

$$\alpha=\left[\frac{\sum_{i=1}^{R+M}t_i^{\beta}}{R}\right]^{\frac{1}{\beta}} \tag{67}$$

Based on Eqs. (62) and (63), the following second partial derivative of log-likelihood can be derived.

$$
\begin{aligned}
\frac{\partial^2 \mathrm{Ln}(L)}{\partial \alpha^2} &= \sum_{i=1}^{R}\left[\frac{\beta}{\alpha^2}-\frac{\beta(\beta+1)}{\alpha^2}\left(\frac{t_i}{\alpha}\right)^{\beta}\right]-\sum_{j=1}^{M}\frac{\beta(\beta+1)}{\alpha^2}\left(\frac{t_j}{\alpha}\right)^{\beta} \\
&\quad+\sum_{l=1}^{P}\frac{1}{[e^{-\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}}-e^{-\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}}]^2} \\
&\Bigg\{\left[e^{-\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}}-e^{-\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}}\right]\Bigg[-\frac{\beta}{\alpha^2}\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}e^{-\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}}-\left(\frac{\beta}{\alpha}\right)^2\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}e^{-\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}} \\
&\quad+\left(\frac{\beta}{\alpha}\right)^2\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{2\beta}e^{-\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}}+\frac{\beta}{\alpha^2}\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}e^{-\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}} \\
&\quad+\left(\frac{\beta}{\alpha}\right)^2\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}e^{-\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}}-\left(\frac{\beta}{\alpha}\right)^2\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{2\beta}e^{-\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}}\Bigg] \\
&\quad-\Bigg[\frac{\beta}{\alpha}\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}e^{-\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}} \\
&\quad-\frac{\beta}{\alpha}\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}e^{-\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}}\Bigg]\left[\frac{\beta}{\alpha}\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}e^{-\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}}-\frac{\beta}{\alpha}\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}e^{-\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}}\right]\Bigg\}
\end{aligned}
\tag{68}
$$

$$
\begin{aligned}
\frac{\partial^2 \mathrm{Ln}(L)}{\partial \beta^2} &= \sum_{i=1}^{R}\left[-\frac{1}{\beta^2}-\left(\frac{t_i}{\alpha}\right)^{\beta}\ln^2\left(\frac{t_i}{\alpha}\right)\right] \\
&\quad-\sum_{j=1}^{M}\left(\frac{t_j}{\alpha}\right)^{\beta}\ln^2\left(\frac{t_j}{\alpha}\right)+\sum_{l=1}^{P}\frac{1}{\left[e^{-\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}}-e^{-\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}}\right]^2} \\
&\Bigg\{\left[e^{-\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}}-e^{-\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}}\right]\Bigg[-\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}\ln^2\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)e^{-\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}} \\
&\quad+\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{2\beta}\ln^2\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)e^{-\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}} \\
&\quad+\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}\ln^2\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)e^{-\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}}-\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{2\beta}\ln^2\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)e^{-\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}}\Bigg] \\
&\quad-\left[-\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}\ln\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)e^{-\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}}+\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}\ln\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)e^{-\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}}\right]^2\Bigg\}
\end{aligned}
\tag{69}
$$

$$
\frac{\partial^2 \mathrm{Ln}(L)}{\partial \alpha \partial \beta} = \sum_{i=1}^{R}\left[-\frac{1}{\alpha}+\frac{1}{\alpha}\left(\frac{t_i}{\alpha}\right)^{\beta}+\frac{\beta}{\alpha}\left(\frac{t_i}{\alpha}\right)^{\beta}\ln\left(\frac{t_i}{\alpha}\right)\right]
$$

$$
\begin{aligned}
&+\sum_{j=1}^{M}\left[\frac{1}{\alpha}\left(\frac{t_j}{\alpha}\right)^{\beta}+\frac{\beta}{\alpha}\left(\frac{t_j}{\alpha}\right)^{\beta}\ln\left(\frac{t_j}{\alpha}\right)\right]\\
&+\sum_{l=1}^{P}\frac{1}{\left[e^{-\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}}-e^{-\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}}\right]^{2}}\left\{\left[e^{-\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}}-e^{-\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}}\right]\right.\\
&\left[\frac{1}{\alpha}\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}e^{-\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}}+\frac{\beta}{\alpha}\ln\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}e^{-\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}}\right.\\
&-\frac{\beta}{\alpha}\ln\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{2\beta}e^{-\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}}-\frac{1}{\alpha}\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}e^{-\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}}\\
&-\frac{\beta}{\alpha}\ln\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}e^{-\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}}\\
&\left.+\frac{\beta}{\alpha}\ln\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{2\beta}e^{-\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}}\right]\\
&-\left[\frac{\beta}{\alpha}\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}e^{-\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}}-\frac{\beta}{\alpha}\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}e^{-\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}}\right]\\
&\left.\left[-\ln\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}e^{-\left(\frac{t_{l\mathrm{L}}}{\alpha}\right)^{\beta}}+\ln\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}e^{-\left(\frac{t_{l\mathrm{U}}}{\alpha}\right)^{\beta}}\right]\right\}
\end{aligned}
\tag{70}
$$

The local estimate of Fisher information matrix for two-parameter Weibull is given by Eq. (71). The covariance matrix is given by Eq. (72).

$$
I(\alpha,\beta)=\begin{bmatrix}-\frac{\partial^2 Ln(L)}{\partial\alpha^2} & -\frac{\partial^2 Ln(L)}{\partial\alpha\partial\beta}\\ -\frac{\partial^2 Ln(L)}{\partial\alpha\partial\beta} & -\frac{\partial^2 Ln(L)}{\partial\beta^2}\end{bmatrix}
\tag{71}
$$

$$
\widehat{\sum}=\begin{bmatrix}\widehat{\mathrm{Var}}(\alpha) & \widehat{\mathrm{Cov}}(\alpha,\beta)\\ \widehat{\mathrm{Cov}}(\alpha,\beta) & \widehat{\mathrm{Var}}(\beta)\end{bmatrix}=I(\alpha,\beta)^{-1}
\tag{72}
$$

Case Study—Lidar Bracket Crack

A python script is customized based on the linear least square regression method and maximum likelihood estimation method covered in previous subsections. scipy.optimize.newton and scipy.optimize.root [40] are used to solve Eqs. (62)–(64), and (66) for maximum likelihood estimation. The parameter estimations from linear least square regression are used as initial guesses for the optimization of maximum likelihood estimation. numpy.linalg.inv [41] is used to get the covariance matrix from local estimates of Fisher information matrix by Eqs. (55) and (72).

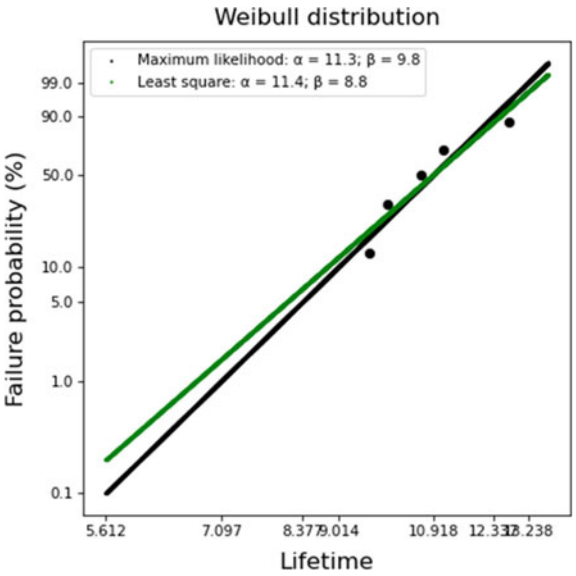


Fig. 26 Life distribution of Lidar bracket crack

The Lidar bracket crack data in Table 19 is used as case study again. To avoid convergence issues related to overflow runtime error, raw data is normalized by 10^3. Hence, the times to failure used for data analysis become 9.586, 9.947, 10.652, 11.146, and 12.731 with units 10^3 miles. As indicated in Fig. 26, the parameter estimations returned by linear least square regression (green color) are $\alpha = 11.4$ and $\beta = 8.8$ while the parameter estimations returned by maximum likelihood estimation (black color) are $\alpha = 11.3$ and $\beta = 9.8$. These parameter estimations by customized python script are the same as those returned by JMP software ($\alpha = 11.3$ and $\beta = 9.8$) and Relyence software ($\alpha = 11.3$ and $\beta = 9.8$).

The local estimate of Fisher information matrix for two-parameter Weibull is given by Eq. (73). The covariance matrix is given by Eq. (74).

$$\widehat{I}(\alpha, \beta)\begin{bmatrix} 3.76065328 & -0.21553316 \\ -0.21553316 & 0.10664804 \end{bmatrix} \tag{73}$$

$$\widehat{\sum} = \widehat{I}(\alpha, \beta)^{-1} = \begin{bmatrix} 0.30074592 & 0.60780037 \\ 0.60780037 & 10.60498755 \end{bmatrix} \tag{74}$$

By using Eq. (57), the two-sided 95% confidence interval of α is [10.3, 12.5] and the two-sided 95% confidence interval of β is [5.1, 18.8].

Life Distribution Comparison

Sometimes, reliability engineers need to compare life distributions to help team down select designs and processes and develop reliability models to predict field failure rate based on testing result under accelerated testing conditions. Assuming

(1) there are k groups (designs, processes, accelerated life testing conditions, etc.).
(2) the m^{th} group ($m = 1, \ldots, k$) has R^m complete data, M^m right censored data, and P^m interval-censored data. Each complete data has lifetime $t_i{}^m$, each right censored data survives $t_j{}^m$, and each interval has bounds $[t_{l\text{L}}{}^m, t_{l\text{U}}{}^m]$.
(3) the m^{th} group follows a Weibull distribution with shape β^m and scale α^m.

Equation (75) provides the log-likelihood for the whole set of data. The subscript "S" means different groups have different or separate shape parameters. Equations (76) and (77) have 2 k equations and 2 k unknowns and can be used to solve α^m and β^m.

$$\begin{aligned}\mathrm{Ln}(L)_S = \sum_{m=1}^{k}\Bigg\{&\sum_{i=1}^{R^m}\mathrm{Ln}\left[\frac{\beta^m}{\alpha^m}\left(\frac{t_i^m}{\alpha^m}\right)^{\beta^m-1} e^{-\left(\frac{t_i^m}{\alpha^m}\right)^{\beta^m}}\right] \\ &- \sum_{j=1}^{M^m}\left(\frac{t_j^m}{\alpha^m}\right)^{\beta^m} + \sum_{l=1}^{P^m}\mathrm{Ln}\left[e^{-\left(\frac{t_{l\mathrm{L}}^m}{\alpha^m}\right)^{\beta^m}} - e^{-\left(\frac{t_{l\mathrm{U}}^m}{\alpha^m}\right)^{\beta^m}}\right]\Bigg\}\end{aligned} \tag{75}$$

$$\frac{\partial \mathrm{Ln}(L)_S}{\partial \alpha^m} = 0, m = 1, \ldots, k \tag{76}$$

$$\frac{\partial \mathrm{Ln}(L)_S}{\partial \beta^m} = 0, m = 1, \ldots, k \tag{77}$$

If the failure mechanisms are the same, shape parameter β^m are expected to be the same or common. Equations (75)–(77) can be modified as Eqs. (78)–(80). The subscript "C" means different groups have common shape parameters. Equations (79) and (80) have $k + 1$ equations and $k + 1$ unknown and can be used to solve α^m and β.

$$\begin{aligned}\mathrm{Ln}(L)_C = \sum_{m=1}^{k}\Bigg\{&\sum_{i=1}^{R^m}\mathrm{Ln}\left[\frac{\beta}{\alpha^m}\left(\frac{t_i^m}{\alpha^m}\right)^{\beta-1} e^{-\left(\frac{t_i^m}{\alpha^m}\right)^{\beta}}\right] \\ &- \sum_{j=1}^{M^m}\left(\frac{t_j^m}{\alpha^m}\right)^{\beta} + \sum_{l=1}^{P^m}\mathrm{Ln}\left[e^{-\left(\frac{t_{l\mathrm{L}}^m}{\alpha^m}\right)^{\beta}} - e^{-\left(\frac{t_{l\mathrm{U}}^m}{\alpha^m}\right)^{\beta}}\right]\Bigg\}\end{aligned} \tag{78}$$

$$\frac{\partial \mathrm{Ln}(L)_C}{\partial \alpha^m} = 0, m = 1, \ldots, k \tag{79}$$

$$\frac{\partial \mathrm{Ln}(L)_C}{\partial \beta} = 0 \tag{80}$$

To decide whether a common shape parameter should be used or not, the likelihood-ratio test (L-R test) can be conducted. In Eq. (81), $\mathrm{Ln}(L)_S$ comes from Eq. (75) and $\mathrm{Ln}(L)_C$ comes from Eq. (78).

$$T = 2 \cdot \left[\mathrm{Ln}(L)_S - \mathrm{Ln}(L)_C\right] \tag{81}$$

Based on Nelson [35], if $T \leq \chi^2(1-\gamma, k-1)$, the k shape parameters do not differ statistically at the 100 $\gamma\%$ level and can be treated as the same or common. $\chi^2(1-\gamma, k-1)$ is the Chi-square percentile with $k - 1$ degree of freedom, which can be evaluated by using python's Scipy package $scipy.stats.chi2.ppf(1-\gamma, k-1)$. In fact, instead of checking $T \leq scipy.stats.chi2.ppf(1-\gamma, k-1)$, we can also evaluate the following relationship, $\gamma \leq 1 - scipy.stats.chi2.cdf(T, k-1)$. If this relationship is met, the k shape parameters can be treated as the same or common.

After shape parameters and scale parameters are estimated, the methods described in Sects. 5.2.1 and 5.2.2 can be used to calculate the confidence intervals of model parameters and reliability metrics. To comment whether two groups are different or not, we should consider the confidence intervals. If the confidence interval of one group envelopes the estimate of another group, the estimates of two groups are comparable and not statistically significantly different. If the confidence intervals of two groups do not overlap, the estimates of two groups are statistically significantly different. For other situations, we cannot make a conclusion and need to use other methods.

Case Study—Glue Process Improvement of Dash Mount Audio Device

Temperature cycle test (−40 °C/85 °C) was used to qualify the dash mount audio device inside the car. The epoxy or glue between the metal plate and plastic enclosure delaminated during the temperature cycle test. To mitigate the risk, the glue dispensing process was changed from old process to new process by optimizing the dispensing pattern and dispensing weight. Samples with the old glue process and new glue process were loaded into the temperature cycle test (−40 °C/85 °C) for life distribution comparison. Table 21 summarized the lifetimes of the dash mount audio devices with the old glue process and the new glue process after temperature cycle test. The failure mode tracked is glue crack.

Figure 27 showed the Weibull life distributions of two glue processes analyzed by using JMP software. (a) used separate shape parameters and (b) used common shape parameters. Shaded areas mean 95% confidence interval. The life distribution of the new glue process shifted to the right in comparison with the one of the old glue processes. When a common shape parameter was used, the confidence intervals of two glue processes do not overlap.

Table 21 Lifetime of dash mount audio device

Glue process	Start (cyc)	End (cyc)	Number of failure	Censor type
Old	0	50	10	Interval
	50	100	10	Interval
	100	200	1	Interval
	50		1	Right
	250		1	Right
New	100	150	4	Interval
	150	200	2	Interval
	250		17	Right

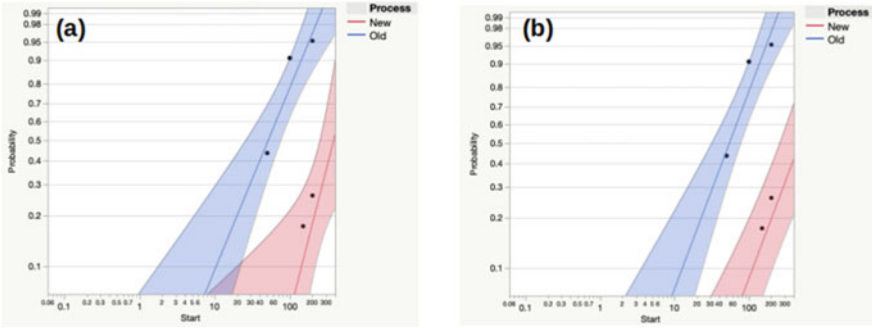


Fig. 27 Weibull life distribution of dash mount audio device **a** with separate shape parameter and **b** with common shape parameter

Table 22 summarized the Weibull parameter estimations and confidence intervals returned by JMP software. By using the "−2LoglikeLihood" for "Location" (common shape parameter and separate scale parameters) and "Location and Scale" (separate shape parameter and separate scale parameter) from JMP, the L-R test has $T = -96.36582 - (-97.40843) = 1.043$, which is also provided by JMP under "L-R ChiSquare" for "Location vs. Location and Scale." For two groups, $scipy.stats.chi2.ppf(1 - 0.05, 2 - 1) = 3.841$, $1 - scipy.stats.chi2.cdf(1.043, 2 - 1) = 0.3071$. Because $T = 1.043 < 3.841$ or $\gamma = 0.05 < 0.3071$, the shape parameters of two groups do not differ statistically at the 5% level and can be treated as the same or common. In JMP software, this conclusion is supported by the fact that "Prob > ChiSq" for "Location vs. Location and Scale" is 0.3072 > 0.05.

Because the confidence intervals of two glue processes do not overlap, the estimates of two glue processes are statistically significantly different. Because the life distribution of the new glue process shifted to the right, the new glue process does improve the robustness of the device.

Table 22 Estimate of dash mount audio device

Fitting method	Parameter (glue process)	Estimate	95% Lower bound	95% Upper bound
With separate β	α(new)	457.8	299.7	1596.9
	β(new)	1.9	0.8	3.8
	α(old)	69.2	43.2	103.3
	β(old)	1.2	0.7	1.7
With common β	α(new)	631	357	1661
	α(old)	71.4	47.6	102.9
	β	1.3	0.9	1.8

Case Study—Developing Reliability Model for Glue Crack in Camera

During the qualification of one type of camera, the glue or epoxy between plastic lens housing and ceramic substrate cracked after temperature cycle testing (−40 °C/85 °C). To predict field failure rate, a reliability model needs to be developed. To estimate the Coffin-Manson exponent m in Eq. (10), accelerated life testing with two conditions, −40 °C/85 °C and −20 °C/65 °C, was conducted. These testing conditions were cherry picked to stay within storage limits of the camera and avoid stress artifacts. Table 23 summarized the lifetime of cameras. The failure mode tracked is glue crack.

Figure 28 showed the Weibull life distributions of cameras under two temperature cycle conditions analyzed by using JMP software. Common shape parameter was used.

Table 24 summarized the Weibull parameter estimations and confidence intervals returned by JMP software. By using the "−2LoglikeLihood" for "Location" (common shape parameter) and "Location and Scale" (separate shape parameters) from JMP, the L-R test has $T = -78.74705 - (-78.77694) = 0.030$. For two groups, $scipy.stats.chi2.ppf(1-0.05, 2-1) = 3.841$, $1 - scipy.stats.chi2.cdf(0.030, 2-1) = 0.8625$. Because $T = 0.030 < 3.841$ or $\gamma = 0.05 < 0.8625$, the shape parameters of two groups do not differ statistically at the 5% level and can be treated as the same or common. In JMP software, this conclusion is supported by the fact that "Prob > ChiSq" for "Location vs. Location and Scale" is 0.8627 > 0.05. For the same failure mechanism, a common shape parameter is expected unless the accelerated testing condition is too harsh and generates other failure modes.

For temperature cycle testing, Eq. (40) can be simplified as

$$\mathrm{Ln}(\alpha) = \gamma_0 + \gamma_1 \cdot [\ln(\Delta T)] \tag{82}$$

By substituting α and ΔT from Table 24 (with common shape parameter) into Eq. (82), we can get matrix (83)

Table 23 Lifetime of cameras

TC condition	Start (cyc)	End (cyc)	Number of failure	Censor type
−40 °C/85 °C	25		50	Right
	50		50	Right
	100		50	Right
	0	150	1	Interval
	150		49	Right
	0	200	1	Interval
	200		49	Right
	0	250	6	Interval
	250		44	Right
−20 °C/65 °C	25		50	Right
	50		50	Right
	100		50	Right
	150		50	Right
	200		50	Right
	350	400	1	Interval
	400	500	1	Interval
	500		48	Right

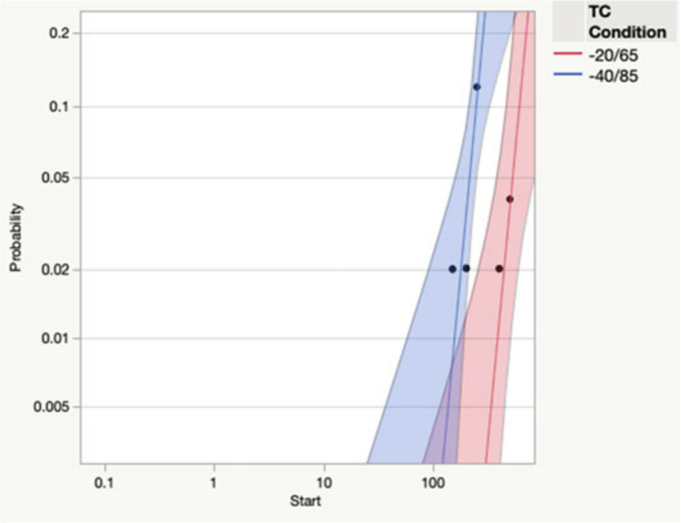


Fig. 28 Weibull life distribution of cameras

Table 24 Estimate of camera

Fitting method	Parameter (TC condition)	Estimate	95% Lower bound	95% Upper bound
With separate β	α(−40 °C/85 °C)	385.7	294.4	1241.7
	β(−40 °C/85 °C)	4.9	1.7	10.5
	α(−20 °C/65 °C)	880.3	586.9	6073.0
	β(−20C/65 °C)	5.7	1.5	18.2
With common β	α(−40 °C/85 °C)	377.6	297.7	755.0
	α(−20 °C/65 °C)	932.9	551.4	2407.2
	β	5.1	2.4	9.8

$$\begin{bmatrix}\mathrm{Ln}(377.6)\\ \mathrm{Ln}(932.9)\end{bmatrix}=\begin{bmatrix}1 & \mathrm{Ln}(125)\\ 1 & \mathrm{Ln}(85)\end{bmatrix}\cdot\begin{bmatrix}\gamma_0\\ \gamma_1\end{bmatrix} \tag{83}$$

By using Python packages, such as numpy.linalg.inv and numpy.matmul, we can get $\gamma_0 = 17.257$ and $\gamma_1 = -2.345$. Hence, the Coffin-Manson exponent $m = -\gamma_1 = 2.345$. The acceleration model for temperature cycle can be expressed as

$$\mathrm{Ln}(\alpha) = 17.257 - 2.345 \cdot [\ln(\Delta T)] \tag{84}$$

By substituting temperature change in the field into Eq. (84), we can get the scale parameter in the field. By using common shape parameter (β) and scale parameter in the field (α), we can estimate the field failure rate.

The procedure described above got the scale parameters and shape parameters for each testing condition by maximum likelihood estimation first and solved the acceleration model parameter second. A better method was presented by Nelson [35]. The idea is to combine Eqs. (40) and (78). The log-likelihood will be the function of acceleration model parameters (γ_i), Weibull shape parameter (β), temperature cycle testing condition (ΔT_i), and lifetime (t_i). Because ΔT_i and t_i are known, the optimization problem will be forcing the first derivatives of log-likelihood relative to γ_i and β to zero. Due to the page limit, the final equations would not be listed here.

Three-Parameter Weibull Distribution

Sometimes, a two-parameter Weibull distribution cannot fit the life distribution data very well due to the existence of failure free time, which can be positive or negative. This is especially true when a burn-in process (such as high temperature stress, reflow stress, temperature cycle stress, and vibration stress) is applied to kill infant mortality failure at the end of the assembly line. A three-parameter Weibull distribution can handle this situation. In Eqs. (85)–(87), t_0 is a failure free time or location parameter. By using these equations and methods described in previous subsections, the likelihood and log-likelihood can be expressed. The optimization of log-likelihood will be forcing the first derivatives of log-likelihood relative to α_i, β_i, and t_0 to zero. Due to the page limit, the final equations would not be provided here.

$$f(t) = \frac{\beta}{\alpha}\left(\frac{t - t_0}{\alpha}\right)^{\beta-1} e^{-\left(\frac{t-t_0}{\alpha}\right)^{\beta}} \tag{85}$$

$$F(t) = 1 - e^{-\left(\frac{t-t_0}{\alpha}\right)^{\beta}} \tag{86}$$

$$R(t) = e^{-\left(\frac{t-t_0}{\alpha}\right)^{\beta}} \tag{87}$$

4.3 *Confidence Interval*

In previous subsections, the confidence interval of the estimate of life distribution was briefly discussed. In this subsection, more discussions are provided.

4.3.1 Central Limit Theorem

Assume a population has mean μ and standard deviation σ. If a sampling with the same sample size n from this population is repeated many times, the distribution of each sampling's mean approaches a normal distribution, no matter which kind of distribution the population has. As indicated in Fig. 29, the mean ($\mu_{\underline{x}}$) of each sampling's mean ($\underline{x}$) is equal to the mean of population and the standard deviation ($\sigma_{\underline{x}}$) of each sampling's mean is equal to the standard deviation of population divided by square root of sampling sample size. This is true for large sampling sample sizes. With the increase of sampling sample size, the spread of each sampling's mean is reduced and each sampling's mean is closer to the mean of population.

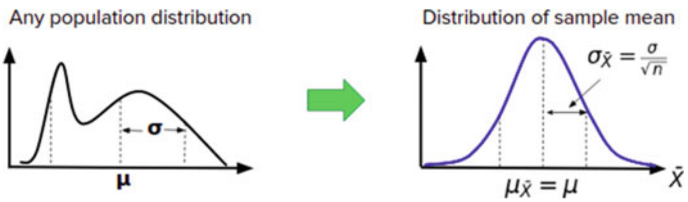

Any population distribution
σ
μ
Distribution of sample mean
$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
$\mu_{\bar{X}} = \mu$
$\bar{X}$

Fig. 29 Central limit theorem

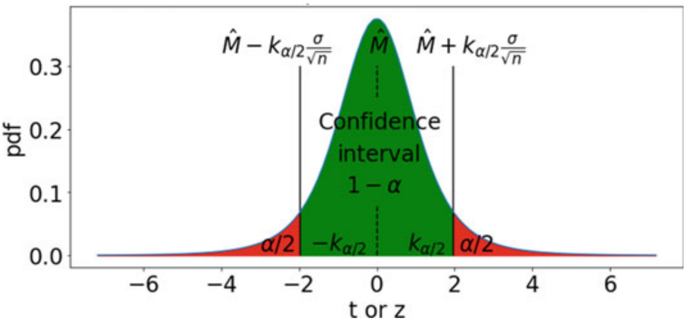


Fig. 30 Confidence interval of mean

4.3.2 Confidence Interval of Mean

Confidence level (CL) is the probability that the confidence interval (CI) contains the population parameter. Significance level is the probability that the population parameter is outside the confidence interval. As shown in Fig. 30, the sum of confidence level and significance level is equal to one (the area under the pdf curve shall be equal to one).

Based on the Central limit theorem, if a sampling with sampling sample size n has mean $\widehat{M}$ and the population standard deviation σ is known, the $100(1-\alpha)\%$ two-sided confidence interval of population mean is

$$\mathrm{CI} \in \left[\widehat{M} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \widehat{M} + Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] \tag{88}$$

$Z_{\frac{\alpha}{2}}$ is $100(1-\alpha/2)$th standard normal percentile and represents the number of standard deviations from mean.

When the population standard deviation σ is not known, t-distribution is used to estimate the confidence interval of population mean. In Eq. (89), s is unbiased standard deviation of sampling data. $t_{\frac{\alpha}{2};n-1}$ is $100(1-\alpha/2)$th percentile of t-distribution with $n-1$ degree of freedom.

$$\mathrm{CI} \in \left[\widehat{M} - t_{\frac{\alpha}{2};n-1}\frac{s}{\sqrt{n}}, \widehat{M} + t_{\frac{\alpha}{2};n-1}\frac{s}{\sqrt{n}}\right] \tag{89}$$

4.3.3 Confidence Interval for Sample Proportion

Sometimes, reliability data is non-parametric and does not have lifetime information. For example, after the test, x samples out of n samples fail, but no failure time information is available.

The estimate of failure probability or proportion is

$$\hat{p} = \frac{x}{n} \tag{90}$$

The standard deviation of failure probability or proportion is

$$\sigma_{\hat{p}} = \sqrt{\frac{(1-\hat{p})^2 n\hat{p} + (0-\hat{p})^2 n(1-\hat{p})}{n}} = \sqrt{\hat{p}(1-\hat{p})} \tag{91}$$

The confidence interval of failure probability or proportion can be estimated by replacing $\widehat{M}$ with $\hat{p}$ and σ with $\sigma_{\hat{p}}$ in Eq. (88).

$$\mathrm{CI} \in \left[\hat{p} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \tag{92}$$

If the population size is limited, population size N can be introduced to correct the confidence interval.

$$\mathrm{CI} \in \left[\hat{p} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})(N-n)}{n(N-1)}}, \hat{p} + Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})(N-n)}{n(N-1)}}\right] \tag{93}$$

Let us look at the following example as a case study.

Free fall test or drop test is widely used to mimic the mishandling of automotive components during an assembly process. Drop height (such as 1 m) and drop medium (such as concrete and carpet) are generally controlled. After a drop test of one type of automotive module, six out of one hundred units failed. When the result was published externally, it said "failure rate was 6%." Is this correct? No. By substituting x, n, and α with 6, 100, and 0.05 in Eqs. (90) and (92), we are 95% confident that the failure rate is between 0.013 and 0.107.

4.3.4 Binomial Bound

In Sect. 4, non-parametric Binomial zero defect sampling was discussed. If no failure is observed, the upper bound of failure rate is given by Eq. (22). This relationship can be approximated by the Rule of Three—if there is zero failure in n samples, we are 95% confident that the failure rate is between 0 and $3/n$. This Rule of Three can be generalized, and we are $[1-\alpha] \times 100\%$ confident that the failure rate is between 0 and $-\ln(\alpha)/n$.

Let us look at the following example as a case study.

Engineers did vibration tests on power supply units, which were used by one type of self-driving car. The result was zero failure out of twenty-two units (0F/22). When the result was published externally, it said "Failure rate was 0%." Is this correct? No. We are 95% confident that the upper bound of failure rate is 3/22 ~ 14%.

4.4 Hypothesis Test

During the down selection of designs, materials, and processes, hypothesis tests are used frequently.

Mean comparison methods include *z*-test, *t*-test, ANOVA, etc. *Z*-test and *t*-test can be used to compare mean of one group with known target or compare means of two groups. If population statistics are known, *z*-test can be used; if population statistics are unknown, *t*-test can be used. ANOVA can be used to compare means among two or more groups.

Median comparison methods include Kruskal–Wallis test, etc.

Variance comparison methods include Chi-square test, *F*-test, Bartlett test, Levene's test, etc. Chi-square test is to compare variance of one group with a known value. *F*-test is to compare variances between two groups. Bartlett test is to compare variances among two or more groups. Usually, it assumes normality. Levene's test also compares variances of two or more groups, but does not need normality assumption.

This subsection will discuss some of these methods. For details, please refer to [36, 42].

4.4.1 Mean Comparison

Student's t Test

Equation (94) is the probability density function of Student's *t*-distribution. ν is the number of degrees of freedom, Γ is the Gamma function, and t is the *t*-value or statistics.

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{94}$$

As indicated in Fig. 31, the pdf of *t*-distribution is symmetric and bell-shaped, but has heavier or longer tails than normal distribution.

To compare the mean ($\underline{x}$) of one sample group with population mean (μ), Eq. (95) can be used. Based on the sampling data (the sample mean, unbiased standard deviation (s)), a *t*-statistics (T) can be calculated. The number of degrees of freedom is sample size (n) minus one.

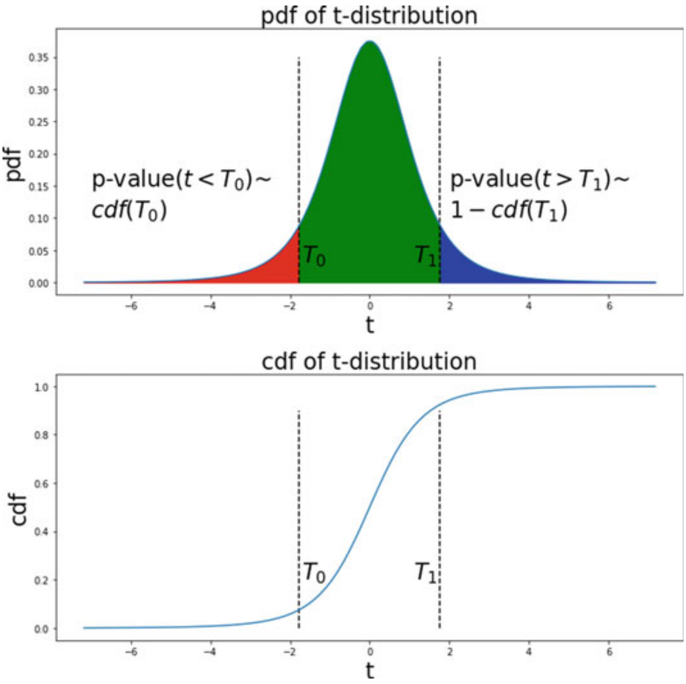


Fig. 31 t-distribution with $\nu = 4$

$$T = \frac{\underline{x} - \mu}{\frac{s}{\sqrt{n}}} \tag{95}$$

To evaluate whether the mean of one sample group and the population mean are statistically equal, we can check the two-sided p-value. If the p-value for absolute value of t-statistics, 1-cdf($|T|$), $< \gamma/2$, we are $100(1 - \gamma)$ confident that the mean of one sample group and the population mean are statistically different. Otherwise, the difference is not significant.

To evaluate whether the mean of one sample group is statistically smaller than the population mean, we can check the one-sided p-value. If the p-value for t-statistics, cdf(T), $< \gamma$, we are $100(1 - \gamma)$ confident that the mean of one sample group is statistically smaller than the population mean. Otherwise, the difference is not significant.

To evaluate whether the mean of one sample group is statistically larger than the population mean, we can check the one-sided p-value. If the p-value for t-statistics, 1-cdf(T), $< \gamma$, we are $100(1 - \gamma)$ confident that the mean of one sample group is statistically larger than the population mean. Otherwise, the difference is not significant.

To compare the means ($\underline{x}_1$, $\underline{x}_2$) of two sample groups which have equal variance, Eq. (96) can be used. S_p is the pooled unbiased standard deviation. Subscripts "1" and “2” represent group 1 and group 2.

$$T = \frac{\underline{x}_1 - \underline{x}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{96}$$

$$S_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{97}$$

$$\nu = n_1 + n_2 - 2 \tag{98}$$

To compare the means ($\underline{x}_1$, $\underline{x}_2$) of two sample groups which have unequal variance, Eq. (99) can be used. Subscripts “1” and “2” represent group 1 and group 2.

$$T = \frac{\underline{x}_1 - \underline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{99}$$

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} \tag{100}$$

To evaluate whether the means of two sample groups are statistically equal, we can check the two-sided p-value. If the p-value for absolute value of t-statistics, 1-cdf($|T|$), < $\gamma/2$, we are $100(1 - \gamma)$ confident that the means of two sample groups are statistically different. Otherwise, the difference is not significant.

To evaluate whether the mean of group one is statistically smaller than the mean of group two, we can check the one-sided p-value. If the p-value for t-statistics, cdf(T), < γ, we are $100(1 - \gamma)$ confident that the mean of group one is statistically smaller than the mean of group two. Otherwise, the difference is not significant.

To evaluate whether the mean of group one is statistically larger than the mean of group two, we can check the one-sided p-value. If the p-value for t-statistics, 1-cdf(T), < γ, we are $100(1 - \gamma)$ confident that the mean of group one is statistically larger than the mean of group two. Otherwise, the difference is not significant.

Sometimes, data type is binomial, either “pass” or “fail.” For example, samples come from class test, free fall test, etc. To utilize the equations above, we can transform the data. For each unit, if it passes the test, it is assigned a numerical value of “0”; if it fails the test, it is assigned a numerical value “1.” Eq. (101) defines failure rate per group ($i = 1, 2$). F_i is the number of failures per group. n_i is the number of samples per group. Equation (102) is the unbiased standard deviation per group. Equation (99) can be rewritten as Eq. (103).

$$p_i = \frac{F_i}{n_i} \tag{101}$$

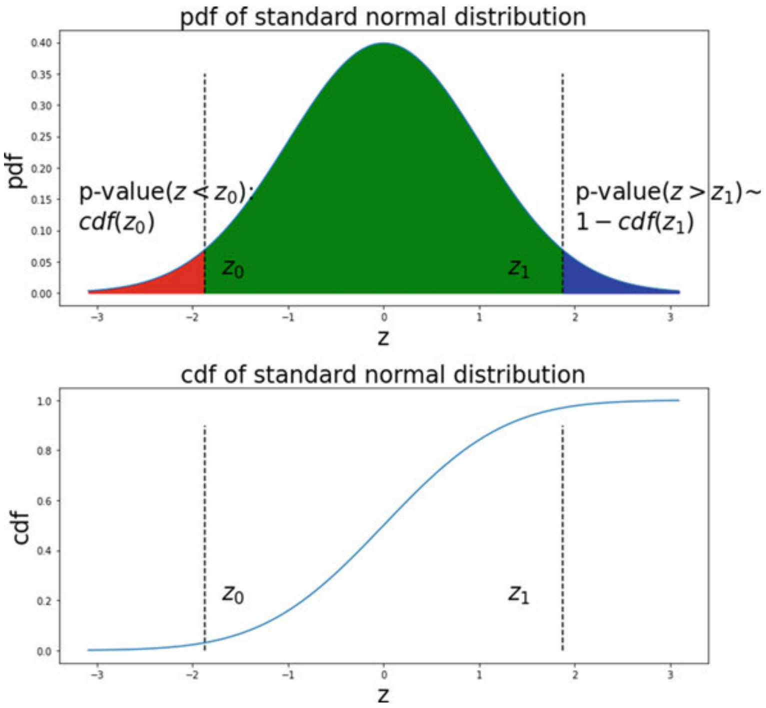


Fig. 32 Standard normal distribution

$$s_i^2 = \frac{n_i p_i (1 - p_i)}{n_i - 1} \tag{102}$$

$$T = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1-1} + \frac{p_2(1-p_2)}{n_2-1}}} \tag{103}$$

Z Test

Equation (104) is the probability density function of a standard normal distribution. Z is the Z-value or statistics (Fig. 32).

$$f(Z) = \frac{e^{-\frac{Z^2}{2}}}{\sqrt{2\pi}} \tag{104}$$

To compare the mean ($\underline{x}$) of one sample group with population mean (μ), Eq. (105) can be used. Based on the sample mean and population standard deviation (σ), a z-statistics can be calculated. The z-statistics tells how far, in standard deviations, the mean of one sample group is from population mean.

$$Z = \frac{\underline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{105}$$

To evaluate whether the mean of one sample group and the population mean are statistically equal, we can check the two-sided *p*-value. If the *p*-value for absolute value of z-statistics, 1-cdf(|Z|), < γ/2, we are 100(1 − γ) confident that the mean of one sample group and the population mean are statistically different. Otherwise, the difference is not significant.

To evaluate whether the mean of one sample group is statistically smaller than the population mean, we can check the one-sided *p*-value. If the *p*-value for *z*-statistics, cdf(*Z*), < γ, we are 100(1 − γ) confident that the mean of one sample group is statistically smaller than the population mean. Otherwise, the difference is not significant.

To evaluate whether the mean of one sample group is statistically larger than the population mean, we can check the one-sided *p*-value. If the *p*-value for *z*-statistics, 1-cdf(*Z*), < γ, we are 100(1 − γ) confident that the mean of one sample group is statistically larger than the population mean. Otherwise, the difference is not significant.

To compare the means ($\underline{x}_1, \underline{x}_2$) of two sample groups with equal variance, Eq. (106) can be used. Subscripts "1" and "2" represent group 1 and group 2.

$$Z = \frac{\underline{x}_1 - \underline{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{106}$$

To compare the means ($\underline{x}_1, \underline{x}_2$) of two sample groups with unequal variance, Eq. (107) can be used. Subscripts "1" and "2" represent group 1 and group 2.

$$Z = \frac{\underline{x}_1 - \underline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{107}$$

To evaluate whether the means of two sample groups are statistically equal, we can check the two-sided *p*-value. If the *p*-value for absolute value of *z*-statistics, 1-cdf(|Z|), < γ/2, we are 100(1 − γ) confident that the means of two sample groups are statistically different. Otherwise, the difference is not significant.

To evaluate whether the mean of group one is statistically smaller than the mean of group two, we can check the one-sided *p*-value. If the *p*-value for *z*-statistics, cdf(*Z*), < γ, we are 100(1 − γ) confident that the mean of group one is statistically smaller than the mean of group two. Otherwise, the difference is not significant.

To evaluate whether the mean of group one is statistically larger than the mean of group two, we can check the one-sided *p*-value. If the *p*-value for *z*-statistics, 1-cdf(*Z*), < γ, we are 100(1 − γ) confident that the mean of group one is statistically larger than the mean of group two. Otherwise, the difference is not significant.

When the data type is binomial, Eq. (107) can be rewritten as Eq. (108).

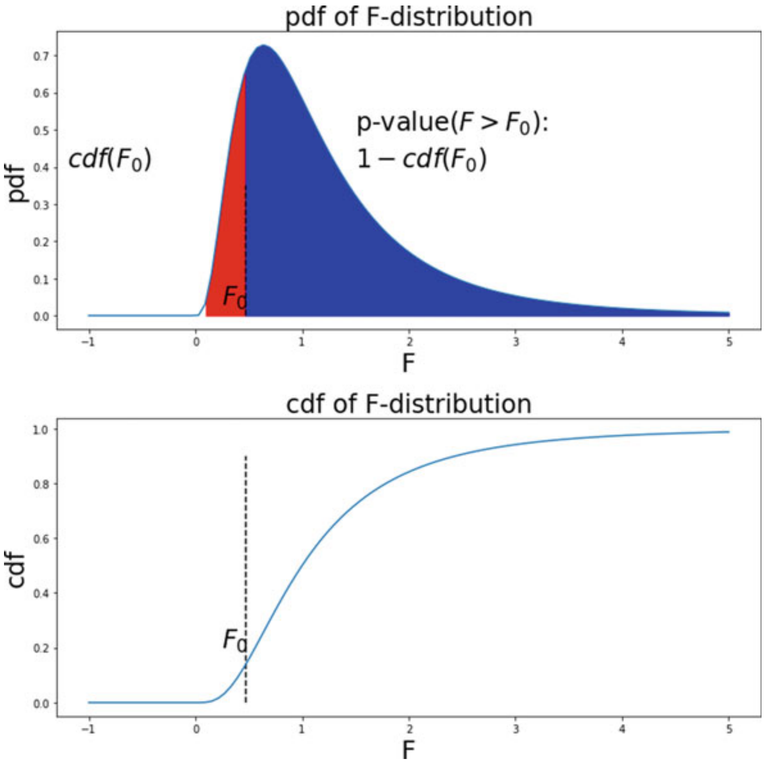


Fig. 33 F-distribution with $\nu_1 = \nu_2 = 9$

$$Z = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \tag{108}$$

ANOVA

Equation (109) is the probability density function of a F-distribution. ν is the number of degrees of freedom, B is the Beta function, and F is the F-value or statistics. “1” means numerator and “2” means denominator (Fig. 33).

$$f(F, \nu_1, \nu_2) = \frac{\nu_2^{\nu_2/2} \nu_1^{\nu_1/2} F^{\nu_1/2-1}}{(\nu_2 + \nu_1 F)^{(\nu_1+\nu_2)/2} B(\nu_1/2, \nu_2/2)} \tag{109}$$

To evaluate whether the means of K groups are equal, one-way analysis of variance (ANOVA) can be conducted by using Eq. (110). ν is the number of degrees of freedom. “bg” means between groups. “wg” means within a group. “SS” means sum of squared deviations. “N_i” means sample size per group. “M_i” means the mean per group. "M_T" means the mean across all of the groups. F is F-statistics or value. The assumptions behind ANOVA include normal distribution and equal standard deviation within groups. If samplings are from the same population, the between group variation should be smaller than within-group variation. A larger F-statistics implies that the samples are drawn from populations with different mean values. If F-statistics >100($1 - \gamma$)th percentile, we are $100(1 - \gamma)$ confident that the means of K groups are statistically different. This can be evaluated by using Python's SciPy package to compare F-statistics and Scipy.stats.f.ppf($1 - \gamma, \nu_{\text{bg}}, \nu_{\text{wg}}$). Alternatively, p-value can be calculated by using 1-Scipy.stats.f.cdf(F-statistics,$\nu_{\text{bg}}, \nu_{\text{wg}}$). If p-value $< \gamma$, we are $100(1 - \gamma)$ confident that the means of K groups are statistically different.

$$F\left(\nu_{\text{bg}}, \nu_{\text{wg}}\right) = \frac{\frac{\text{SS}_{\text{bg}}}{\nu_{\text{bg}}}}{\frac{\text{SS}_{\text{wg}}}{\nu_{\text{wg}}}} \tag{110}$$

$$\nu_{\text{bg}} = K - 1 \tag{111}$$

$$\nu_{\text{wg}} = \sum_{i=1}^{K} N_i - K \tag{112}$$

$$\text{SS}_{\text{bg}} = \sum_{i=1}^{K} N_i (M_i - M_T)^2 \tag{113}$$

$$\text{SS}_{\text{wg}} = \sum_{i=1}^{K} \sum_{j=1}^{N_i} \left(x_j^{(i)} - M_i\right)^2 \tag{114}$$

$$M_i = \frac{\sum_{j=1}^{N_i} x_j^{(i)}}{N_i} \tag{115}$$

$$M_T = \frac{\sum_{i=1}^{K} N_i M_i}{\sum_{i=1}^{K} N_i} \tag{116}$$

4.4.2 Variance

Chi-Square

Equation (117) is the probability density function of a χ^2-distribution. ν is the number of degrees of freedom, Γ is the Gamma function, and χ is the χ^2-value or statistics (Fig. 34).

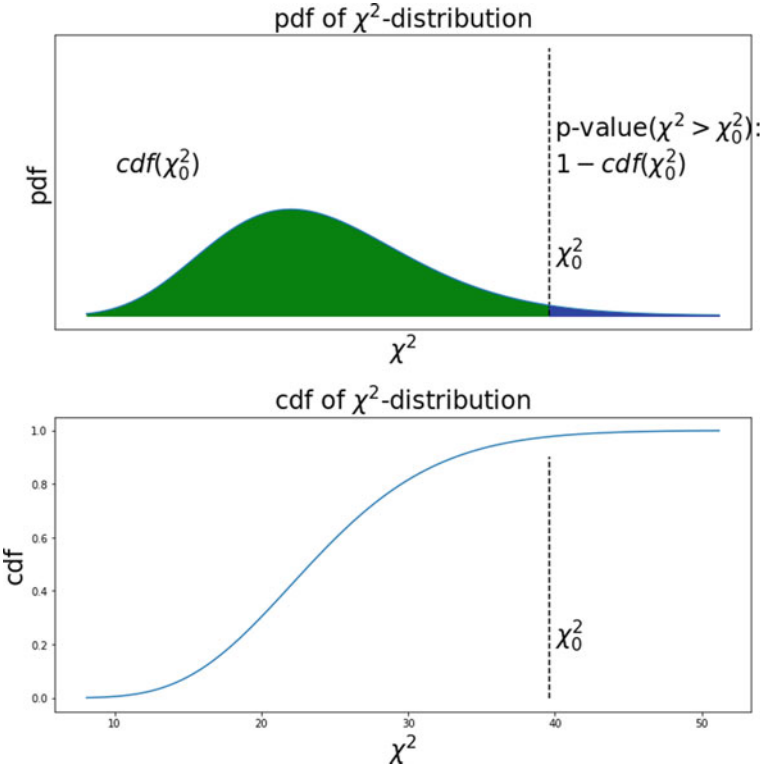


Fig. 34 χ^2-distribution with $\nu = 24$

$$f(\chi, \nu) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} \chi^{\nu/2-1} e^{-\chi/2} \tag{117}$$

To check whether a sampling variance (s) is equal to a specified variance (σ), Eq. (118) can be used.

$$\chi^2 = \nu \frac{s^2}{\sigma^2} \tag{118}$$

$$\nu = n - 1 \tag{119}$$

To evaluate whether the sampling variance and specified variance are statistically equal, we can do a two-sided Chi-square test. If the χ^2-statistics falls between the $100\gamma/2$th percentile and $100(1 - \gamma/2)$th percentile, we are $100(1 - \gamma)$ confident that the sampling variance and specified variance are statistically equal. Otherwise, they are statistically different. This can be evaluated by using Python SciPy package to check whether the following relationship is met or not, scipy.stats.chi2.ppf($\gamma/2$, ν) < χ^2-statistics < scipy.stats.chi2.ppf($1 - \gamma/2$, ν).

To evaluate whether the sampling variance is statistically smaller than the specified variance, we can check the one-sided p-value. If the p-value for χ^2-statistics, cdf(χ^2-statistics), $< \gamma$, we are $100(1 - \gamma)$ confident that the sampling variance is statistically smaller than the specified variance. Otherwise, the difference is not significant. This can be evaluated by using Python SciPy package to check whether the following relationship is met or not, scipy.stats.chi2.cdf(χ^2-statistics, ν) $< \gamma$.

To evaluate whether the sampling variance is statistically larger than the specified variance, we can check the one-sided p-value. If the p-value for χ^2-statistics, 1-cdf(χ^2-statistics), $< \gamma$, we are $100(1 - \gamma)$ confident that the sampling variance is statistically larger than the specified variance. Otherwise, the difference is not significant. This can be evaluated by using Python SciPy package to check whether the following relationship is met or not, 1-scipy.stats.chi2.cdf(χ^2-statistics, ν) $< \gamma$.

F-test

The pdf of F-distribution was discussed in Eq. (109).

To compare the variances of two sampling groups, Eq. (120) can be used. "s" means unbiased standard deviation. Subscripts "1" and "2" represent group 1 and group 2.

$$F = \frac{s_1^2}{s_2^2} \tag{120}$$

$$\nu_1 = n_1 - 1 \tag{121}$$

$$\nu_2 = n_2 - 1 \tag{122}$$

To evaluate whether the variances of two sample groups are statistically equal, we can do a two-sided F-test. If the F-statistics falls between the $100\gamma/2$th percentile and $100(1 - \gamma/2)$th percentile, we are $100(1 - \gamma)$ confident that the variances of two sample groups are statistically equal. Otherwise, they are statistically different. Alternatively, we can also look at the p-value, which is equal to cdf(1/F-statistics) + 1-cdf(F-statistics) assuming F-statistics > 1. If the p-value $< \gamma$, we are $100(1 - \gamma)$ confident that the variances of two sample groups are statistically different. Otherwise, they are statistically equal.

To evaluate whether the variance of group one is statistically smaller than the variance of group 2, we can check the one-sided p-value. If the p-value for F-statistics, cdf(F), $< \gamma$, we are $100(1 - \gamma)$ confident that the variance of group one is statistically smaller than the variance of group two. Otherwise, the difference is not significant.

To evaluate whether the variance of group one is statistically larger than the variance of group two, we can check the one-sided p-value. If the p-value for F-statistics, 1-cdf(F), $< \gamma$, we are $100(1 - \gamma)$ confident that the variance of group one is statistically larger than the variance of group two. Otherwise, the difference is not significant.

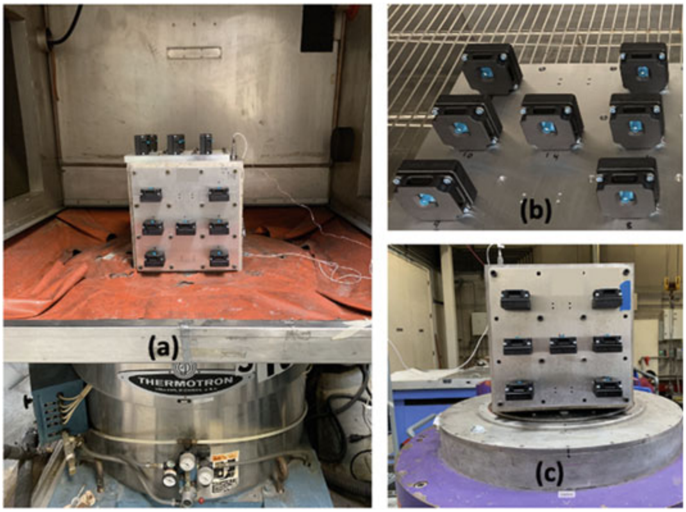


Fig. 35 **a** Combined environment reliability test (CERT), **b** temperature humidity cycle (THC) test, and **c** random vibration test

4.4.3 Case Study

Camera Connector Sealing Material Down Selection

Camera plays a crucial role in autonomous vehicles. When cameras are exposed to the environment, water ingress and dust ingress can be potential concerns. During camera development, a series of reliability tests were developed to down select the adhesives to seal the camera connector. Figure 35a shows a combined environment reliability test (CERT), which was composed of a thermal shock test (−40 °C => 85 °C, 12 cycles) and a customized random vibration test in parallel. Figure 35b, c show the waterfall test, which was composed of a temperature humidity cycle (THC) test (−40 °C/uncontrolled humidity => 85 °C/85%RH,14 cycles) and a customized random vibration test in sequential order. To save test time and utilize a vertical shaker, a metal cube was used during the vibration test. Cameras attached to different surfaces of the metal cube experienced different vibration directions. By rotating the metal cube, cameras went through all of the vibration directions. Before and after the reliability test, pressure leakage of the camera was monitored. If the pressure drop or leakage rate were too big, the connector sealing was defined as a failure. Table 25 summarized the test result. Cameras with silicone sealing did not fail (failure rate $p_1 = 0\%$), while cameras with polyurethane sealing had 5 failures (failure rate $p_2 = 35.7\%$).

Based on Eq. (103), the *t*-statistics was −2.687 and the degree of freedom is 13. The *p*-value can be calculated by using Python's SciPy package, Scipy.stats.t.cdf(−2.687, 13) ~ 0.0093. Because the *p*-value is less than 0.05, we

Table 25 Camera connector sealing material down selection

Test type	Adhesive	Sample size	Number of failures
Waterfall	Silicone	7	0
	Polyurethane	7	4
CERT	Silicone	7	0
	Polyurethane	7	1

Table 26 Wire pull strength (N)

Unit	1	2	3	4	5	6	7	8	9	10
Terminal 1	21.4	26.6	27.2	18.6	15.1	21.8	22.3	21.7	19.8	20.8
Terminal 2	13	17.5	19.5	14.2	17.3	17.4	17.3	17.4	28.7	18

are 95% confident that cameras with silicone sealing have a smaller failure rate than those with polyurethane sealing.

Based on Eq. (108), the Z-statistics was −2.789. The p-value can be calculated by using Python's SciPy package, Scipy.stats.norm.cdf(−2.789) ~ 0.0026. Because the p-value is less than 0.05, we are 95% confident that cameras with silicone sealing have a smaller failure rate than those with polyurethane sealing.

Power Supply Wire Pull Strength

Pull test was used to down select two types of wire terminal design. Table 26 summarized the wire pull strength results. The mean pull strength for Terminal Type 1 is 23.61 N. The mean pull strength for Terminal Type 2 is 19.83 N.

Based on Eqs. (99) and (100), the t-statistics was 2.02019 and the degree of freedom was 17.49064. The p-value can be calculated by using Python's SciPy package, 1-Scipy.stats.t.cdf(2.02019, 17.49064) ~ 0.0295. Because the p-value is less than 0.05, we are 95% confident that the pull strength of Terminal Type 1 is larger than that of Terminal Type 2.

Based on Eq. (107), the Z-statistics was 2.422. The p-value can be calculated by using Python's SciPy package, 1-Scipy.stats.norm.cdf(2.422) ~ 0.0077. Because the p-value is less than 0.05, we are 95% confident that the pull strength of Terminal Type 1 is larger than that of Terminal Type 2.

The calculations above are based on equal variance. In fact, F-test can be used to check whether the variances from two designs are statistically significantly different. Based on Eq. (120), F-statistics was 0.7085. The degrees of freedom are 9 and 9. The 2.5th percentile can be calculated by using Python's SciPy package, Scipy.stats.f.ppf(0.025,9,9) = 0.2484. Similarly, the 97.5th percentile can be calculated by using Scipy.stats.f.ppf(0.975,9,9) = 4.0260. Because 0.7085 is between 0.2484 and 4.0260, we are 95% confident that the variances of two terminal types are statistically equal. This can be further validated by checking the p-value.

Scipy.stats.f.cdf(1/1.4116,9,9) + 1-Scipy.stats.f.cdf(1.4116,9,9) = 0.6159. Because 0.6159 > 0.05, we are 95% confident that the variances of two terminal types are statistically equal.

5 Failure Analysis (FA) and Corrective/Preventive Actions (CAPA)

5.1 Introduction

During reliability testing, failures will likely occur. In some sense, one can argue that the main purpose of reliability testing during the development phases is to produce failures more efficiently. While design engineers may abhor them, failures are of incalculable value to reliability engineers since that is the source of all the learnings. Failures will also come from the field as the product gets launched for a variety of reasons. This is to be expected even with a perfect reliability development process. After all, the goal of reliability is not to completely eliminate field failures, but rather to maintain the failure rate to an acceptable level based on cost and schedule trade-offs. In reality, of course, the reliability development process will never be perfect and new failure modes will almost always pop up unexpectedly, indicating a gap in our understanding. Thus, performing detailed FA and identifying CAPA on these failures will help reliability engineers better refine the process and the tests, improve the design, and prevent similar issues in the future.

To achieve the goal of identifying root cause and solutions, a systematic approach is therefore needed for FA/CAPA [43]. This will help reduce the cycle time and minimize waste, leading to robust products with low cost and short development time.

5.2 General Process Flow

The proposed FA/CAPA process flow in Fig. 36 generally follows the widely adopted 8D problem-solving methodology [44]. Each step will be covered in detail in later sections, but here is a general overview:

- Failure confirmation is to confirm the failure symptoms reported before proceeding any further. It is a rather straightforward yet very critical step.
- Determining the exact failure components helps narrow down the scope of the issue and ensure the team focuses on the right problem. This is achieved in the second step of failure component identification.

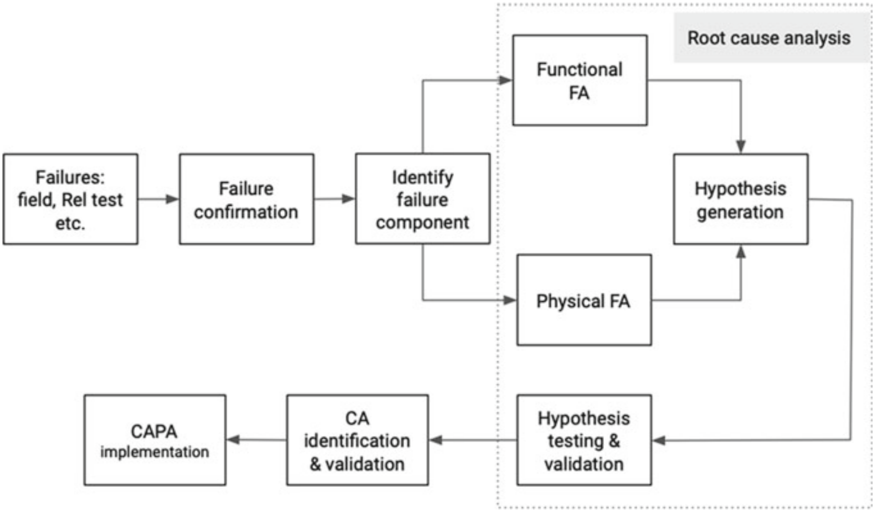

Root cause analysis
Functional FA
Failures: field, Rel test etc.
Failure confirmation
Identify failure component
Hypothesis generation
Physical FA
CAPA implementation
CA identification & validation
Hypothesis testing & validation

Fig. 36 FA/CAPA process (CA = corrective action, CAPA = corrective/preventive actions)

- Root cause analysis uses all available FA tools including functional testing, physical analysis, hypothesis generation, validation, etc., to isolate the true root cause.
- After the root cause is identified and validated, corrective actions will be generated and validated to confirm its effectiveness in resolving the issue.
- Implementing corrective and preventive actions is the last step to ensure the learnings are incorporated into actual products. Purpose of the preventive actions is to apply the learnings proactively to other relevant products.

5.2.1 Failure Confirmation

The first step in any FA work is always failure confirmation. This is not only to confirm the unit indeed failed, but also the same failure symptoms are replicated. While it may seem obvious, it is actually very critical in the sense that it can prevent a lot of waste. Over 50% of consumer electronics, for example, returned from the field are actually NTFs (no trouble found) for a variety of reasons including user errors, improper settings, abuse of return policy, etc. Therefore, isolating the true failures will have significant implications in terms of cost, engineering resources, equipment times, etc. Another important aspect of failure confirmation is to ascertain that the symptoms are indeed as reported. This is especially true for field failures since users and field personnel generally lack the technical know-how and can easily get confused with different issues. Confirming the failure symptoms helps better define the problem and thus ensure FA activities are properly focused.

5.2.2 Identify Failure Component

After a failure is confirmed, the next step is to isolate the problem further down to component level. Since a unit, e.g., PCBA, sometimes can have thousands or more components, it requires thorough technical understanding of the system. Many times, it requires several rounds of gradually narrowing down the problem. For example, when the compute module inside a self-driving car fails, one may first have to determine whether the GPU, motherboard, power supply, etc., failed by replacing them with known good ones, then isolate the problem further by separating analog circuit issues from digital ones, and then lastly determine which specific components failed.

It is always a good idea to replace the failed components with known good ones and retest them. This is to ensure the said components indeed caused the failure symptoms reported. It is quite possible that a failure event can cause multiple components to fail. Or worse, one component failure can cause other components to fail. In such cases, failure to identify all failed components can miss the real failure component and therefore lead to trouble in the next step of root cause analysis.

5.2.3 Root Cause Analysis

Root cause analysis (RCA) is generally one of the most critical and challenging steps in the whole FA flow. Similar to real-world detective work, one needs to be very thorough and methodical in collecting and analyzing evidence, generating, and testing all hypotheses before a conclusion with high confidence can be reached.

During the evidence collection and analysis phase, reliability and FA engineers utilize a lot of functional and physical measurement tools to gather as much relevant data as possible. Functional measurement data at component level includes curve tracing, studying timing information, measuring impedance versus frequency, testing component performance under high temperature, high humidity conditions, etc. Physical analysis includes optical/X-ray inspection, morphology study, material analysis, cross sectioning, etc. The goal is to collect all the relevant evidence that is needed for forming a hypothesis list in the next step. Currently, there are so many FA tools available and will not be discussed in detail here.

With all the functional and physical information, it is now time to come up with a list of hypotheses that are consistent with all the symptoms observed. Such a list needs to be generated together with the design and manufacturing engineering teams to ensure its comprehensiveness. This will also get their buy-in to help during the hypothesis validation process. For each hypothesis on the list, corresponding actions need to be defined. The actions will then need to be followed up such that each hypothesis can be either rejected or proved. Sometimes, the list can be too long to be taken on in parallel. Then, prioritization based on the team's consensus is a good idea. This allows the investigation effort to be focused on the high priority one first.

To fully validate the root cause, it is generally a good idea to recreate the same failure symptoms once the root cause candidates are identified. Without going through such a step, there is always the danger that some critical contributors are missed. After completing this step, there should be high confidence that the true root causes have been identified and it is time to define the corrective actions needed to fix the problem.

5.2.4 Containment and Corrective Actions

Containment is short-term solution to minimize the reliability risks to customers before the long-term corrective actions can be implemented. For example, if a new test is developed and demonstrated to be effective in capturing the failures before shipping, the containment action would be to use the test to screen failures out. For containment action to be effective, it needs to be comprehensive. Using the same example above, not only new builds need to be screened, but all the ones in inventory. A decision would also have to be made if the units already deployed in the field need to be recalled for screening. Such a decision can never be taken lightly, of course, given the significant cost and reputation implications. In such a scenario, a risk assessment utilizing stress-strength analysis, as discussed in Sect. 2, is typically warrantied to quantify the risk-reward trade-off and facilitate decision-making by management.

Corrective actions are the long-term solution to the issue by eliminating the true root cause. Once the true root causes are identified, the team needs to determine the proper actions to fix it through either design, material, or process changes. Effective corrective actions, however, require thorough validation. The goal of the validation is two folds. First, it confirms that the underlying root causes are indeed resolved. Needless to say, implementing a solution only to realize later that the problem is not solved is a tremendous disservice to the reliability team and the company. Second, the validation also helps assure the team that no other unintended problems arise because of the changes. While seemingly straightforward, it is generally not easy to do in practice given the enormous pressure the team would no doubt be under. But all of us have heard stories from the news about companies that claimed to have fixed a field issue, but have to retract it later. It falls upon the reliability team to insist upon performing thorough validation on the corrective actions. Hopefully, it is also obvious to the reader now that one important value of effective containment action is to buy the team precious time to get the corrective actions ready.

5.2.5 Preventive Action Implementation

The purpose of preventive action is to apply the same learning from the root cause and corrective action investigation to other products that are potentially subjected to the same failure mode. Needless to say, this has significant value for the company. Too many times, the author has seen people from the same company learn the same lesson repeatedly due to lack of communication or poor information sharing. It falls upon the reliability engineer to broadcast the new learnings to the entire engineering community and work together with each product team potentially affected to implement relevant design and process changes.

It is also a good idea to document the whole learning experience thoroughly and put them in FRACAS database for future references. It is very easy to get the learnings lost as engineering team members come and go. The value of the information accumulated over time cannot be overstated not only for preventing future failure, but also as a training resource for any new team members.

5.3 Case Study

5.3.1 Power Supply Controller Failure Due to Solder Bridging

During temperature cycling (TC) testing of a power supply module, ~30% of the units were observed to fail with no output voltage after 200 cycles. The failures were then taken out of the chamber and the symptoms confirmed at the bench level. Further functional and electrical FA were able to isolate the failure to the controller IC. In fact, resistance measurements showed two of the IC pins were electrically shorted. As indicated in Fig. 37a, visual inspection suggests solder bridging as a potential cause.

In order to validate the hypothesis, two more actions were taken: (1) X-ray imaging of the failure area (Fig. 37b). The solder bridging between the two leftmost pins can now be easily observed. (2) Take a good unit and manually short the same two pins. Same failure symptoms were observed. These results convinced the reliability engineer that solder bridging is indeed the root cause.

As a short-term containment action, the vendor agreed to add 100% optical inspection as part of the process flow to prevent defective material from escaping. Existing materials already shipped are also returned to the vendor for screening. For corrective action, the vendor agreed to reoptimize the solder paste printing and reflow process

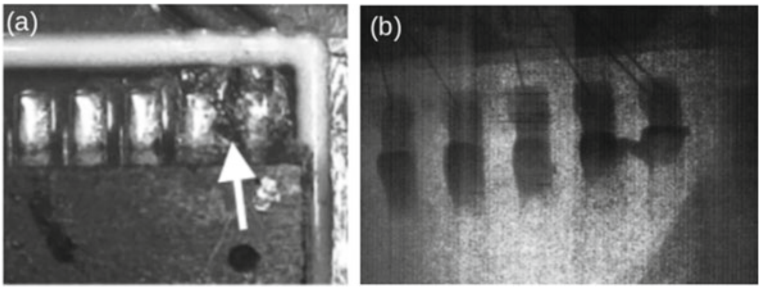


Fig. 37 **a** Optical inspection and **b** X-ray inspection of solder bridging failure

to resolve the issue. The learnings are also applied to other electronic products in the company as preventive action.

5.3.2 Waterblock Coolant Leakage

Due to the massive computing activities, hybrid cooling with air cooling and liquid cooling has been introduced to autonomous vehicles. Cold plates are attached to CPUs, GPUs, and power systems. Liquid coolant circulates through cold plates and carries heat away. When several GPUs are bundled in parallel to enhance the computing capability, manifolds are used to integrate individual waterblock together and simplify coolant loop design. Figure 38 is one off-the-shelf (OTS) manifold/waterblock structure. Manifold is sitting at the top of several waterblocks. EPDM gaskets are sandwiched between manifold and waterblock. When the screws are tightened, the EPDM gaskets are compressed which generate a repulsive force on the manifold/screw. As indicated in Eqs. (123)–(125), this repulsive force determines the torque of the screw, which is proportional to the reciprocal of EPDM thickness.

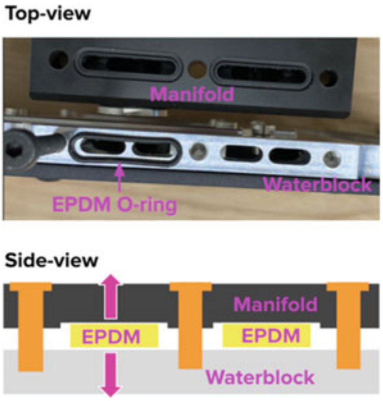
$$\sigma = E \times \varepsilon = E \times \frac{L - H}{L} \tag{123}$$

$$F_{\mathrm{EPDM}} = \sigma \times A = \frac{E \times A \times (L - H)}{L} \tag{124}$$

$$T = c \times D \times F = c \times D \times \frac{2 \times F_{\mathrm{EPDM}}}{3} = C_0 - C_1 \times \frac{1}{L} \tag{125}$$

In these equations, σ is EPDM stress, E is EPDM modulus, L is initial EPDM thickness, H is final EPDM thickness, F_{EPDM} is EPDM force, A is EPDM contact area with waterblock/manifold, T is torque of screw, c is friction coefficient, F is screw tension force, D is screw diameter, and C_0 and C_1 are two constant coefficients.

Fig. 38 Waterblock/manifold



As indicated in Fig. 39, coolant leakage surrounding the joints between manifold and waterblock was observed after the temperature cycle test of one computing system.

As shown in Fig. 40, Fishbone diagram was used to brainstorm potential root causes. The leading hypothesis is the compression set and permanent plastic deformation of EPDM gasket at high temperature. When compression set or permanent plastic deformation occurred, the equivalent initial thickness in Eq. (125) was reduced, decreasing the torque of the screw. Once the torque of the screw was reduced, the EPDM gasket might not be compressed enough to seal the gap between manifold and waterblock.

To validate this hypothesis, a series of manifold/EPDM/waterblock sandwich samples were loaded into a thermal oven for a high temperature storage test at 55 °C. At different test durations, different samples were pulled out of the oven for characterizations, including EPDM thickness measurement by optical microscope and screw torque measurement by torque screwdriver. Figure 41a shows that the logarithm of mean EPDM thickness decreases linearly as a function of high temperature storage test time. In other words, the EPDM thickness follows exponential distribution. This data supported our hypothesis—EPDM has compression set at high temperature.

Figure 41b shows that the mean of screw torque decreases linearly as a function of the reciprocal of mean EPDM thickness. This is consistent with Eq. (125) and also

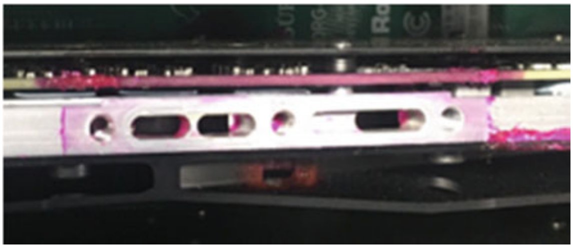


Fig. 39 Coolant leakage in waterblock/manifold (pink color)

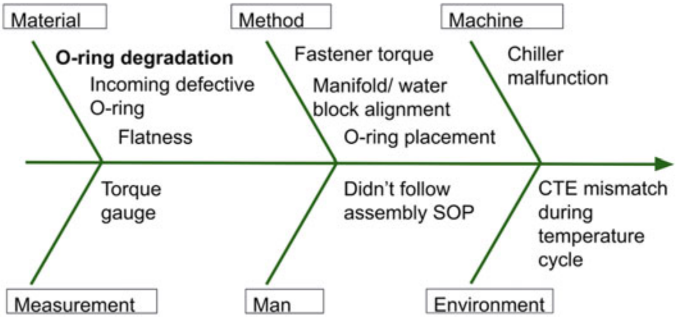


Fig. 40 Fishbone diagram of waterblock coolant leakage

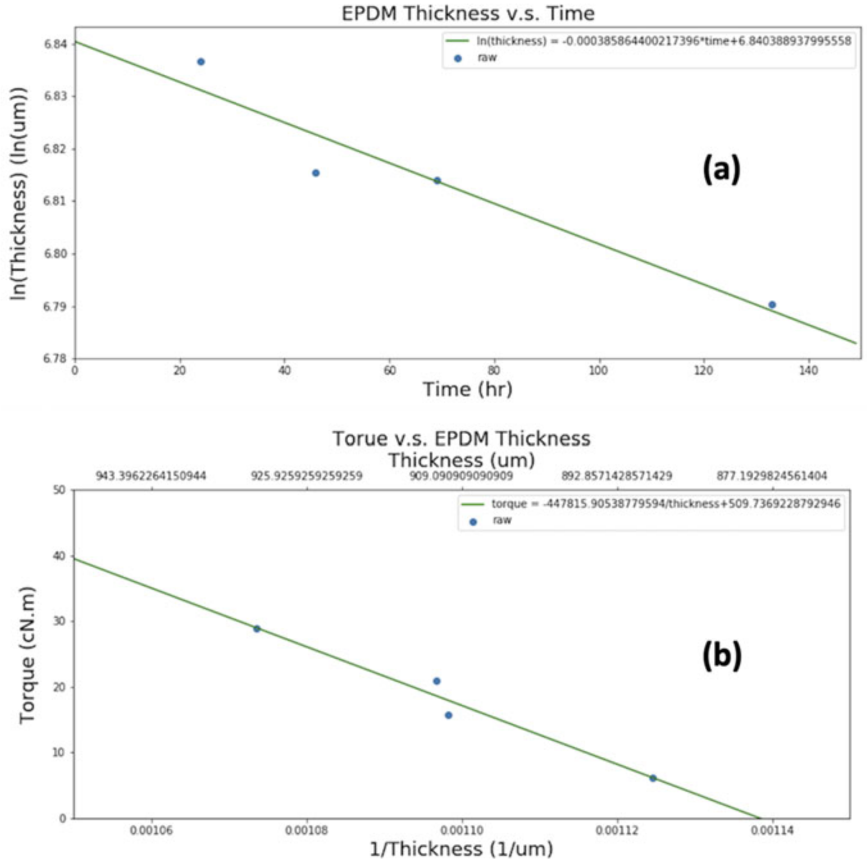


Fig. 41 **a** EPDM gasket thickness versus time at 55 °C, **b** screw torque versus reciprocal of EPDM thickness

supported our argument above—screw torque decreased when EPDM had thickness reduction induced by compression setting at high temperature.

Because gasket material cannot be changed in the short term, one process mitigation adopted was to use higher initial screw torque which can compensate for the impact of EPDM compression set.

6 System Level Reliability and Modeling

6.1 Introduction to Reliability Block Diagram (RBD)

With all modules' reliability performances quantified, one should be able to model and predict system level reliability. This is especially valuable in cases where system level data is expensive or impossible to collect which is the case for self-driving cars. Having the ability to model will also provide critical insight during earlier development when the entire system is not yet available.

To model system level reliability, two methods, reliability block diagram (RBD) and fault tree analysis (FTA), are generally used [11, 45]. RBD represents the modules or failure modes with blocks and is defined in success space. FTA, on the other hand, uses nodes and is defined in failure space. Since RBD is easier to construct and understand for system reliability analysis, we will only cover it here.

There are several different types of RBD configurations normally used including series, parallel, *k*-out-of-*n*, and complex systems, etc. We will cover each of them as well as the dynamic system model in detail in the ensuing sections. One thing to note is that RBD configuration can be quite different from functional block diagrams (FBD). For example, all 4 tires on a self-driving car are functionally parallel to one another, but for RBD, they are actually in series since the failure of any would cause the car to fail.

6.2 Series Configuration

When all modules have to function in order for the entire system to work, the system is considered to be in series configuration reliability wise. In this case, the reliability of the system is simply the product of the reliability of all components. Assuming R_i represents the reliability of the ith component, the system reliability R_{S} pictured in Fig. 42 is:

$$R_S = R_1 \times R_2 \times \cdots \times R_n \tag{126}$$

For example, in a three-components system, if R_1, R_2, and R_3 are 99.9, 98.9, and 98.5%, then the system reliability is:

$$R_S = R_1 \times R_2 \times R_3 = 99.9\% \times 98.9\% \times 98.5\% = 97.3\% \tag{127}$$

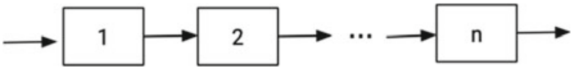
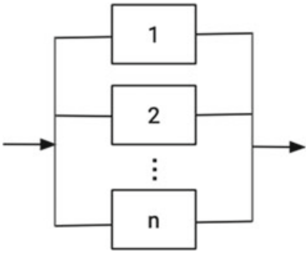


Fig. 42 Series RBD configuration

Fig. 43 Parallel RBD configuration



Since each component's reliability is always less than 1, overall system reliability is lower than that of the least reliable one. Adding additional components will further reduce the system reliability. It is always a good idea to use as few components as possible during the design phase.

6.3 Parallel Configuration

If only the failure of all the units causes the failure of the system, the arrangement is considered a parallel configuration. Assuming R_i, again, represents the reliability of the ith component, then system reliability R_S pictured in Fig. 43 is:

$$R_S = 1 - (1 - R_1) \times (1 - R_2) \times \cdots \times (1 - R_n) \tag{128}$$

For example, in a three-components system, if R_1, R_2, and R_3 are 99.9, 98.9, and 98.5%, then the system reliability is:

$$R_S = 1 - (1 - R_1) \times (1 - R_2) \times (1 - R_3)$$

$$= 1 - 0.1\% \times 1.1\% \times 1.5\% = 99.99998\% \tag{129}$$

Different from that of the series one, overall system reliability for parallel configuration is higher than the most reliability component. One caution as noted earlier is that parallel reliability configuration is different from parallel functional configuration. One has to fully understand how the system works in order to determine whether true reliability parallelism exists.

6.4 k-Out-of-n Configuration

Another configuration that reliability engineers will run into is k-out-of-n, where at least k modules out a total of n parallel ones are needed to make the system work,

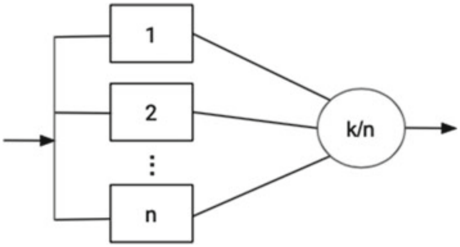


Fig. 44 *k*-out-of-*n* configuration

e.g., airplane engines. For simplicity reasons, assume all components are identical with a reliability of R as pictured in Fig. 44. The system reliability R_S is then:

$$R_s = \sum_{r=k}^{n} \binom{n}{r} R^r (1 - R)^{n-r} \tag{130}$$

For example, in a 2-out-of-4 system, assume R = 84%, then the system reliability is:

$$R_s = \sum_{r=2}^{4} \binom{4}{r} 0.84^r (1 - 0.84)^{4-r} = 0.986 \tag{131}$$

Both series and parallel configurations are special cases of *k*-out-of-*n*. When $k = n$, all modules have to work for the system to function and, therefore, it becomes series configuration. When $k = 1$ and $n > 1$, it then becomes parallel configuration.

The mathematical analysis for such a system with non-identical components is very complicated and will not be covered here. Interested readers can refer to [46] for more details.

6.5 Combined Configurations

Any of the aforementioned configurations can be used simultaneously in a system. In such cases, reliability of each individual configuration can be separately calculated first and then combined to produce the system reliability results. For example, the system illustrated in Fig. 45 is a combination of series and parallel configuration. Assume R_1, R_2, and R_3 are 99.9, 98.9, and 98.5%, we first calculate the reliability of branch R_{12} which is a serial configuration:

$$R_{12} = R1 \times R2 = 99.9\% \times 98.9\% = 98.8\% \tag{132}$$

We can now calculate the reliability of the parallel system between R_{12} and R_3 as:

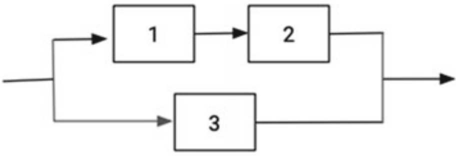


Fig. 45 Combination of series and parallel configurations

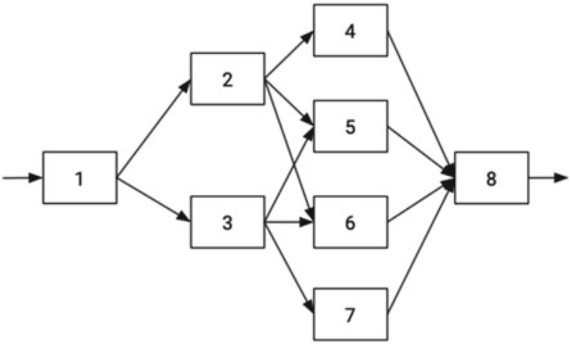


Fig. 46 Complex reliability system

$$R_S = 1 - (1 - 98.8\%) \times (1 - 98.5\%) = 99.98\% \tag{133}$$

6.6 Complex Reliability Systems

There are some complex system configurations that cannot be simply represented by anyone that we discussed previously. A good example of such a system would be a network like the one illustrated in Fig. 46.

For such complex systems, special methods are needed in order to calculate the reliability. These methods include decomposition or Bayes' theorem method, event space, and path-tracing or mini cut set method, etc. Details of each method are outside the scope of this book chapter, but interested readers can refer to [47] for more information. In most cases, a manual solution is not realistic and specialized computer software programs would be needed.

6.7 Dynamic System Reliability Models

Previous discussions are all for static system reliability calculation where the reliability values of all components are at a specific given time. A dynamic system reliability model would be needed to incorporate time dependency. Fortunately, for the above configurations, we can simply introduce a time variable into the corresponding formula to get the dynamic model. For example, for series configuration discussed in Eq. (126), the formula would become:

$$R_S(t) = R_1(t) \times R_2(t) \times \cdots \times R_n(t) \tag{134}$$

However, there are special dynamic cases that reliability engineers will encounter, not covered by the ones discussed above. These would include load sharing and standby redundant systems, etc.

For the load sharing case, two or more components are connected in parallel. Under normal circumstances, each component will share the load equally and is derated. If a subset of the components fails, the total load will be redistributed to the remaining ones. With higher load now, the remaining functional ones will be subjected to a higher failure rate than normal. Thus, a different system reliability model would be needed to take into consideration the effect of failure rate change.

For standby redundant systems, as its name implies, there are components on standby mode and will only be switched on when the primary components fail. An example of a standby system would be the backup LV battery inside a self-driving car that supplies power to safety–critical units when the main HV battery circuits fail.

To analyze such dynamic systems, one needs to know or assume failure distributions of all the components. In the case of a standby redundancy system, the reliability of the sensing and switching system would also need to be included. To simplify the calculation, we are going to assume the same exponential failure distribution for all components in this section. Non-identical components or other failure distributions, e.g., Weibull or Lognormal, can be incorporated using the same logic.

For load sharing case, let us assume two units in the system with a constant failure rate of λ_1 under normal operating condition when both are working. If one fails, the other would continue to operate with the entire load at a constant failure rate λ_2 where $\lambda_2 > \lambda_1$. For a mission time of T, the system would be considered a success if (1) both components operate successfully for the entire duration at failure rate λ_1 or (2) one fails at time $t < T$ and the remaining component continues to function for a duration of $T - t$ at a higher failure rate of λ_2, and thus, we get the following formula:

$$R_s(t) = e^{-2\lambda_1 t} + \frac{2\lambda_1\left(e^{-\lambda_2 t} - e^{-2\lambda_1 t}\right)}{2\lambda_1 - \lambda_2} \text{ if } 2\lambda_1 - \lambda_2 \neq 0 \tag{135}$$

For the case where $2\lambda_1 - \lambda_2 = 0$, the above equation simplifies to:

$$R_s(t) = e^{-2\lambda_1 t} + 2\lambda_1 t e^{-\lambda_2 t} \tag{136}$$

For standby redundant systems, let us assume there are only 1 primary and one standby component as illustrated in Fig. 47. In this case, the system at mission time T is considered a success if either component #1 operates successful for the entire

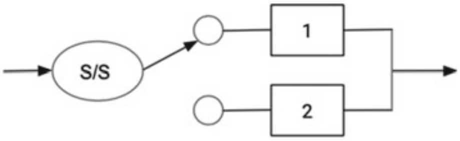


Fig. 47 A standby redundant system example

duration, or if it fails at time $t < T$, but component #2 operates successfully for the remaining duration of $T - t$. Thus, system reliability can be calculated as:

$$R_s(t) = e^{-\lambda t}(1 + R_{\mathrm{S/S}}\lambda t) \tag{137}$$

where $R_{\mathrm{S/S}}$ refers to the reliability of the sensing and switching mechanism.

7 Repairable System

7.1 Introduction

For repairable systems, a different set of reliability metrics are needed to assess its robustness [45]. The main goal for such a system is to ensure the control of failures. If the frequency and nature of these failures can be accurately predicted, various types of strategies can be set up to minimize the impact of these failure events to customers. These would include different types of redundancy, fault tolerance methods, preventive maintenance plan, spare part management, efficient repair processes, etc. Reliability for repairable systems is typically measured with three parameters: mean-time-between-failure (MTBF) measuring the frequency of failure events, mean-time-to-repair (MTTR) characterizing how quickly the system can be repaired once a failure event happens, and availability (A), the percentage of time that the system is either operating or capable of operating. In the ensuing sections, we will only define these three parameters and give examples on how to calculate them. Readers who are interested in learning more about how to use these metrics to define system requirements, optimize design, and improve repair process, can refer to Ref. [16].

7.2 Mean Time Between Failure (MTBF)

MTBF is defined as the average duration between consecutive failures. Mathematically, it is the ratio of the cumulative observed time divided by the total number of failures:

$$\text{MTBF} = \frac{\text{total operating time}}{\text{number of failures}} \tag{138}$$

MTBF is independent of time only if the failures follow exponential distribution. In such cases, MTBF is simply the inverse of failure rate λ:

$$\text{MTBF} = \frac{1}{\lambda} \tag{139}$$

Assume a reliability laboratory has 5 temperature/humidity chambers of the same type running on average 65 h per week. Over a 2-year period, there have been a total of 12 reported failures. In order to calculate the MTBF, we need to determine the total time first. This can be calculated as 52.14 weeks per year times 65 h per week for each of the chambers. We get the MTBF as:

$$\text{MTBF} = \frac{52.14 \times 65 \times 2 \times 5}{12} = 2824.25\,\text{h} \tag{140}$$

The MTBF of 2824.25 h means that if the laboratory were to purchase another chamber of the same type and use it exactly the same way, one should expect a failure roughly every 43.5 weeks.

7.3 Mean Time to Repair (MTTR)

MTTR measures the efficiency of system repairs once a failure event happens. It is the ratio of the total repair time divided by the number of failures:

$$\text{MTTR} = \frac{\text{total repair time}}{\text{number of failures}} \tag{141}$$

Assuming repair time follows exponential distribution, we can define a repair rate μ:

$$\mu = \frac{1}{\text{MTTR}} \tag{142}$$

Then, the maintainability equation can be written as:

$$M(t) = 1 - e^{-\mu t} \tag{143}$$

$M(t)$ describes the probability that a failure event can be repaired within any given time duration t.

Let us go back to the temperature/humidity chamber example in Eq. (140). Assuming the total repair time for the 12 failure events are 300 h, then MTTR is:

$$\text{MTTR} = \frac{300}{12} = 25\,\text{h} \tag{144}$$

The maintainability of any failure event for 30 h, i.e., the probability of getting the chamber back to the operating condition in less than 30 h is:

$$M(30) = 1 - e^{-\frac{30}{25}} = 69.9\% \tag{145}$$

7.4 *Availability*

Availability here is defined as the percentage of time that a system is operationally capable of performing intended tasks. At a high level, it can be defined as:

$$A = \frac{\text{uptime}}{\text{uptime} + \text{downtime}} \tag{146}$$

However, depending on the definition of downtime and the purpose of the measurement, there are multiple definitions available.

Instantaneous Availability

It is defined as the probability of a system being operational at any specific time t. Assuming exponential distribution for MTBF and MTTR, then mathematically it can be written as:

$$A(t) = \frac{\mu}{\mu + \lambda} + \frac{\lambda}{\mu + \lambda} e^{-(\mu+\lambda)t} \tag{147}$$

where
μ = unit repair rate = 1/MTTR.
λ = unit failure rate = 1/MTBF.

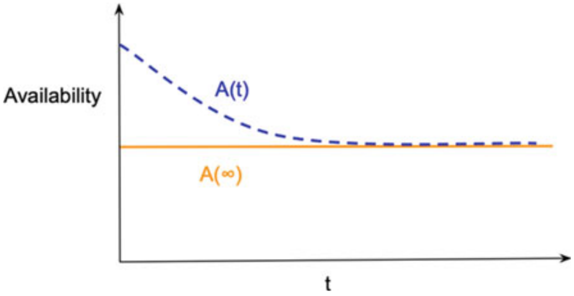
Steady State Availability

Also called inherent availability, this is the most commonly used definition across industries. It is defined as the availability of the system when time goes to infinity:

$$A(\infty) = \lim_{t\to\infty} A(t) = \frac{\text{MTBF}}{\text{MTBF} + \text{MTTR}} \tag{148}$$

The relationship between instantaneous availability and steady state can be seen from Fig. 48.

Fig. 48 Instantaneous availability $A(t)$ versus steady state availability $A(\infty)$



Again, going back to the temperature/humidity chamber example in Eqs. (140) and (144), since MTBF is 2824.25 h and MTTR 25 h, then steady state availability is:

$$A(\infty) = \frac{2824.25}{2824.25 + 25} = 99.1\% \tag{149}$$

Average Availability

Up-time or average availability is the percentage of time from 0 to T that the system is available. Mathematically, it can be expressed as:

$$\overline{A}(T) = \frac{1}{T}\int_{0}^{T} A(t)\mathrm{d}t \tag{150}$$

Achieved Availability

When both corrective and preventive induced downtime are included in the calculation, we get achieved availability:

$$A_A = \frac{\text{MTBE}}{\text{MTBE} + \underline{M}} \tag{151}$$

where MTBE is the mean time between maintenance events, regardless whether it is preventive or corrective. $\underline{M}$ is the mean maintenance downtime including both preventive and corrective activities.

Operational Availability

When all sources of downtime including, e.g., administrative downtime, are included, one gets operational availability. This is the availability that the customer experiences. But since it includes downtime that cannot be controlled by the manufacturer, it is primarily of interest to the customers.

$$A_o = \frac{\text{uptime}}{\text{Total operating cycle time}} \tag{152}$$

8 Summary

In this chapter, some role and responsibility of reliability engineers across the product life cycle are reviewed.

In the beginning, risk assessment methodologies including failure mode and effect analysis (FMEA), fault tree analysis (FTA), and stress-strength analysis are discussed. Camera, Cartop advertising LED display, and cold plate are used as case studies.

Second, accelerated life testing (ALT) and highly accelerated life testing (HALT) are discussed. Widely adopted models for ALT are presented with usage cases explained. Stress profiles customizations, the impact of stress profiles on network and multimedia PCB boards, a magnetic sensor design with mechanical, thermal and humidity stresses, thermal step test of Radar, and voltage step test of LED panel are employed as case studies to provide further insight to the reader.

Third, various reliability statistics topics are covered, including sample size calculation, life distribution analysis by Linear least square regression and Maximum likelihood estimation, confidence interval calculation method, and hypothesis test methods for mean and variance comparison. Lidar bracket crack, glue crack of dash mount audio device, glue crack of cameras, camera connector sealing material down selection, and power supply wire pull strength are used as case study.

Fourth, the topic of failure analysis (FA) and Return Merchandise Authorization (RMA) is presented, starting with general process flow, followed by detailed discussion of each step. Case studies include solder bridging induced power supply controller failure, and waterblock coolant leakage failure.

This chapter ends with system reliability metrics, reliability block diagram (RBD) methods, and repairable system. Various system configurations and reliability analysis approaches are discussed including series, parallel, *n*-out-of-*k*, dynamic system reliability models, etc.

Due to the page limitation, many roles and responsibilities are not discussed here and will be presented separately in the future, including warranty analysis, prognostic health management, etc.

Acknowledgements Authors would like to thank Steven Khau, Hannah Talisse, Marco Marroquín, Alex Granieri, Austin Ng, Azhar Meyer, Ben Rhoads, Catalin Popovici, Corey Bang, David Milton, George Hansel, Helen Lurie, Jason Syu, Kevin Danford, Kevin Page, Leo Hsieh, Lina Tawfik, Minh Nguyen, Randy Mallari, Richard Wimmer, and Vincent Wong from Lyft Inc., for technical discussions and test supports. Special thanks are to Angel Shi and Charlotte Shi for the grammar check.

Editors would like to thank Dongwen Gan from Amazon and Yan Li from Intel for their critical reviews.

References

1. C. Carlson, “Effective FMEAs: Achieving Safe, Reliable, and Economical Products and Processes using Failure Mode and Effects Analysis”, *Wiley*, (2012).
2. “Failure modes and effects analysis (FMEA and FMECA)”, IEC 60812, (2018).
3. AIAG & VDA FMEA Handbook, (2019).
4. “Potential Failure Mode and Effects Analysis (FMEA) Including Design FMEA, Supplemental FMEA-MSR, and Process FMEA”, SAE J1739, (2021).
5. https://relyence.com/.
6. https://www.reliasoft.com/.
7. https://www.plato-e1ns.com/fmea/.
8. https://www.apis-iq.com/.
9. H. Shi, H. Talisse, S. Khau, M. Marroquín, “Hardware reliability in robo-taxi”, 2021 IEEE 71st Electronic Components and Technology Conference (ECTC).
10. Clifton A Ericson II, “Fault Tree Analysis Primer”, *Create Space Independent Publishing Platform,* (2011).
11. Fundamentals of reliability, ReliaSoft (2017).
12. M. Baro-Tijerina, G. Duran-Medrano, “Stress/strength models to estimate systems reliability R(t)=P(x,y)”, International journal of engineering research & technology, vol.7 issue 03, March 2018, pp 356–361.
13. H. W. McLean, “HALT, HASS, and HASA Explained: Accelerated Reliability Techniques”, Revised Edition, ASQ Quality Press (May 12, 2009).
14. J. W. McPherson, “Reliability Physics and Engineering: Time-To-Failure Modeling”, 2nd Edition, Springer (2013).
15. http://reliawiki.org/index.php/Time-Varying_Stress_Models.
16. D. Crowe, A. Feinberg, Design for reliability, CRC press LLC, (2001).
17. L. Coffin, “A Study of the Effects of Cyclic Thermal Stresses on a Ductile Metal,” Transactions of the ASME, Vol. 76, 1954, pp. 931–950.
18. “Failure Mechanisms and Models for Semiconductor Devices”, JEP122F, JEDEC NOVEMBER 2010.
19. https://www.asminternational.org/documents/10192/1849770/05224G_Chapter14.pdf.
20. “Road vehicles — Environmental conditions and testing for electrical and electronic equipment — Part 3: Mechanical loads”, ISO 16750-3 (2007).
21. “General Specification for Electrical/Electronic Components - Environmental/Durability”, GMW3172 (2008).
22. T. Irvine, https://www.vibrationdata.com/.
23. “Road vehicles — Environmental conditions and testing for electrical and electronic equipment — Part 4: Climaticl loads”, ISO 16750-4 (2003).
24. https://www.ncdc.noaa.gov/
25. “Standard Practice for Operating Fluorescent Ultraviolet (UV) Lamp Apparatus for Exposure of Nonmetallic Materials”, ASTM G154 − 12a
26. https://en.wikipedia.org/wiki/Solar_irradiance.
27. https://en.wikipedia.org/wiki/Ultraviolet.
28. B. S. B. Wachler, “Assessment of levels of ultraviolet A light protection in automobile windshields and side windows”, JAMA Ophthalmology, p. 772–775, Vol. 134, Number 7, July 2016.
29. “A choice of lamps for the QUV accelerated weathering tester”, Q-Lab technical bulletin LU-8160, 2019.
30. D. S. Steinberg, “Vibration analysis for electronic equipment”, 3rd edition, John Wiley & Sons, Inc., ISBN 978-0-471-37685-9, 2000.
31. https://www.jmp.com/en_us/home.html
32. https://www.python.org/
33. https://www.r-project.org/

34. https://www.microsoft.com/en-us/microsoft-365/excel
35. W. B. Nelson, “Accelerated Testing: Statistical Models, Test Plans, and Data Analyses”, 1st edition, Wiley-Interscience, ISBN 978-0471697367, 2004.
36. W. B. Nelson, “Applied Life Data Analysis’, 1st Edition, Wiley, ISBN 978-0471094586, 1982.
37. G. Yang, “Life Cycle Reliability Engineering”, 1st edition, Wiley, ISBN 978-0471715290, 2007.
38. B. Bertsche, “Reliability in Automotive and Mechanical Engineering”, Springer, ISBN 978-3540681892, 2008.
39. P. P. O’Connor, A. Kleyner, “Practical Reliability Engineering”, 5th Edition, Wiley, ISBN 978-0470979815, 2012.
40. https://docs.scipy.org/doc/scipy/index.html
41. https://numpy.org/doc/stable/reference/generated/numpy.linalg.inv.html
42. B. L. Amstadter, “Reliability mathematics Fundamentals; Practices; Procedures”, McGraw-Hill, Inc., 07-001598-8, 1971.
43. W. Ireson, C. Coombs Jr., R. Moss, Handbook of reliability engineering and management, 2nd edition, McGraw-Hill, (1996).
44. C. Visser, 8D Problem Solving Explained: Turning Operational Failures Into Knowledge to Drive Your Strategic and Competitive Advantages, CreateSpace Independent Publishing Platform, 2017.
45. D. Benbow, H. Broome, The certified reliability engineer handbook, ASQ quality press, 2008.
46. P. J. Boland, F. Proschan, “The Reliability of K Out of N Systems”, The Annals of Probability, Vol. 11, No. 3 (Aug., 1983), pp. 760–764.
47. https://weibull.com/hotwire/issue2/hottopics2.htm.

Failure Analysis in Advanced Driver Assistance Systems

Yan Li and Hualiang Shi

Abstract Failure analysis (FA) could provide timely feedback to process optimization and solution paths for system failures; thus, it is critical for the development of advanced driver assistance systems (ADAS). In this chapter, failure analysis flows starting from systems or boards until components or packages and dice are introduced. Electrical fault isolation (FI) techniques designed to locate subtle defects inside complicated semiconductor devices are reviewed. Physical failure analysis approaches adopted to provide artifact free nanometer scale analysis are discussed. Material analysis methods assisting in thorough root cause investigation are presented. Non-destructive and high-resolution imaging tools with the potential of significantly shortening failure analysis through put time are demonstrated. Case studies are used to illustrate strategies and methodologies in ADAS failure analysis.

1 Introduction

Advanced driver assistance systems (ADAS), consisting of electronic systems that aid drivers in driving and parking functions, have been a rapidly growing area in semiconductor industry [1]. It involves various types of sensors and high-performance computers having very harsh use conditions, such as elevated temperature and humidity, big temperature variation, and corrosive environment. Electrical tests are performed at the end of production line (EOL) as well as post-reliability tests to ensure the quality and reliability of ADAS. Failure analysis of the failed devices in electrical tests at either EOL or post-reliability tests is critical for the development of ADAS, as it could reveal the root causes of failures, determine failure mechanisms,

Y. Li (✉) · H. Shi
Intel Corporation, Chandler, Arizona, USA
e-mail: yan.a.li@intel.com

H. Shi
e-mail: hualiang.shi@gmail.com

Y. Li
Level 5 Self-driving Division, Lyft, Palo Alto, USA

Y. Li and H. Shi (eds.), *Advanced Driver Assistance Systems and Autonomous Vehicles*,
https://doi.org/10.1007/978-981-19-5053-7_14

define factors causing failures, provide timely feedback to process optimization, and suggest solution paths for system failures.

Since there are many kinds of sensors, components, and computers in ADAS, detailed failure analysis flows at system, board, component, package, and die levels are necessary to help pinpoint defects causing the failures. Electrical failure analysis is indispensable for successful failure analysis. Fault isolation at system level can help locate the failures to board level. Board level fault isolation is used to detect if the failures could be in the packages or components on the boards, the solder joints connecting the packages or components to the boards, or the trace routings inside the board layers. Component or package level fault isolation is employed to see if the failures could be in the multiple die stacks in the packages or components, the interconnects connecting the die stacks to the organic substrates inside the packages or components, or the trace routing in the package or component layers. Die level fault isolation is utilized to find the failures inside Si layers, which could be either in the transistors, the back end, or the far back-end layers inside Si.

Electrical failure analysis (EFA) techniques, such as curve trace, time-domain reflectometry (TDR), lock-in thermography (LIT), magnetic field imaging (MFI), infrared emission microscopy (IREM), thermally induced voltage alteration (TIVA), nanoprobing, e-beam imaging, and e-beam probing are reviewed.

Physical failure analysis (PFA) is decisive to precisely reveal the defects after the electrical fault isolation. Depending on the locations of the defects, various sample preparation techniques including conventional mechanical polishing, laser ablation, ion milling, reactive ion etching, focused ion beam (FIB), and plasma-FIB can be used to perform cross section or de-layering inside boards, packages, or dice. After the defects are exposed by PFA, defect imaging techniques, such as scanning electron microscopy (SEM) and transmission electron microscopy (TEM) are employed to inspect the defects at high magnification.

Material analysis is important for in-depth study of the disclosed defects. Energy-dispersive X-ray spectroscopy (EDX) is typically used in SEM and TEM to reveal the chemical elements in the defects and surrounding layers. To study the chemical states of organic defects, Fourier transform infrared spectroscopy (FTIR) and atomic force microscopy-based infrared spectroscopy (AFM-IR) can be employed. If failures are related to surface or interface contaminations, X-ray photoelectron spectroscopy (XPS) and time-of-flight secondary ion mass spectrometry (TOF–SIMS) can be very helpful. Backscatter diffraction (EBSD) provides crystallographic orientation of interconnects and is sometimes beneficial for fundamental understanding of failures inside solder joints, Cu vias, and through silicon vias (TSV) in ADAS.

Failure analysis in complex ADAS is typically time consuming. Non-destructive and high-resolution imaging techniques, such as optical microscope and infrared (IR) imaging, scanning acoustic microscopy (SAM), 2D X-ray radiography, and 3D X-ray computed tomography (CT) are found to be very effective to detect failures in electronic systems. Depending on the field of view of the technique, defects could be revealed non-destructively without any fault isolation. Integrating the non-destructive and high-resolution imaging techniques into the failure analysis flows results in skipping some complicated steps in electrical failure analysis and physical failure analysis; thus, it significantly improves the through put time (TPT) of ADAS failure analysis.

This chapter provides an overview of the failure analysis flows, introduces technical details of the tools and techniques typically used in electrical failure analysis and physical failure analysis, highlights a few non-destructive and high-resolution imaging approaches for TPT reduction. Case studies in ADAS industry are adopted to demonstrate the applications of the failure analysis flows and techniques. Future development trends in failure analysis of ADAS are also discussed to overcome current challenges and gaps.

2 Failure Analysis Flow

Failure analysis flow is a diagnostic flow chart carefully designed to detect the defects causing electrical failures in complex electronic systems. As illustrated in Fig. 1, it typically involves with problem statement review, failure verification, non-destructive imaging investigation, electrical failure analysis, physical failure analysis, and material analysis [2].

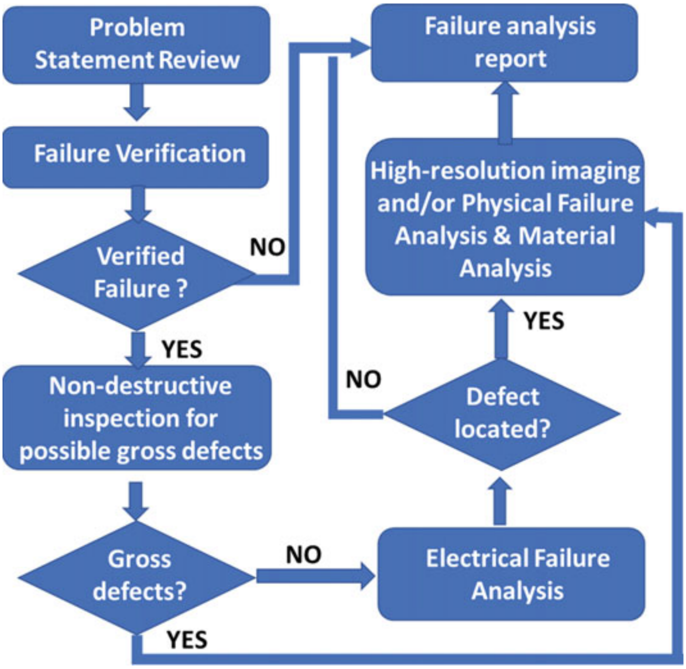

Problem Statement Review
Failure Verification
Verified Failure ?
NO
YES
Non-destructive inspection for possible gross defects
Gross defects?
NO
YES
Electrical Failure Analysis
Defect located?
NO
YES
High-resolution imaging and/or Physical Failure Analysis & Material Analysis
Failure analysis report

Fig. 1 Schematic of failure analysis flow (Adapted from ref. [2])

Although often overlooked by rushed analysts, problem statement review is the most critical step for successful failure analysis [2]. A good problem statement may involve the following factors. (1) Failure occurrence condition and history. For example, does it happen at EOL, reliability tests, or customer return? Failures happened at EOL typically indicates process defects. Process details of the failing units and baseline defect library are very helpful for defect identification. Reliability failures are typically either initiated or propagated at the rigorous test conditions. Various reliability tests can lead to different types of defects; thus, it needs distinct failure analysis techniques to reveal them efficiently [3]. For customer returned units, communications about the unit history at the customer site is crucial. Detailed information about the thermal and test history of the failed system is beneficial to successful failure analysis. (2) Failure rate analysis. A gross failure rate at EOL typically indicates a major process issue and gross defects. While a small failure rate at EOL suggests that it could relate to tiny and random defects. For reliability failures, the "bathtub curve" of failure rate vs reliability test time as shown in Fig. 2 illustrates three types of failures in reliability tests [3]. Infant mortality failures are defect-driven failure modes, having very high failure rates at early read outs of reliability tests, suggesting process defects. Intrinsic failures have a constant failure rate during reliability tests, which are typically inherent to the system design and stress. Wear-out failures start to show up with high failure rates at extended test time indicating the end of system life. Failure hypothesis and failure analysis flow could be different for different types of failures as shown in Fig. 2. For customer returns, failure rate analysis at customer site can reveal information about possible process or test issues. (3) Historical failure analysis data leveraging. Historical failure analysis data from a similar electrical system is very valuable for the design of failure analysis flow. Similar failure analysis flow can be followed, and sometimes cases could be signature closed based on the historical data of failure locations or failure signatures, without complicated electrical and physical failure analysis [3].

Because of common tester contact issues, test program immaturity, and failure intermittency, failure verification is required before any other time-consuming failure analysis steps [2]. For parametric failures in electronic systems, including both open

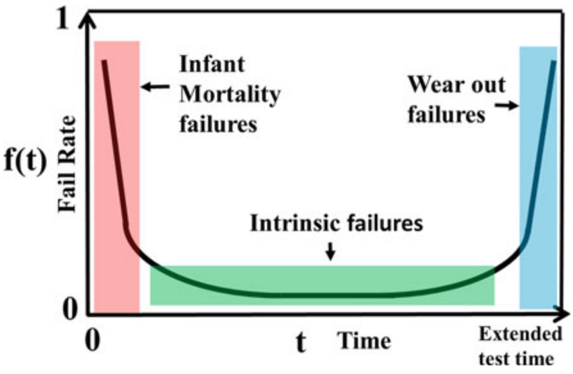


Fig. 2 Schematic of bathtub curve showing failure rate versus reliability test time (Adapted from ref. [3])

and short failures, I–V curve, resistance, or currents of the failed device are usually obtained and compared with those from a test good device. To verify functional failures, debug testers and test programs are utilized, and the test results are compared between the failed and passing units. Results from verified failures typically show distinct deviation from those of good units.

After failures are verified, non-destructive imaging techniques with large field of view are employed to inspect the whole "area of interest (AOI)," which could cause the electrical failures. For example, the whole system can be inspected by optical microscopy to find any possible handling damages. SAM can be used to scan the whole package and all dice on a device for gross delamination. 2D X-ray can be used to check solder joint integrity in the entire bump area. IR imaging can be used to exam all the dice for internal die cracks. If a gross failure is identified in the AOI by non-destructive imaging techniques, the unit can directly go to further physical failure analysis and material analysis for root cause understanding, skipping electrical fault isolation, as illustrated in Fig. 1. However, if gross failures are not found, detailed electrical failure analysis is indispensable.

Electrical failure analysis can perform fault isolation and locate a tiny defect in a very complicated electrical system. TDR, LIT, and MFI are common tools for FA of parametric failures in electronic systems. Optical fault isolation tools, e-beam imaging, and probing techniques are typically used in FA of device functional failures. Probing at micron or nanoscale is also very effective to locate the failures. A combination of various types of fault isolation tools is normally applied to pinpoint the failures in ADAS.

Physical failure analysis and material analysis are crucial to reveal defect details and suggest failure root causes. The selected physical failure analysis techniques need to have a relatively short through put time (TPT) and reveal defects without introducing artifacts to the AOI. For gross defects in organic PCB board, substrates, or big solder joints, mechanical cross section or planar grinding with relatively lower cost and shorter TPT is preferred. For subtle defects in organic substrates, micron scale interconnects, and die layers, combined physical failure analysis techniques, including mechanical, laser ablation, ion milling, and FIB are typically adopted to disclose defects with artifact free results as well as short TPT.

Failure analysis report is the last step of the flow chart and includes problem statements, steps performed in the analysis, results, and conclusions. Comprehensive failure analysis reports are important for further communications with partners to investigate root causes and serve as historical data for reference in the future.

2.1 Failure Analysis of Systems or Boards

Failure analysis of systems or boards follows the same flow chart as shown in Fig. 1. After the careful review of the problem statement, failure verification at system or board level is needed. System level tests can verify the failures and help isolate the

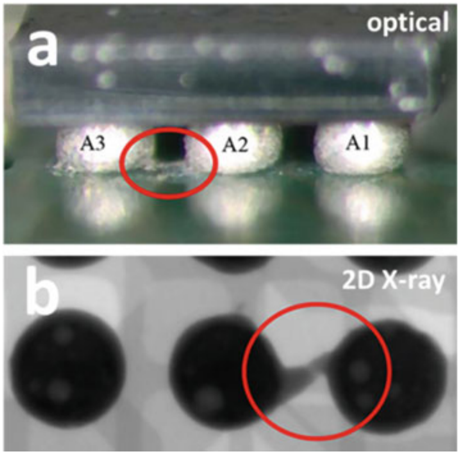


Fig. 3 Optical (**a**) and 2D X-ray image (**b**) of solder residue (highlighted by red circles) between two solder joints causing short failures (Adapted from ref. [1])

failures to different modules or boards. Board level tests can verify failures at board level and further isolate them to components or packages.

After the failures are verified at board level, a through non-destructive inspection, including optical microscopy, 2D X-ray, SAM, and IR imaging of the entire failed unit is very effective to capture any gross defects, like solder joint open and shorts, board damage, board layer delamination, and die internal cracks. [2]. Figure 3 illustrates the solder residue between two solder joints causing short failures, exposed by optical and 2D X-ray inspection.

If gross defects could not be revealed by non-destructive inspection of the whole board area, electrical failure analysis at board level can help identify a much smaller "AOI" having the defects. LIT or MFI, not requiring any sample preparation, is typically used at board level to perform the fault isolation. Once the failure location is identified, non-destructive imaging techniques with much higher resolution are employed to reveal the defects within the smaller AOI. Figure 4a, b shows LIT results performed on the failing net of a board. The hot spot highlighted by an arrow as shown in Fig. 4b indicates the short failure location on the board, which is a capacitor. High-resolution 2D X-ray imaging is performed on the failing capacitor and reveals an internal crack, which leads to the short failure [2].

Detailed physical and material analysis of defects can help with root cause investigation. Mechanical cross section or planar polishing is typically used in board level failure analysis, as defects at board level are relatively bigger compared with those at package or die level. Ion milling cross-section techniques can be applied if the analysis could not tolerant mechanical polishing induced artifacts. [4] As illustrated in Fig. 5a, the artifact free cross section of a solder joint with ~300um in diameter is performed by ion milling cross section. Figure 5b illustrates the high-quality electron backscatter diffraction (EBSD) mapping of the solder joint showing grain structure, size, and crystallographic orientation of the Sn bump and Cu pads. Figure 5c displays the strain contour mapping of the solder join from the high-resolution EBSD analysis.

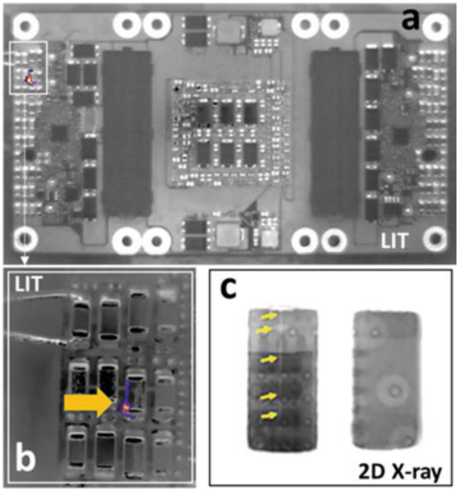


Fig. 4 **a** and **b** LIT performed at board level showing a hot spot on a single capacitor indicating the location of the short failure. **c** High-resolution 2D X-ray image of the failing capacitor shows a crack led to the short failure. (Adapted from ref. [2])

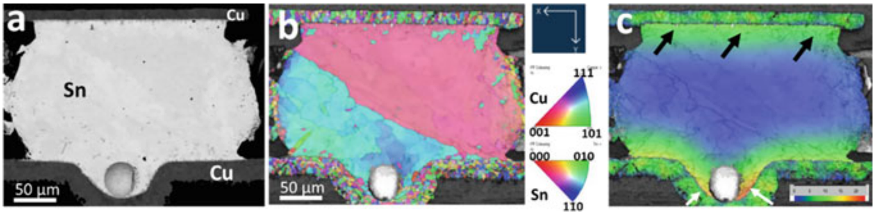


Fig. 5 **a** SEM image of a solder joint after ion milling cross section. **b** Electron backscatter diffraction (EBSD) mapping of the solder joint showing grain structure, size, and crystallographic orientation of the Sn bump and Cu pads. **c** Strain contouring EBSD map of the solder joint; Strain accumulates at the interface of the solder bump and copper pad (black arrows) and at the area between the solder bump void and copper pad (white arrows) (Adapted from ref. [4])

The mechanical cross-sectioned solder joints are not qualified for the same EBSD analysis presented in Fig. 5, due to the shadowing effect resulting from different material removal rates during mechanical polishing [4].

2.2 *Failure Analysis of Packages*

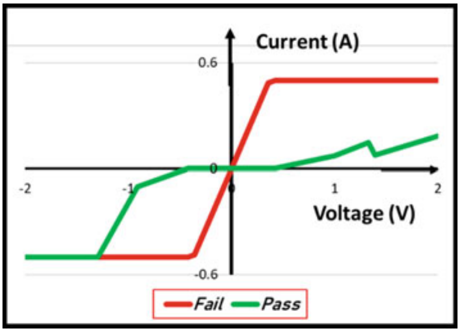
Failed components or packages verified at board level are typically transferred to the package failure analysis team. Package failure analysis investigates failures in Cu vias, interconnects, traces, and components in dice or organic substrates, as well as solder joints or wire bonds connecting packages to boards. [3] The package failure analysis flow is the same as shown in Fig. 1.

After conscientious problem statement review, failure verification is performed at package level. Depending on the failure modes, failure verification of packages used in ADAS sometimes requires the use of evaluation boards (EVB) to replicate the module level functionality as well as automatic test equipment (ATE) tests required for the analysis. Development of an EVB socketing is sometimes needed for package failure verification [5]. In most of the cases, curve tracing, which shows the current–voltage (I-V) characteristics of a microelectronic device, is usually employed to verify, and isolate open or short failures in packages. Figure 6 illustrates the I-V curve trace collected from a package having short failures and a test good (or passing) package, with the current value along the y-axis and the voltage level along the x-axis [6]. The green I-V curve from the passing unit is nonlinear and displays the unique operating characteristics of the semiconductor device having Si diodes and transistors. While the red I-V curve from the failing unit is linear before reaching the maximum current setting limits, indicating the electrical characteristics of a resistor. The short failure is not likely at die level, as the Si diode is not turned on. They could locate at either the interconnects or the organic substrate.

Post-failure verification, non-destructive inspection techniques are applied to detect any possible gross defects. Figure 7a displays the C-mode SAM (CSAM) image of a unit with open failures. It captures a big underfill void in the failing area. 2D X-ray imaging with a titled view is performed in the UF voiding area, which discloses the solder extrusion between two interconnects, as illustrated in Fig. 7b [3].

If gross defects are detected during the non-destructive inspection, high-resolution imaging or physical failure analysis on the defects is needed. As illustrated in Fig. 7c, high-resolution 3D X-ray CT is performed around the solder extrusion area and reveals an open bump. Failure analysis report is written up and suggests the root cause of the open failure demonstrated in Fig. 7 is due to underfill voiding. During reflow reliability stress, molten solder from an interconnect located inside the UF void extrudes into the open space in the underfill and causes an open failure [3].

Fig. 6 I-V curve trace from a passing (in green) and a failing unit (in red) (Adapted from ref. [6])



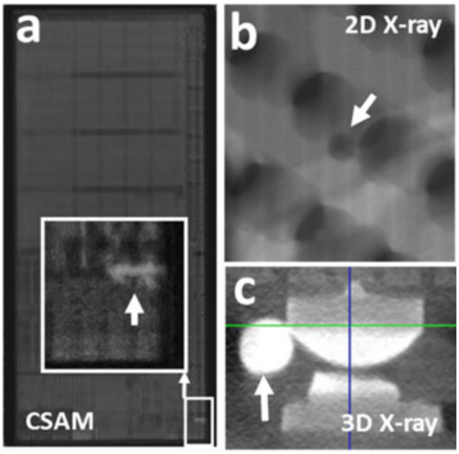


Fig. 7 **a** CSAM images showing a UF void in the failing area. **b** 2D X-ray image with titled view discloses the solder extrusion in the failing area having the UF void. **c** 3D X-ray CT around the solder extrusion revealed an open bump (Adapted from ref. [3])

If gross defects are not detected during the non-destructive inspection, a detailed electrical failure analysis is needed. For short failures in packages, LIT is typically adopted to help define the short failure location. For open failures, TDR is often used to isolate the failures [3]. As demonstrated in Fig. 8, TDR spectra (red solid line) are collected from the failing pin of the package with open failures and compared with that (green dash dot line) from a test good or golden unit collected from the same package pin. The failed package is with a "package on package" architecture. As illustrated by the schematic picture in Fig. 8, the die connects an organic interposer through first level interconnects (FLI). The solder joints between the interposer and organic substrates are called middle level interconnects (MLI). The whole package connects with the board through second level interconnects (SLI). To locate open failures with TDR technique, reference spectra taken from the same pin of a bare substrate (orange dash line) and a package without die (blue dot line) are taken and define the MLI, and FLI locations in TDR spectra, respectively. By comparing the TDR spectra from the failed unit with all the reference TDR spectra, the open failure location is estimated to be very close to the FLI connecting to the failing pin [3].

After the detailed electrical failure analysis, a relatively small AOI is defined, which enables high-resolution imaging or physical failure analysis in the defective area. Figure 9 demonstrates the high-resolution 3D X-ray CT images around the FLI connecting to the failing pin in Fig. 8. It clearly shows a trace crack close to the FLI. After analyzing the failure location commonality and unit stress history. The failure analysis report is created and suggests the root cause of the open failure is due to package stress induced Cu trace crack post-extended temperature cycling reliability stress [3].

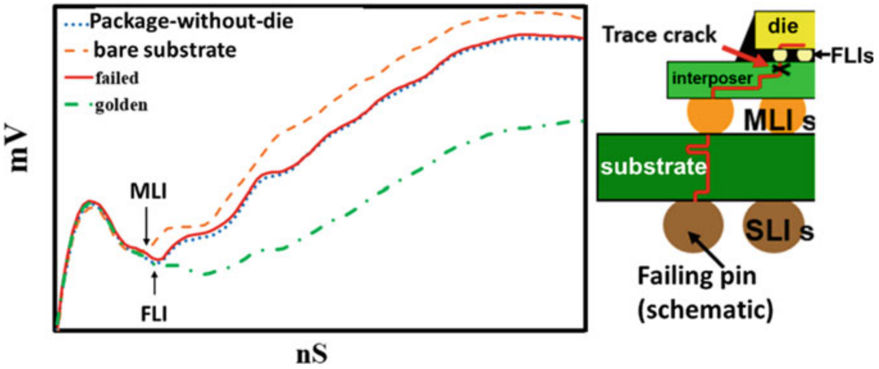


Fig. 8 TDR spectra from a failed unit, a golden unit, a bare SIP substrate, and a SIP package without FPGA die, showing the open failure is close to the FLI solder joint in the FPGA package of a SIP package (adapted from ref. [3])

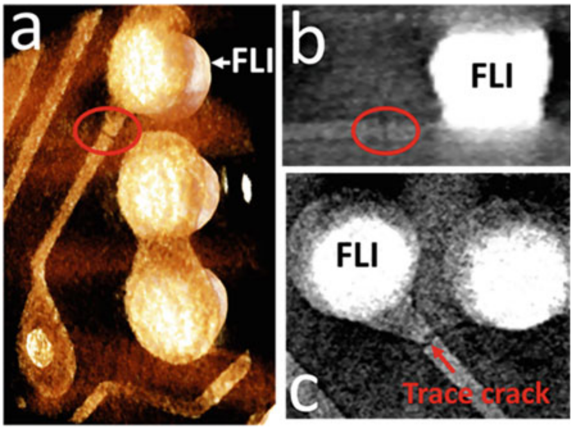


Fig. 9 **(a)** 3D view, **(b)** virtual cross-sectional view, and **(c)** virtual planar view of the 3D X-ray CT showing the trace crack in a package (Adapted from ref. [3])

In some cases, physical failure analysis and material analysis of defects are essential for root cause understanding. Depending on defect types and scales, different sample preparation methodologies can be applied to reveal the defects with minimum artifacts. For big defects at interconnects and organic substrates, mechanical cross section and planar polishing are used to reduce cost and TPT. For subtle defects in interconnects and substrates, or any defects in Si layers, FIB, plasma-FIB, ion etching, and ion milling cross-sectional techniques are needed to precisely reveal the defects without artifacts. As illustrated in Fig. 10a, FIB cross section is used to expose die level interlayer dielectric (ILD) delamination. To expose subtle defects in large features, plasma-FIB or ion milling cross section is employed to speed up the milling process. Figure 10b and (c) display the cross-sectional SEM images of through silicon vias (TSV), which are the interconnects between stacked dice of three-dimensional (3D) packages, performed by plasma-FIB and ion milling, respectively

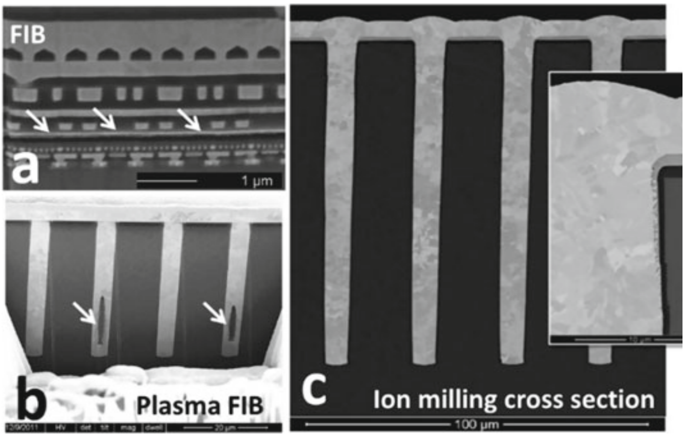


Fig. 10 **a** SEM image of Si layer ILD delamination by FIB cross section. **b** and **c** SEM images of TSVs by plasma-FIB and ion milling cross section, respectively (Adapted from ref. [3, 7, 8])

[3, 7, 8]. Material analysis techniques, such as SEM–EDX, TEM–EDX, FTIR, AFM-IR, XPS, TOF–SIMs, and EBSD are frequently used in package failure analysis for root cause investigation.

2.3 Failure Analysis of Si Device

Units with die level failures are usually transferred to device failure analysis or product failure analysis team. Device failure analysis investigates failures in transistors, back-end, and far back-end circuitry in Si. It also follows the failure analysis flow shown in Fig. 1. Defects in Si transistors and die layer routes are typically in nanometer or sub-micron scale, requiring dedicated efforts in electrical failure analysis (EFA) and physical failure analysis (PFA).

Failure verification needs to be performed to confirm die level failures post the thorough problem statement review. The failure verification typically involves a debug tester or ATE [9]. As illustrated in Fig. 11a, if defects are between externally accessible signals, static or power-off I-V curve tracing could be used to verify and isolate the failures. Figure 6 shows an example of static I-V curves demonstrating short failures in packages. For failures between internal signals shown in Fig. 11b, which are not accessible externally, debug testers or ATE is needed for failure verification. Figure 12 displays powered I-V curves collected on a debug tester for the failed unit (in red) and the passing unit (in blue). I-V curve from the failed unit clearly shows higher current comparing with that from the golden reference, indicating leakage failures [9]. Static I-V curves collected from the same set of units, without the debug tester, do not show any difference between the failing and the passing, not able to verify the failure [9].

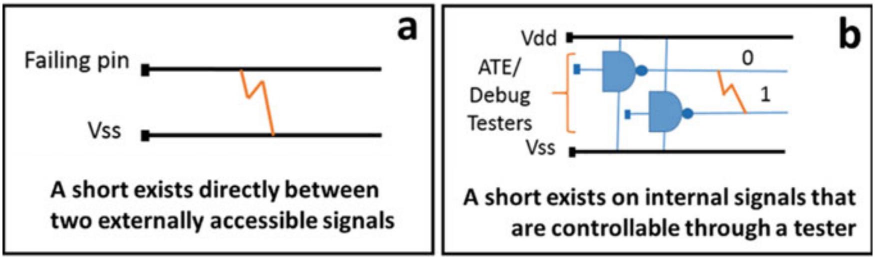

a
Failing pin
Vss
A short exists directly between two externally accessible signals
b
Vdd
ATE/
Debug
Testers
Vss
0
1
A short exists on internal signals that are controllable through a tester

Fig. 11 **a** Schematic of shorts between two externally accessible signals. **b** Schematic of shorts between internal signals that are controllable through testers (Adapted from ref. [9])

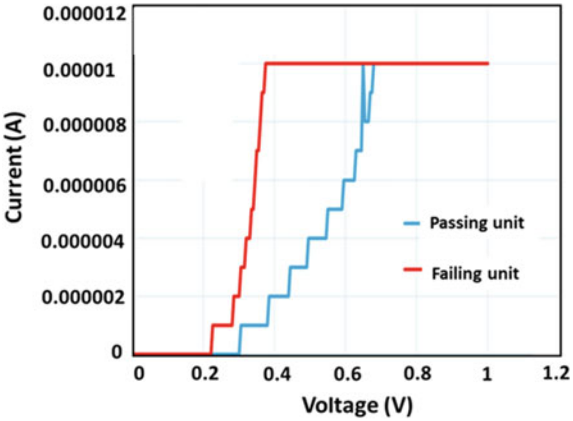


Fig. 12 Powered I-V curves collected on a debug tester shows leakage in the failing unit (Adapted from ref. [9])

Post-failure verification, non-destructive inspection techniques, like CSAM, optical microscopy, and IR imaging in the entire die area are helpful to capture any possible gross defects. Figure 13a displays the CSAM image of a failing die with three “white spots” in the AOI, indicating Si circuitry defects. Figure 13b shows the FIB cross section on one of the “white spots,” indicating Si layer delamination. The delamination could propagate to a Cu via nearby and cause the open failure [3].

If gross defects are not disclosed by the non-destructive inspection, exhaustive EFA is required to locate the subtle Si defects. Optical-based techniques, like TIVA, IREM or PEM, are typically used for the fault isolation (FI). Sample preparation for the optical-based FI usually requires die thinning to less than 40um thickness of bulk Si. If defects are between logic circuitries, as illustrated in Fig. 11b, the optical FI needs to be performed on the samples at the “power-on” state [9]. Figure 14 presents the TIVA data collected by powering on the device under test (DUT) and enforcing the leakage pattern on the failed pins, which clearly identifies two failing locations.

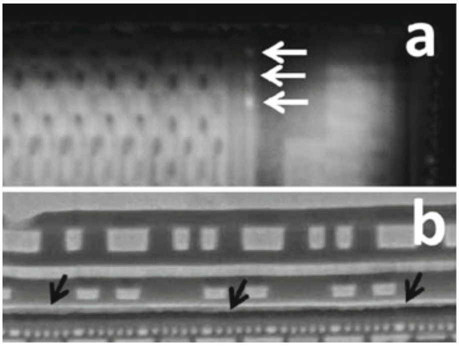


Fig. 13 **a** CSAM image of a failing die showing three "white spots" in the AOI. **b** FIB cross section on one of the defects in **(a)** showing subtle Si layer delamination (Adapted from ref. [3])

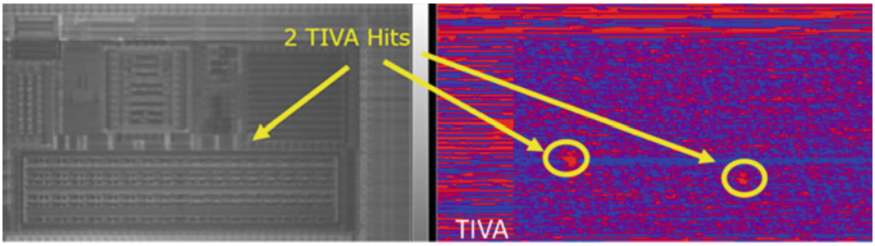
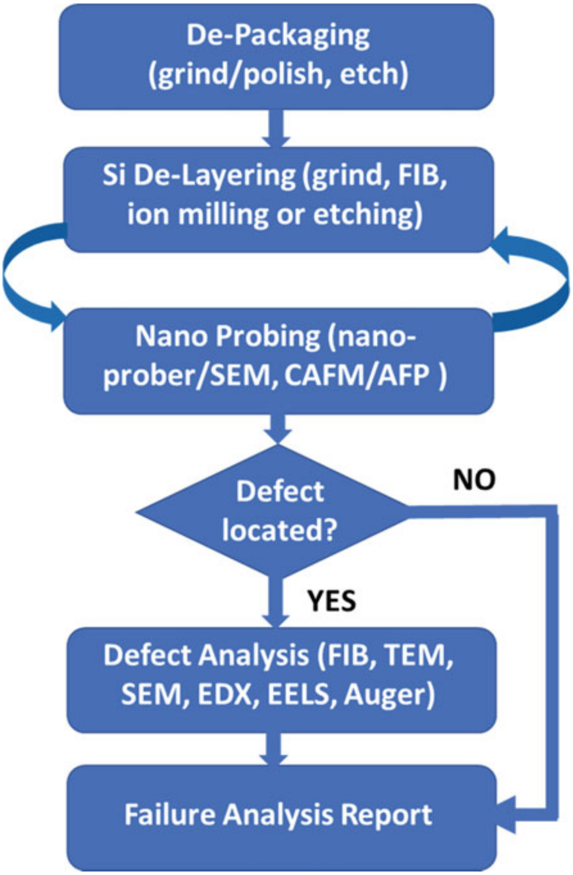


Fig. 14 TIVA data collected by powering on the sample and executing the leakage pattern to the affected IO pins (Adapted from ref. [9])

Static TIVA data on the same DUT without powering on does not show any defects [9].

PFA of Si failures to reveal nanometer scale defects is very time consuming; it is recommended that after the failure is isolated to a small area, high-resolution, and non-destructive imaging techniques, such as nanoscale 3D X-ray CT, high-resolution CSAM or IR imaging to be used to reveal any sub-micron scale Si circuitry abnormity or delamination [2]. If defects cannot be detected by the non-destructive techniques, detailed PFA is needed to expose the defects. As illustrated schematically in Fig. 15, the PFA process starts from de-packaging of the device [10]. Mechanical grinding and polishing, reactive ion dry etching or chemical wet etching techniques are utilized to separate the failed die from the package. Si de-layering techniques, like mechanical grinding, FIB, ion milling or etching, are employed to take off Si circuitry layer by layer. Nanoprobing techniques, like SEM-based nanoprobing, conducting atomic force microscopy (CAFM), or atomic force probing (AFP) are used to locate the defects at each layer until the defects are revealed. Thorough defect analysis is performed after identifying the defect. FIB-SEM is used for planar or cross-sectional view of the defects. If higher resolution is needed, TEM sample preparation around the defect area is applied utilizing the FIB tool. A thin slide of the sample having the defect is sent for TEM analysis. E-beam-based material analysis techniques, such as EDX, Auger electron spectroscopy (AES or Auger), electron energy loss spectroscopy (EELS), are frequently used to reveal the chemical information of the defects. As illustrated in Fig. 16a, I-V curve obtained by nanoprobing at transistor level identifies the short in one of the transistors located in the AOI. FIB-SEM is used for lifting out a thin slide of the failing area for TEM analysis. The planar TEM image shown in Fig. 16b clearly shows the abnormity across fin of the failing fin field effect transistor (FinFET). Figure 16c shows the TEM–EDX mapping of Ti element in the failing area showing extra titanium extending across fin, which leads to the leakage failure of the transistor. Further root cause investigation confirms that the leakage is due to an electrostatic discharge (ESD) event on the failing unit [9]. A through failure analysis report based on all the data is written and published to the technology development team for failure mode and mechanism documentation and further process improvement.

Fig. 15 Schematic of detailed physical failure analysis flow for die level device failures (adapted from ref. [10])


De-Packaging (grind/polish, etch)
Si De-Layering (grind, FIB, ion milling or etching)
Nano Probing (nano-prober/SEM, CAFM/AFP)
Defect located?
NO
YES
Defect Analysis (FIB, TEM, SEM, EDX, EELS, Auger)
Failure Analysis Report

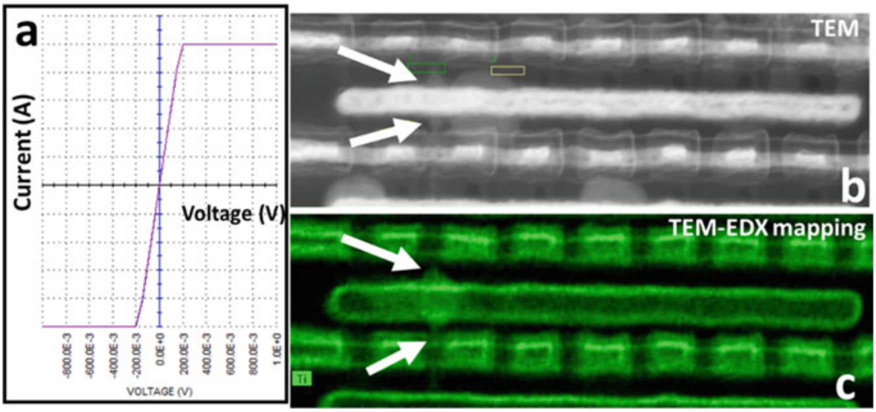


Fig. 16 **a** I-V curve of transistor level nanoprobing showing short in the failing transistor. **b** TEM image shows damages across fin of the transistor. **c** TEM–EDX mapping of Ti element showing extra titanium extending across fin (Adapted from ref. [9])

3 Electrical Failure Analysis (EFA) and Fault Isolation Techniques

3.1 Non-destructive Fault Isolation Tools

Efficient EFA is the key to failure analysis success. Board level and package level failure analysis heavily rely on non-destructive fault isolation techniques which can locate defects by accessing DUT failing pins without any sample preparation.

3.1.1 I-V Curve Tracing

I-V curve tracing displays the I-V characteristics of semiconductor devices and is used non-destructively at board or package level for failure verification. Failures are confirmed and isolated by comparing the I-V curve collected from the failing pins of DUT with that from the same pins of a good unit. As illustrated in Fig. 11a, static curve tracing can be performed on units with failures exist in externally accessible signals. If failures are at internal signals, curve tracing needs to be exercised while DUT is in the "power-on" state. Figure 6 shows static I-V curves from a passing unit and a failing unit. The I-V curve of the failed unit confirms the short failure and indicates that the defect is either in the package or the Si circuitry before reaching any Si diodes or transistors, because the red I-V curve is linear, demonstrating the electrical characteristics of a resistor [6]. Figure presents powered I-V curves collected on a debug tester from a failed unit and a passing unit. It clearly reveals the higher current in the failing unit indicating the leakage failure. Static I-V curves from the same set of units cannot verify the leakage failure. The defects verified in Fig. 12 are due to transistor level shorts, as presented in Fig. 16, and cannot be exposed by performing static I-V curve tracing [9].

3.1.2 Time-Domain Reflectometry (TDR) and Electro-Optic Terahertz Pulse Reflectometry (EOTPR)

To locate board and package level open and high resistance failures, non-destructive FI techniques, such as TDR and EOTPR, are typically employed [11]. Comparing with EOTPR, TDR is a conventional tool with a lower cost.

With 35–40 ps rise time, a step electrical pulse is injected into semiconductor devices. The collected reflected signal from the device is analyzed to obtain the impedance variation along the circuit. The failures can be located by comparing the reflected signals between the failed unit and reference samples. The spatial resolution of TDR could be estimated as 1/10–1/5 of the TDR rise time [12] and could be calculated by the following equation [13]:

$$\lambda_{\text{resolution}} = \frac{0.35}{\text{BW}_{\text{system}}} \cdot \frac{c}{2\sqrt{\varepsilon_{\text{eff}}}} \tag{1}$$

where c is velocity of light in free space, ε_{eff} is the effective dielectric constant of the device, and $\text{BW}_{\text{system}}$ is the "overall system bandwidth." The spatial resolution of a TDR system with 18–20 GHz $\text{BW}_{\text{system}}$ and 35–40 ps rise time in flip chip packages is about 500 μm [11]. As illustrated in Fig. 8, TDR data collected non-destructively at package level isolates an open failure in a semiconductor device. The locations of MLI and FLI in the time domain are defined by the TDR spectra from reference units. The failure is located to be close to the FLIs connecting the die to the interposer, by comparing the TDR spectrum of the failing unit with the reference spectra.

Developed as the next generation of TDR, EOTPR has a higher resolution. A 40 GHz to 4 THz electrical pulse with a sharp peak is produced and is injected into the device under test [11]. The resolution of EOTPR could be about 10 μm, much higher than that of TDR [14], mainly due to the much faster rise time of 5.7 ps and the THz range system bandwidth. As the input signal is a pulse, in EOTPR spectra, open failures in device show as peaks, while short failures as valleys. The same set of data in Fig. 8 is collected by EOTPR and displays in Fig. 17. The failure is isolated to be right before the FLI interconnect, by comparing with the blue spectrum, which is used to define the FLI location in the device.

By transforming the time-domain EOTPR raw data into the distance domain data through the phase velocity of the electromagnetic wave, the exact failure location can be calculated. The phase velocity of the electromagnetic wave can be expressed as [15]:

$$v_p = \frac{c}{\sqrt{\varepsilon_{\text{eff}}}} \tag{2}$$

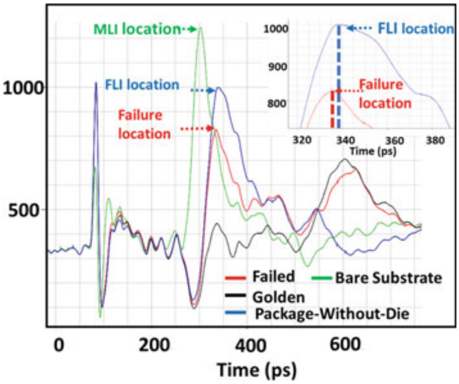


Fig. 17 EOTPR spectra from a failed unit, a golden unit, a bare substrate, and a package without die (adapted from ref. [3])

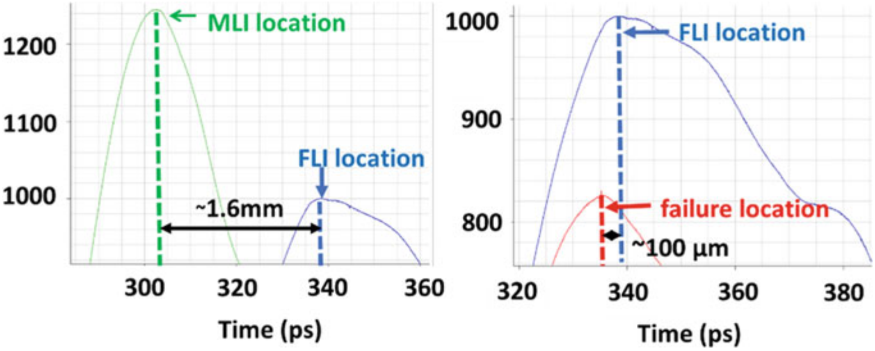


Fig. 18 Failure location can be estimated using known package structure locations in EOTPR spectra (Adapted from ref. [3])

where c is velocity of light in free space, v_p is the phase velocity, and ε_{eff} is the effective dielectric constant of the package, which is distinct for different packages. Figure 18 illustrates that the ε_{eff} can be estimated by the MLI and FLI locations in the EOTPR time-domain spectra and the actual distance between them. In the package design file, the actual distance is measured as 1.6 mm. Based on the estimated ε_{eff}, the distance between the FLI and the failure location can be calculated as about 100 μm [3].

3.1.3 Lock-In Thermography (LIT)

LIT is a very promising technique to obtain z-dimensional information of defects as well as the x and y location data, which is the key to precisely locate shorts and resistive failures in semiconductor devices. It is developed to locate the dissipated heat by resistive failures using the real-time graphical lock-in methodology, which is designed to detect small signal buried in random noise of higher magnitude [16]. LIT can detect the *z*-dimensional locations of defects because the heat propagation is time depended. The phase shift between the excitation signal and thermal response is determined by the time delay of the underlying thermal diffusion process. It is related to the frequency of the stimulation signal, which is the lock-in frequency, as well as thermal properties and thicknesses of different layers in the device. By comparing to reference units with defects at each relevant *z* position, the *z* position of hot spots within the device can be identified by their lock-in frequency vs phase shift curves. Figure 19 [16, 17] displays the experimental and simulation LIT data. It illustrates that hot spots in the stacked dice of a package are differentiated by LIT. The result demonstrates that the Z-dimensional information of the defect can be effectively revealed by LIT, along with the *x* and *y* locations. Figure 4 presents the application of LIT to board level leakage failures, showing a hot spot on the failing capacitor. LIT can efficiently provide three-dimensional leakage failure locations but is not valid for open failure isolation, as well as short failures having resistance less than one ohm [3].

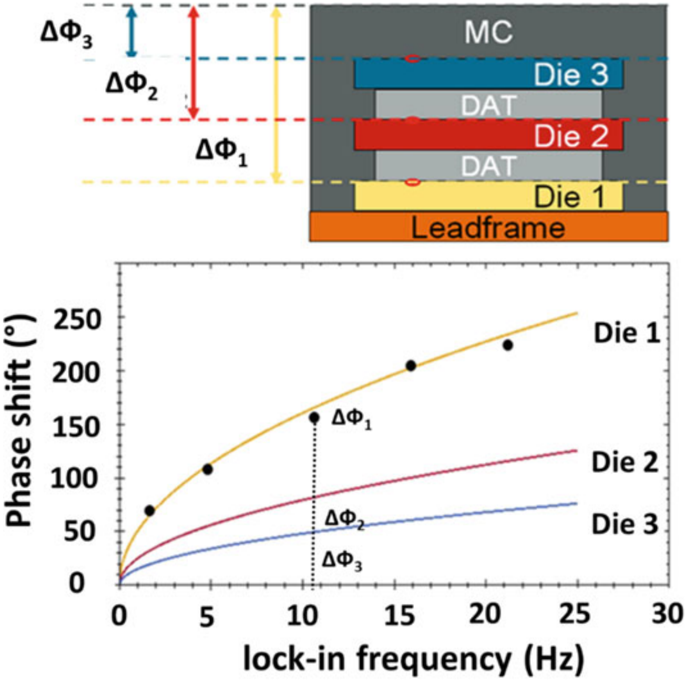


Fig. 19 Measurement results of lock-in frequency vs phase shift curves from reference units with short failures at die 1, die 2, and die 3 of a package with stacked die configuration showing in dots compared to a calibrated simulation model (lines) (Adapted from ref. [16, 17])

3.1.4 Magnetic Field Imaging (MFI) or Magnetic Current Imaging (MCI)

MFI or MCI refers to fault isolation (FI) techniques using magnetic sensors to map the magnetic field produced by a current injected into the failing structure of the DUT [18–20]. Failure locations are identified by converting the magnetic field image into current map and compare it with that from a good unit or the device design circuitry. Multiple types of magnetic sensors are promising for semiconductor device fault isolation, including superconducting quantum interference device (SQUID) [2], giant-magneto resistive effect (GMR) [2], tunnel-magneto resistive effect (TMR) [21], and quantum diamond microscope (QDM) based on nitrogen vacancy magnetometry [22, 23].

SQUID microscopy (SSM), with magnetic sensor operating at kHz frequency range, has been integrated into short and high resistance failure FA flow for more than twenty years [3, 18]. Additionally, SSM can isolate hard open failures by increasing the bandwidth of SQUID electronics to RF range. It is also called space domain reflectometry (SDR) for the high frequency application of SSM [19]. SSM detects the magnetic field generated by the input current inside the failed unit. A Fourier transform inversion technique is then used to process the magnetic field signal to obtain current density map of the sample. By comparing a circuit diagram or the current density map from a golden unit with that from the failed unit, the x and y locations of the failure can be identified. It is known that the distance between the minimum and the maximum magnetic field around a straight current path in the unit is twice the total distance from the SQUID sensor to the current path [20]. The separation from the sample surface to the current path, which is the z-dimensional information of the defect, could be estimated, as the distance from the sensor to the sample surface is fixed for each SSM scan.

Figure 20 demonstrates the detailed process of defining the *x*, *y*, and *z* locations of a short failure in a device with stacked die configuration by SSM. The x and y location of the short can be determined by the current analysis, as detailed in Fig. 20a. The distance between the minimum and the maximum magnetic field around the current path is measured to be 1906 µm as illustrated in Fig. 20b. The sensor to current path distance, which is half of 1906 µm, is calculated to be 953 µm. The sensor to sample distance, which is fixed in this SSM scan, is 393 µm. The *z* location of the defect, which is the distance from the current to the Si surface, is estimated to be 560 µm, by subtracting the sensor to sample distance (393 µm) from the sensor to current path distance (953 µm). As shown in Fig. 20c, the failure is isolated to be in the microbump area between the two stacked dice, based on the feature dimensions of the device [20].

Unlike the conventional method, which provides a 2D current path, a different algorithm could be used to calculate a 3D current path from the collected magnetic field data by SSM [24]. It has been proven that this alternative way of magnetic field data analyzing is promising to map 3D current path in 3D packages with stacked die configuration [24].

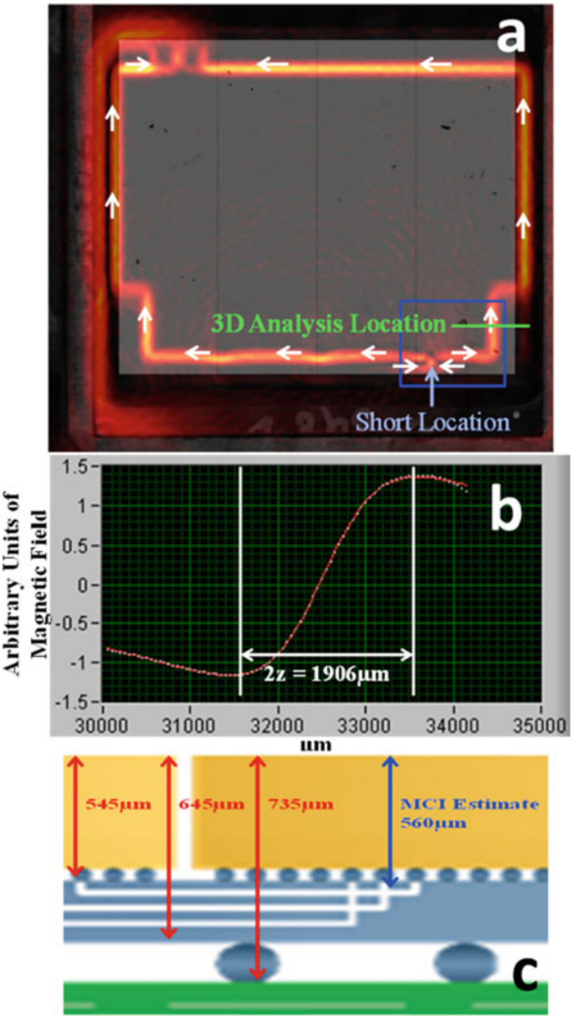


Fig. 20 **a** SSM analysis of a short failure in a package with stacked die configuration. X and y location of the short can be determined by the current analysis. **b** and **c** Z dimension analysis of the current path location shown in (**a**). The distance between the minimum and maximum magnetic field along the green line in (**a**) is measured as 1906 µm. The distance between the sample surface and the failure location is then estimated to be 560 µm, by subtracting the sensor to sample distance (393 µm) from the sensor to current path distance (953 µm, half of 1906 µm (Adapted from ref. [20])

MFI provides current mapping instead of locating the dissipated heat from short defects and is a complementary technique to LIT. There are some semiconductor device failures which could not be located by LIT. For example, high resistance routing in the failing structure, like TSV, could shadow the dissipated heat coming from short defects and preventing a good fault isolation by LIT. Short failures with extremely small resistance are also very hard to get a hot spot using LIT. If the same failing structure has multiple short defects, the z-dimensional information of the defects provided by LIT is not accurate. MFI could be an excellent alternative technique in these cases to effectively identify three-dimensional locations of defects.

3.2 Optical Fault Isolation Tools

Optical fault isolation tools commonly used in die level EFA can be divided into two main categories of techniques: passive and active techniques [25]. A popular passive technique is infrared emission microscope (IREM) or photon emission microscopy (PEM), which is based on electroluminescence. Its failure detection is achieved with a photon-sensitive camera in conjunction with a microscope module. IREM can pinpoint the defective location by detecting a very faint photon emission from the defective circuit when a device under test (DUT) is electrically powered [26–29].

Active techniques are implemented in a scanning optical microscope with a laser source, termed as laser scanning microscope (LSM). Laser beams are used to stimulate failures sensitive to carrier generation or thermal impetus; the amplified signals are utilized to locate defects in a DUT. The transmittance of near infrared (NIR) light to Si is the highest around the Si bandgap, ~1107 nm wavelength, and drops dramatically with increasing doping concentrations. A 1064 nm wavelength laser is used for carrier stimulation, and a 1340 nm laser source is used for thermal stimulation. These wavelengths were chosen to be as close to Si bandgap as possible to optimize laser delivery through the backside [25].

Optical beam induced resistance change (OBIRCH) and thermally induced voltage alteration (TIVA) are fault isolation techniques using thermal stimulation laser source of 1340 nm wavelength. OBIRCH uses a constant voltage source to bias the DUT and monitors the current change due to the thermal stimulation. While TIVA uses a constant current source for biasing and investigates the voltage change, optical beam induced current (OBIC) and light induced voltage alteration (LIVA) utilize carrier stimulation laser of 1064 nm wavelength. LIVA uses a constant current source to monitor changes in voltage, similar with TIVA. Soft defect localization (SDL) and laser assisted device alternation (LADA) use testers or ATE to detect the laser induced changes in DUT. SDL is with thermal stimulation laser source of 1340 nm wavelength, while LADA is with carrier stimulation laser of 1064 nm wavelength. They both are very effective for the isolation of marginal or soft failures in DUT [2, 26].

Semiconductor devices, especially central processing units (CPU), are usually made with heavily doped Si. When applying the optical fault isolation techniques, mirror- finish die thinning to less than 40um thick Si is typically required. Sometimes, an anti-reflection coating to Si backside is utilized to maintain a uniform laser power delivery through the Si backside by reducing the reflectance [30].

3.2.1 Infrared Emission Microscope (IREM) or Photon Emission Microscopy (PEM)

IREM or PEM is the most common passive optical technique for die level device fault isolation [26]. As illustrated schematically in Fig. 21, electroluminescence from the DUT, the photon emission due to electrical stimulation by either static or dynamic powering up the DUT, is detected by the photon-sensitive detector and presents as emission sites [26]. By comparing the same IREM data obtained from a failing vs a passing unit, the abnormal emission observed on the failing unit is identified as possible failing locations. IREM is typically used first during the die level EFA, as it is non-invasive, easy to use, and having high observability [27].

Figure 22 demonstrates the application of IREM in a static random-access memory (SRAM) device failure analysis [26]. After die backside sample preparation, static IREM analysis is performed on the DUT first. But it cannot detect any abnormal emission site, as the static power-up mode does not exercise the failure. As presented in Fig. 22a, the subsequent dynamic IREM analysis, by running a repeating functional test pattern on the DUT, shows an abnormal photon emission site on the failing unit. PFA reveals a defect at transistor gate level, which causes the short failure and the abnormal electroluminescence, highlighted by the red arrow in Fig. 22b [26].

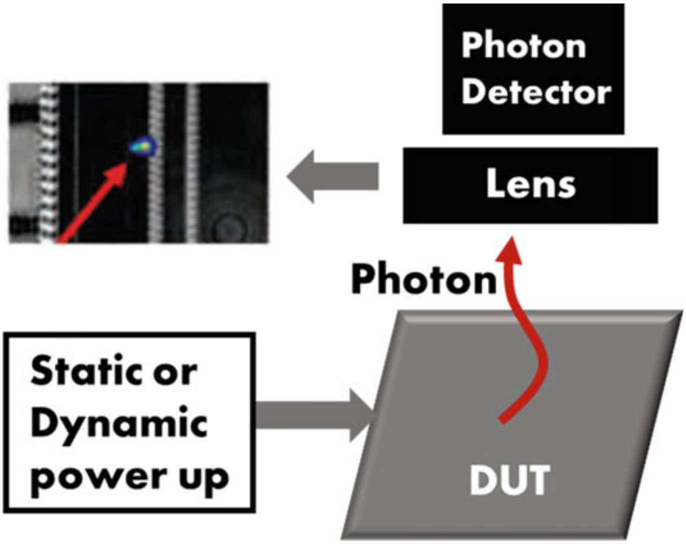


Fig. 21 Schematic of a IREM or PEM setup (Adapted from ref. [26])

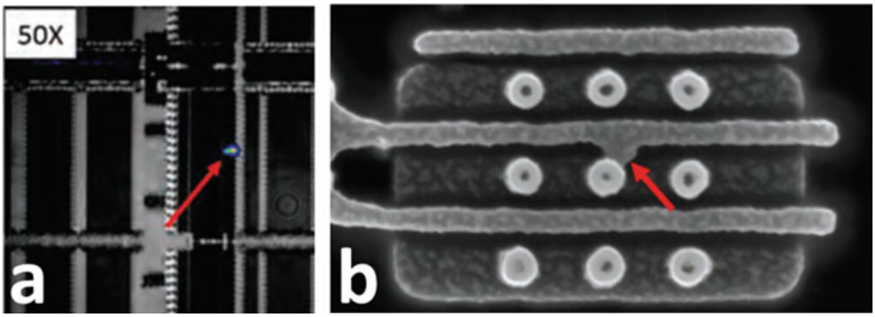


Fig. 22 **a** Dynamic IREM (or PEM) image from the failing unit showing an emission site. **b** SEM image of the abnormal emission site post-PFA displaying the defect leading to the short failure (Adapted from ref. [26])

3.2.2 Thermally Induced Voltage Alteration (TIVA) and Seebeck Effect Imaging (SEI)

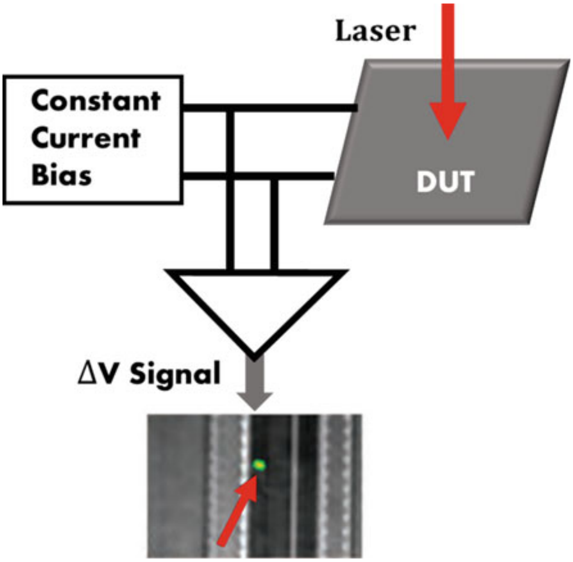
Thermally induced voltage alteration (TIVA) and Seebeck effect imaging (SEI) are popular active optical fault isolation techniques developed for localizing short and open failures in semiconductor devices [31, 32]. As illustrated in Fig. 23, a laser source of 1340 nm wavelength is used to scanning the DIU powered by a constant current supply [26]. TIVA imaging is very effective to isolate short failures; laser illumination on the defects causing short failures changes the resistance of the circuit and the corresponding IC voltage [31]. SEI utilizes the Seebeck effect; localized heating around the defects causing open failures generates a thermal gradient that produces a voltage gradient [31]. The constant current biasing approach provides an extremely sensitive method for detection of subtle changes in the IC power demand. Short (by TIVA imaging) and open failures (by SEI) are localized by monitoring the voltage change, caused by the laser induced thermal stimulation to the defects of DIU.

As illustrated in Fig. 11b, if defects are located on internal signals, debug tester is needed to execute the failing pattern in DIU. Figure 14 displays the TIVA date collected from the powered-on DIU, showing the short defects at transistor level. PFA on the localized defect revealed protruded Ti across fin of the failed FinFET transistor [9].

3.3 Nanoprobing and E-beam Imaging

I-V curve tracing illustrates the electrical character of semiconductor device and is indispensable for failure verification and localization. Figures 6 and 12 display the I-V curve tracing obtained non-destructively from the DUT. [6, 9]. For die level failure analysis, after failure is isolated to a small area by EFA, exhaustive PFA is needed

Fig. 23 Schematic of a TIVA setup (Adapted from ref. [26])



to reveal the nanometer scale defects within die. As demonstrated in Fig. 15, after de-packaging of the DUT, Si De-layering either from the front side or back side is performed to expose the Si circuitry layer by layer [10]. Nanoprobing on Si circuitry is performed layer by layer in the AOI to understand the electrical characteristic of the DUT and locate the defects. There are two major techniques to perform nanoprobing, atomic force microscopy (AFM)-based methods and SEM originated approaches [2]. Conductive AFM (CAFM), atomic force probing (AFP), scanning capacitance, and electrostatic force microscopy are all AFM-based nanoprobing techniques [2]. They have been widely used in die level FA for transistor level probing. SEM-based nanoprobing utilize piezoelectric-driven probes inside the SEM chamber to precisely land on sub-micron or nanometer scale Si circuitry. I-V curves are collected during nanoprobing to verify and localize the failures. Figure 16a illustrates the I-V curve by transistor level nanoprobing, showing the short failure in the defective transistor [9].

For SEM-based nanoprobing setup, it is convenient to use e-beam imaging-based fault isolation techniques to localize open and short failures [33]. Nanoprobers are used to land on the failing nets and monitor for current, while an e-beam is scanning over the DIU. The absorbed electrons inside the conductors of DIU produce an electron beam absorbed current (EBAC) and generate resistive contrast imaging (RCI), which is used to locate open and short failures. When e-beam is scanning over the DIU, heat is generated within the e-beam reaction volume, which can change the resistance of metal interconnects and short defects in DIU. The resistance change is detected by e-beam induced resistance change (EBIRCH) and is effective for short localizations. At sufficiently high accelerating voltages, the e-beam reaches Si and creates electron–hole pairs, which are detected to yield e-beam induced current

(EBIC), a successfully technique for p–n junction defect identification. The e-beam imaging techniques complement the I-V analysis of nanoprobing and are regularly used in die level failure localization [2].

3.4 E-beam Probing

Die level failure analysis methodologies discussed so far involve optical fault isolation, de-packaging, Si de-layering, and nanoprobing. E-beam probing is the alternative approach for semiconductor device die level failure analysis [10, 34–37]. As demonstrated in Fig. 24, Si backside of the DUT is thinned down to the transistor level and docked to a test board inside a SEM. The test board is connected via feedthroughs to a debug tester sitting outside of the SEM vacuum chamber [10]. Test patterns are exercised on the DIU while electron beam scans across the die. Secondary electron yield is mainly a function of primary electron beam energy, local surface potential, topography, and material electronic property. It is also modulated by device local bias and provides a dynamic voltage contrast imaging on the DIU under test. Failure locations with nanometer resolution can be identified by analyzing the secondary electron voltage contrast imaging.

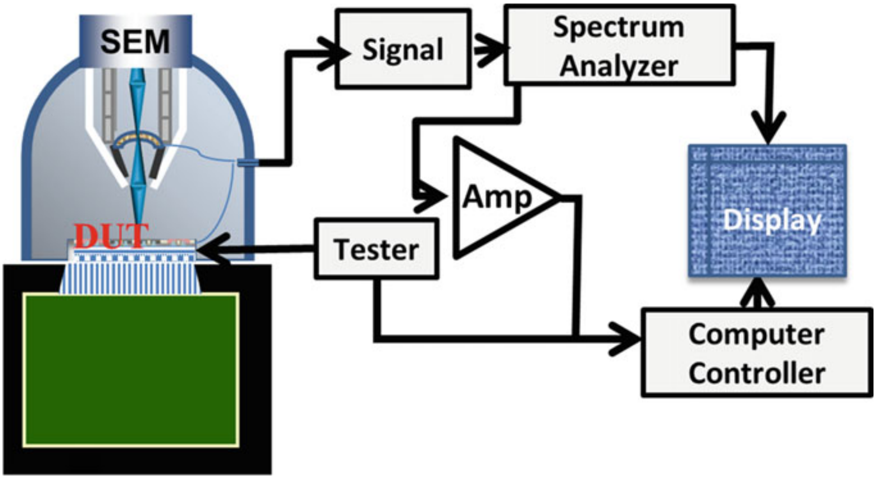


Fig. 24 Schematic of e-beam probing (adapted from ref. [10])

4 Physical Failure Analysis

Root cause understanding of failures is critical for process improvement and optimization during semiconductor device product development, which typically needs physical failure analysis (PFA) to disclose the details of defects. Thorough PFA typically utilizes techniques of sample preparation, defect imaging, and material analysis.

4.1 Sample Preparation Techniques

4.1.1 Mechanical Polishing

Mechanical polishing has been the “classical” method in PFA for cross section, planar grinding, de-packaging, and Si die backside thinning. As the size of interconnects in recent advanced packages decreases to a few micrometers, getting good alignment as well as artifact free finishing is very challenging by this classical mechanical polishing technique. However, if using the conventional Ga ion beam-based focused ion beam (FIB) technique, it is very time consuming to cross-sectional TSVs with more than one hundred micrometers in length [38]. To fill in the technical gaps, a couple of advanced sample preparation techniques is developed in the industry, including plasma-FIB, femtosecond (fs) and nanosecond (ns) laser ablation, and broad-beam ion milling.

4.1.2 Laser Ablation Techniques

Laser ablation techniques could provide a very short TPT, due to their fast material removal rate. However, they could not be used as an independent cross-sectional technique, because of the relatively low spatial resolution and laser induced thermal damage [39]. To clean up the μm range heat affected zone and improve resolution, FIB or ion milling is applied post the laser ablation. Integrating conventional FIB and SEM with femtosecond (fs) laser ablation has proven to be a very effective cross-sectional technique. It has much better location control, very small laser induced damage to the sample, and a high material removal rate of 103 $\mu m^3/s$. The system has been used to perform artifact free TSV cross sections, having comparable TPT to the plasma-FIB system [39].

4.1.3 Plasma-FIB

The maximum beam current is about 65nA in a conventional Ga ion beam-based FIB system. While a commercially available plasma-FIB system with an inductively coupled plasma (ICP) source can provide a focused Xe beam in a wide beam current range from several pA up to 2μA [36]. Thus, the TPT of sample preparation process with similar finishing could be 20–100 times faster using plasma-FIB system, comparing with conventional FIB. It has been demonstrated that plasma-FIB combined with SEM technique can reveal subtle defects in a large cross-sectional area with high quality and fast TPT. For instance, delamination or cracks in TSVs, microbumps, and stacked dice. Figure 10b displays the plasma-FIB cross section of TSVs with voids [36]. Additionally, plasma-FIB has demonstrated its extensive applications, such as Si de-layering, cross section, and Si dry etching, in die backside sample preparation for die level failure analysis [2, 40].

4.1.4 Broad-Beam Ion Milling

Broad-beam ion milling is a low-cost sample preparation technique for package and board level failure analysis, comparing with the advanced plasma-FIB and fs laser systems. Before performing ion milling, units are typically cross sectioned either by mechanical polishing or laser ablation close to the area of interest. Within a couple of hours or even shorter time, about 1mm^2 cross-sectional area with artifact free final finishing could be obtained by ion milling. The ion milling method minimizes curtaining artifacts in FIB cross sections. Because it has varied incident angles and the large area of homogenous ion dose. The ion milling technique has been employed for TSV cross sectioning with fast TPT and artifact free sample finishing [3, 8].

SEM images of TSVs in a package, cross sectioned by ion milling technique are presented in Fig. 10c. Within one hour, a large area of artifact free cross-sectional view in the device is disclosed by ion milling cross section [8]. TSV seed layer details are clearly displayed in the SEM image (the white frame of Fig. 10c).

4.2 Defect Imaging Techniques

4.2.1 Scanning Electron Microscopy (SEM)

Once defects are revealed during PFA, defect imaging techniques are employed for detailed analysis. Optical microscopes are typically used for gross defects. For submicron defects, SEM is utilized in a daily base. Secondary electron imaging is often applied to obtain morphology details of defects; backscattered electron imaging is helpful to highlight elemental difference; low energy SEM is required when imaging low k ILD layers in Si, which are prone to have E-bema induced damages [2].

4.2.2 Transmission Electron Microscopy (TEM)

Despite the limitation of very small (typically sub-micron scale) field of view, TEM can provide atomic-level spatial resolution and is utilized daily in die level failure analysis to expose nanometer scale defects. Before the adoption of FIB technique, TEM sample preparation used to be extremely time consuming. FIB-SEM enormously reduces the TPT of TEM sample preparation. A very thin slide (<100 nm) of the sample containing the defects is lifted out by FIB-SEM for TEM analysis. Figure 16b displays the TEM image of the FinFET transistor having nanometer scale defects across fin, causing the short failure [2, 9].

4.3 Material Analysis Techniques

For in-depth root cause investigation, it is crucial to have thorough material analysis around the defects. Elemental, crystallographic orientation, chemical composition and state, as well as spectral fingerprints of materials around the area of interest, could be obtained through combined applications of material analysis techniques, including energy-dispersive X-ray spectroscopy (EDX), electron backscatter diffraction (EBSD), X-ray photoelectron spectroscopy (XPS), time-of-flight secondary ion mass spectrometry (TOF–SIMS), Fourier transform infrared spectroscopy (FTIR), and atomic force microscopy-based infrared spectroscopy (AFM-IR) [3].

4.3.1 Energy Dispersive X-Ray Spectroscopy (EDX)

In conjunction with electron microscopy, including both SEM and TEM, EDX can produce spectra of X-ray counts versus X-ray energy. X-rays with characteristic energy unique to the ionized atoms in the sample are generated because of the interaction between the electron beam and the sample. Elemental composition of the sample about 1–3 μm in depth could be obtained by SEM–EDX, depending on the incident electron beam energy and the reaction volume with the sample [41]. SEM–EDX is typically employed as a non-destructive material analysis technique to quickly characterize materials on the sample surface, with μm range spatial resolution. It is also often used to disclose chemical composition of the defects revealed by PFA. TEM–EDX can provide much higher spatial resolution than SEM–EDX. Because the reaction volume of the high energy electron beam is much smaller in the ultra-thin TEM samples. However, TEM sample preparation is time consuming and only a small area could be inspected each time. Figure 16c illustrates the TEM–EDX mapping around the FinFET transistor short defect, showing extra titanium extending across fin, leading to the short failure.

4.3.2 Fourier Transform Infrared Spectroscopy (FTIR) and Atomic Force Microscopy-Based Infrared Spectroscopy (AFM-IR)

FTIR can provide unique IR spectral fingerprint of molecules as the pattern of absorption peaks in the IR spectra, by measuring the amount of infrared light absorbed by a sample as a function of the IR frequency. It has been widely used for chemical analysis in semiconductor devices, especially for the characterization of organic materials [42]. However, it is not capable for sub-micron chemical characterization. Because the diffraction of the long IR wavelengths (2.5–20 μm) limits the spatial resolution of FTIR to a couple of microns.

By combining the spatial resolution of atomic force microscopy (AFM) with the chemical analysis capability of infrared (IR) spectroscopy, AFM-IR technique is very useful to characterize organic contaminations in the sub-micron scale [43]. Regional absorption of the IR radiation in the sample is created, when a tunable infrared laser focuses onto a small area of a sample near the AFM probe tip. By locally detecting thermal expansion of the sample resulting from the IR absorption, the AFM probe tip acts as an IR detector. The AFM-IR technique can overcome the spatial resolution limits of conventional FTIR technique, as the AFM tip can detect the thermal expansion in nanometer scale.

4.3.3 X-ray Photoelectron Spectroscopy (XPS) and Time-Of-Flight Secondary Ion Mass Spectrometry (TOF–SIMS)

In electron spectroscopy for chemical analysis (ESCA), also known as XPS, the sample surface is excited with monoenergetic Al kα X-rays, which generates photoelectrons. By measuring the energy of photoelectrons emitted from the sample surface, XPS can provide quantitative elemental information along with chemical state of a sample. The average analysis depth is about 5 nm [44].

TOF–SIMS uses a time-of-flight analyzer to measure the exact mass of ions and clusters emitted from the sample surface excited with a finely focused ion beam. It can provide elemental, chemical state, and molecular information from sample surfaces with about 1 nm analysis depth. By combining TOF–SIMS measurements with sputtering, thin film characterization with depth distribution information is obtained [45].

XPS and TOF–SIMS are surface-sensitive techniques with an analysis depth of less than 2–5 nm, in contrast to SEM–EDX, which has an analysis depth of 1–3 μm. Therefore, they are better suited for the compositional analysis of thin layers. Additionally, both techniques can be employed to characterize molecular information in organic materials, which is not possible if using SEM–EDX. For root cause understanding of failures in devices related to impurity or contamination at surface and interface, XPS or TOF–SIMS characterization of surface layers or thin film structures is critical.

4.3.4 Electron Backscatter Diffraction (EBSD)

Backscatter Kikuchi diffraction (BKD), also known as EBSD, is used to define the crystallographic orientation of microstructures in materials. Combining with SEM, EBSD data is typically acquired from a cross-sectioned sample. For the root cause understanding of electromigration (EM) induced open interconnects, it is critical to have detailed EBSD data [46–48]. Because of the much smaller interconnect dimension in advanced microelectronic packages, EM in semiconductor device interconnects becomes a major reliability concern. Due to Sn self-diffusion through a void nucleation and propagation mechanism, EM can cause solder joint open failures. The electron wind can also cause failures due to depletion of under bump metallization (UBM) by accelerating the dissolution of UBM on the cathode side. The UBM dissolution is found to be closely related to Sn grain orientations [46]. The time to dissolve 2 μm Ni UBM can vary from 100 hours (when Sn c-axis is in parallel to electron flow direction) to 1800 hours (when c-axis is perpendicular to electron flow), at given temperature and current density [47]. To disclose the relation between metal grain orientation and EM induced open failures, EBSD is widely used to analyze the Sn grain orientation distributions of samples [48].

High-resolution EBSD is also used for solder joint strain characterization for further reliability performance improvement [4]. As illustrated in Fig. 5c, the strain contour of the solder joint is obtained from the high-resolution EBSD analysis shown in Fig. 5b. Strain accumulates at the area between the solder bump void and copper pad and could initiate solder joint cracks under extended reliability stress [4].

Being a major concern for yield and reliability of packages with stacked die and TSV configuration, TSV extrusion is shown to be related to the mechanical properties of TSV [49]. Cu grain size distribution analysis has been performed by EBSD in two types of TSVs with different reliability performance.

The EBSD data suggests that TSVs with smaller and more uniform Cu grains have a higher yield strength are better for inhibiting extrusion [49].

5 Non-Destructive and High-Resolution Imaging Techniques

5.1 *Optical and Infrared (IR) Imaging*

During the failure analysis flow, shown in Fig. 1, optical and IR imaging is typically used to inspect the entire DIU for any possible gross defects before the detailed FA investigation [2]. Adding the inspection step can save enormous time in FA if a gross defect is captured. As illustrated in Fig. 3a, optical inspection detects solder residue between two solder joints, which caused the electrical short [2].

IR imaging can be used to examine die active side circus from the back side of the die or wafer, because Si is transparent at infrared wavelengths longer than 1.1 microns

[2]. Due to band gap shifts and free carrier absorption or scattering, doped silicon is much less transparent than un-doped silicon. Advanced IR imaging techniques, such as near-IR imaging, shortwave-infrared (SWIR) imaging, at longer wavelengths (the 1300–1500 nm range), or infrared confocal imaging with improved signal to noise ratio, could be used to non-destructively examine semiconductor devices with doped Si [50–52]. Real-time inspection of die front side circus, enabled by the IR imaging, is very efficient to detect internal die cracks, die active side cracks, and die active layer circus defects. If using regular optical microscope, these defects located inside Si are invisible. Figure 25 demonstrates an IR image directly taken from die backside of the DIU, without any Si die thinning sample preparation. The dark shadow pointed by the red arrows in the image indicates an internal die crack, which propagates into die active area and damages Si circuitry, leading to the electrical failures [2].

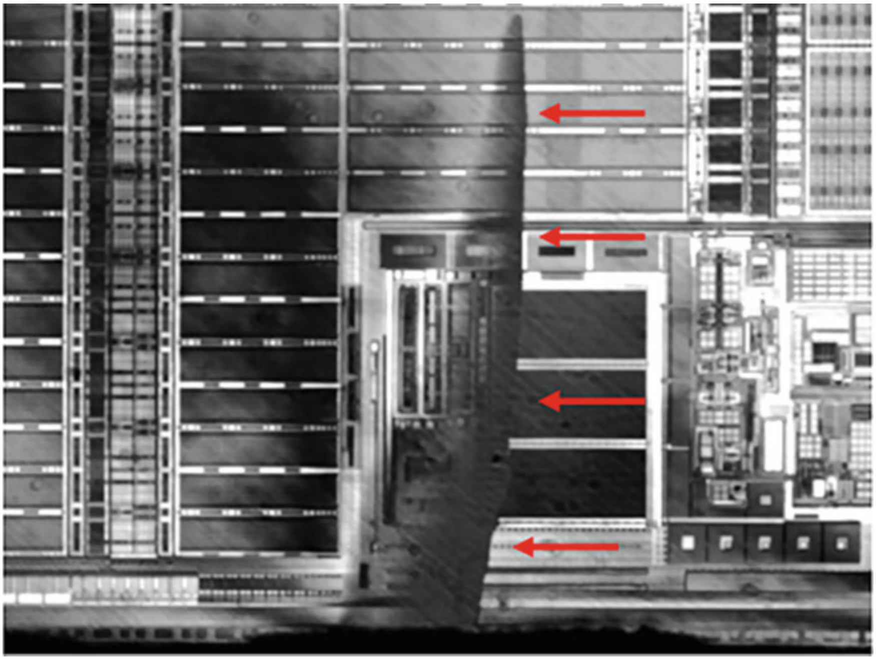


Fig. 25 The IR image taken from die backside showing an internal die crack highlighted by the red arrows (Adapted from ref. [2])

5.2 Scanning Acoustic Microscopy (SAM)

SAM is a widely used tool for the non-destructive detection of delamination and voids in various packaging materials [53–56]. Excited by an extremely short electrical discharge, the acoustic transducer with piezoelectric element transmits an ultrasonic pulse into a semiconductor device, as illustrated in Fig. 26. Water is employed as the coupling medium for the transmission of the acoustic perturbation from the transducer to the package. The reflection and transmission of acoustic wave from medium 1 to medium 2 are described by the reflection coefficient "*R*" and transmission coefficient "*T*" at any interface, which are calculated according to the following equations.

$$R\left(90^{\circ}\right) = \frac{Z_2 - Z_1}{Z_2 + Z_1} \tag{3}$$

$$T\left(90^{\circ}\right) = \frac{2Z_2}{Z_1 + Z_2} \tag{4}$$

Z_1 and Z_2 are the acoustical impedance of medium 1 and medium 2, respectively. The acoustic impedance Z is estimated as

$$Z = \rho c \tag{5}$$

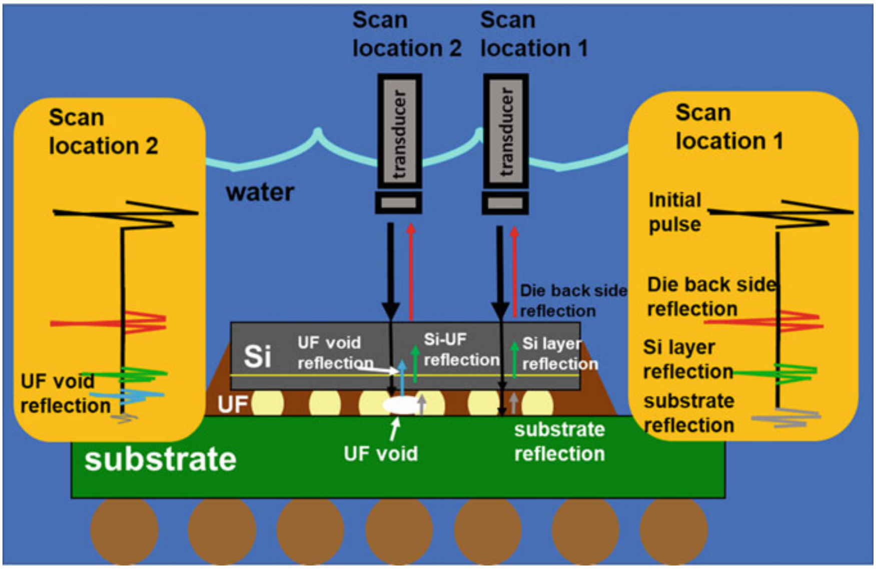


Fig. 26 Schematic of the SAM technique (adapted from ref. [3])

"c" is the acoustic wave velocity in the medium, while "ρ" is the volumetric mass density of the medium. For longitudinal acoustic waves traveling in an isotropic and homogeneous medium, c could be expressed as

$$c = \sqrt{\frac{E(1-\mu)}{\rho(1+\mu)(1-2\mu)}} \tag{6}$$

where "μ" is the Poisson's coefficient; "E" stands for Young's modulus. Acoustic Impedance is characteristic of materials and is defined by Young's modulus, Poisson's coefficient, and density [57, 58].

The acoustic impedance of air, which is about $0.389 \cdot 10^{-3}$ [Kg/m^{2}s] × 10^6, is very small comparing with other materials [58]. According to Eqs. (3) and (4), assuming medium 2 is air, which means $Z_2 \approx 0$. The transmission coefficient "T" and reflection coefficient "R" from any medium to air are equal to 0, and -1, respectively, representing nearly 100% reflection. For that reason, SAM is a very effective technique for air gap defect detection in semiconductor devices, like delamination, blisters, cracks, and voids.

Figure 26 demonstrates the pulse-echo method of SAM technique. It provides information associated with both intensity and time of flight of the ultrasonic beam [59–61]. The same ultrasonic transducer is used for transmitting the acoustic pulse, as well as receiving echoes occurring at boundaries of different acoustic impedance values. At the scan location 2 of Fig. 26, the echo reflected by a UF void shown in blue has a unique time of flight compared with the die back side. Therefore, it contains the z-dimensional information of the defect, which could be revealed by either B-scan SAM or C-scan SAM (CSAM) with tomographic acoustic micro-imaging (TAMI) technique [62–64]. B-scan SAM could present a cross-sectional view of defects and is virtual cross sectioning of the CSAM plane [64]. While CSAM presents a plane view of the size and location of features in the test specimen. TAMI CSAM provides images from multiple depths within the sample simultaneously, by setting a sequence of multiple gates of data acquisition adjusted to resolve the interfaces of interest.

Both lateral and axial resolution of the CSAM scan need to be enhanced to detect subtle defects in semiconductor devices. CSAM transducers define CSAM scan resolution. Lateral or X–Y-dimensional resolution is determined by transducer beam spot size, the smaller the better. While Z-dimensional or axial resolution, which reveals the depth of delamination gaps, is related to transducer frequency, the higher the better [64, 65]. However, the bigger attenuation of higher frequency acoustic waves in the medium limits the non-destructive applications of high frequency transducers, especially for samples with thick Si. The focal length of CSAM transducers needs to be optimized for a given Si thickness to perform non-destructive CSAM analysis. For high-resolution and non-destructive CSAM imaging of semiconductor devices, high frequency transducers with focal length designed for a specific die thickness would be desired [65].

GHz-SAM technique is invented to further improve SAM resolution [7, 66]. To ensure a much higher spatial resolution of ~1 μm, the frequency of the acoustic

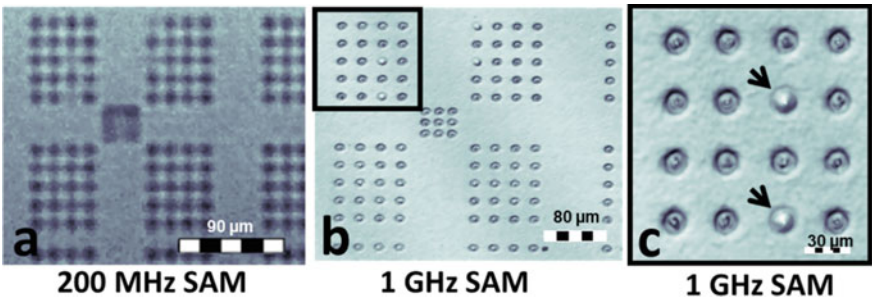


Fig. 27 CSAM images of TSVs taken by **(a)** 200 MHz transducer, **(b)** and **(c)** 1 GHz transducer (Adapted from ref. [7])

transducer is extended to the GHz domain. On the other hand, the penetration depth of GHz-SAM is much smaller due to the high frequency. Only a drop of water is used to cover the sample surface in a GHz-SAM setup, instead of immersing the sample fully in water. GHz-SAM has been used to detect TSV voids, as demonstrated in Fig. 27. The GHz-SAM image shown in Fig. 27b, c clearly displays the TSVs with voids, compared with the CSAM image using a 200 MHz transducer shown in Fig. 27a. FIB cross section shown in Fig. 10b confirms the GHz-SAM signal, showing the voids with about 1 µm in diameter and 10 µm in depth [7]. GHz-SAM could also provide additional information besides the greatly improved image spatial resolution, for example, the stress around TSVs induced by defects. It is possibly due to the reflection of surface waves generated by the GHz transducer [7].

Through transmission SAM (TSAM) technique, especially with unique transducer design and improved signal to noise ratio, is very helpful to reveal defects embedded deep in devices, such as blisters in organic substrate layers, and defects in embedded Si chips. There are two transducers aligned facing each other from both sides of the sample in the TSAM setup, as shown in Fig. 28 [59]. The receiver transducer detects the intensity of the crossing wave at the rear side of the specimen, while the emitter transducer generates a pulse propagating through the specimen. During the acoustic pulse propagates through the sample, the beam energy attenuates. At location 2 of Fig. 28, the propagation of the acoustic wave is interrupted by a crack or an air gap, whereas at location 1, the beam goes through a clean path. The crack at location 2 significantly attenuates the amplitude of the transmitted pulse and gives a “dark contrast” in the TSAM image, indicating defects.

New TSAM emission transducers with frequency in the range of 20–50 MHz, which are directly focused without lenses, can combine with a focused 10 MHz reception transducer to improve the TSAM spatial resolution up to 50 µm. The TSAM setup has been employed to image a PCB having embedded chips. Defects around the four die corners of an embedded chip in the PCB are disclosed by the non-destructive TSAM technique, as shown in Fig. 29 [67].

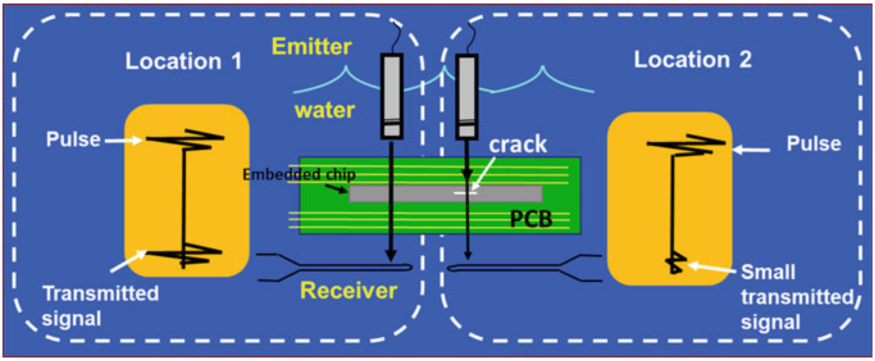


Fig. 28 Schematic of the TSAM technique (Adapted from ref. [3])

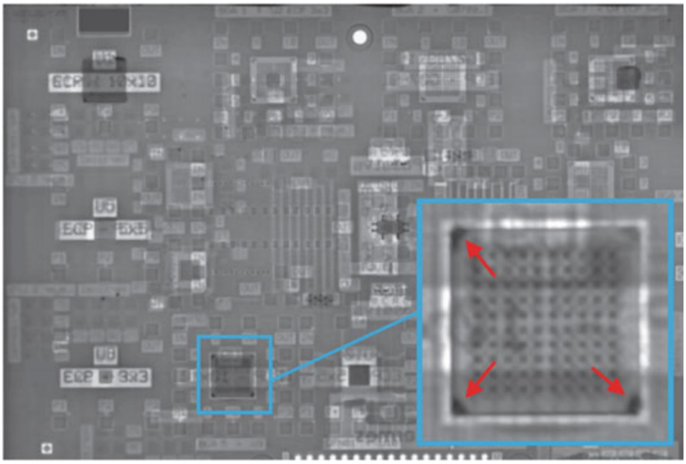


Fig. 29 TSAM of a PCB with embedded chips showing defects in one embedded chip (Adapted from ref. [66])

5.3 2D X-ray Radiography

2D X-ray radiography is a commonly used tool in the semiconductor industry for the solder joint defect inspection of devices. As illustrated in Fig. 30, samples are put between the detector and the punctiform X-ray source [3, 68]. The ratio of the distance between the source and the detector (D), to the distance from the source to the sample (d) is the geometric magnification M. To get high-resolution 2D X-ray images, the X-ray beam spot size is set to be very small, and the placement of the sample is very close to the source [68]. As demonstrated in Fig. 30, 2D X-ray radiography is a projection of a 3D object. Therefore, information in depths and volume needs to be achieved by observing the object from several different angles.

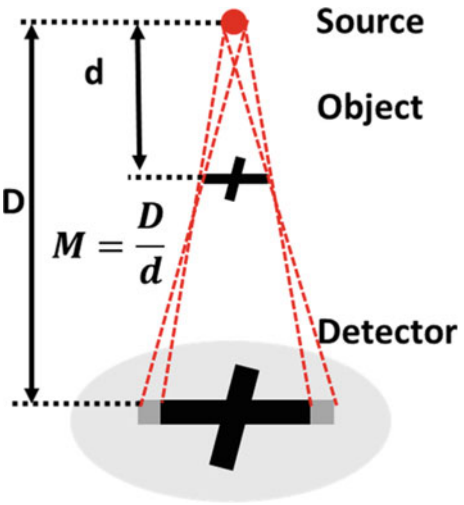


Fig. 30 Schematic of a 2D X-ray radiography setup (Adapted from ref. [3])

2D X-ray has been employed to effectively detect solder void evolution post-multiple reflows in packages [69]. With optimized imaging conditions, such as sample tilt and rotation angles, real-time non-destructive 2D X-ray imaging could be used to scan the failing area and reveal package defects, such as solder joint bridging, solder non-contact open, and non-wet. By cross section, the same solder joints post-non-destructive 2D X-ray imaging, correlation between 2D X-ray images and FLI solder joint defects can be established, as demonstrated in Figs. 31 and 32. By comparing the image contrast and solder joint geometry shape with those from a golden solder joint, defective solder joints, like partial non wet, complete non-wet, non-contact open, and solder joint bridging could be easily analyzed via real-time non-destructive high-resolution 2D X-ray imaging with a very short TPT [70]. Enabled by the real-time, non-destructive, and high-resolution 2D X-ray imaging technique, in situ study of package failures happened during reliability tests, such as reflow, could provide direct observation of how the failure happens; thus, it reveals root causes and solution paths [71].

However, 2D X-ray imaging may not be applicable for micrometer or sub-micron scale defects in semiconductor devices, such as microvoids in Cu vias, substrate trace cracking, subtle solder extrusion between solder interconnects. Because the small contrast difference in a single 2D X-ray image is shadowed by other layers or interconnects in packages [72].

5.4 3D X-ray Computed Tomography (CT)

3D X-ray CT has been widely used to detect defects in micrometer or sub-micrometer scale [73–75]. Figure 33a illustrates the schematic of Zeiss Xradia 3D X-ray CT

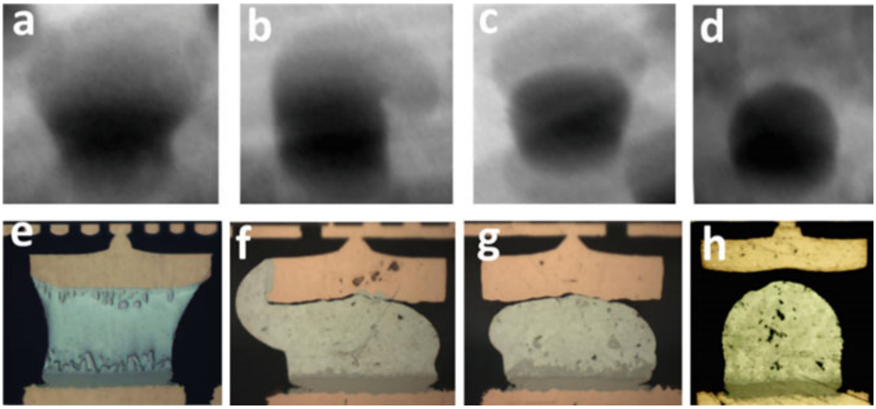


Fig. 31 2D X-ray images taken non-destructively and the corresponding optical images of FLI solder interconnects cross-sectioned post-2D X-ray imaging: **(a)** and **(e)** normal or defect free, **(b)** and **(f)** partial non-wet, **(c)** and **(g)** complete non-wet, **(d)** and **(h)** non-contact open (Adapted from ref. [3])

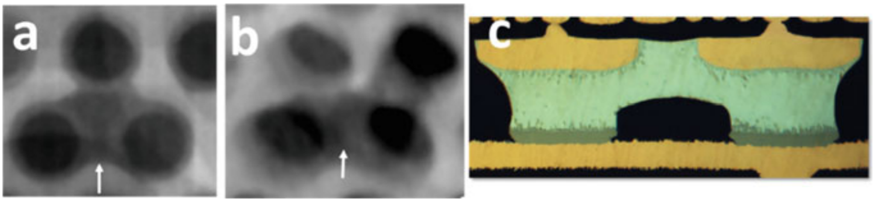


Fig. 32 2D X-ray images of FLI solder joint bridging taken non-destructively with a top-down view **(a)**, a tilted view **(b)**, and the corresponding optical image taken from the sample cross-sectioned post-non-destructive 2D X-ray imaging **(c)** (Adapted from ref. [3])

setup. An X-ray source is used to radiate the object at different tilt angles. A rotating stage is employed to provide angular displacement in equally spaced angles. While a detector is applied to collect the 2D X-ray images at each angle. After the scan data collection, all 2D images collected at various angles are mathematically superimposed and processed to obtain a three-dimensional image of the sample, as illustrated schematically in Fig. 33b. Because the acquired 3D X-ray CT data has volumetric information of the device, analysts can freely display virtual planar or cross-sectional view at any given location of the three-dimensional data set.

High-resolution 3D X-ray imaging combined with non-destructive FI techniques is very effective to detect trace and interconnect failures in packages. Failure locations identified by EOTPR (data is shown in Figs. 17 and 18) are imaged by 3D X-ray CT, and the results are shown in Fig. 9. The substrate trace crack at the failure location is displayed as "a 3D view," "a virtual cross-sectional view," and "a virtual planar view," respectively [75]. To achieve sub-micron resolution without the need for placing the sample close to the source, an X-ray detector with proprietary X-ray

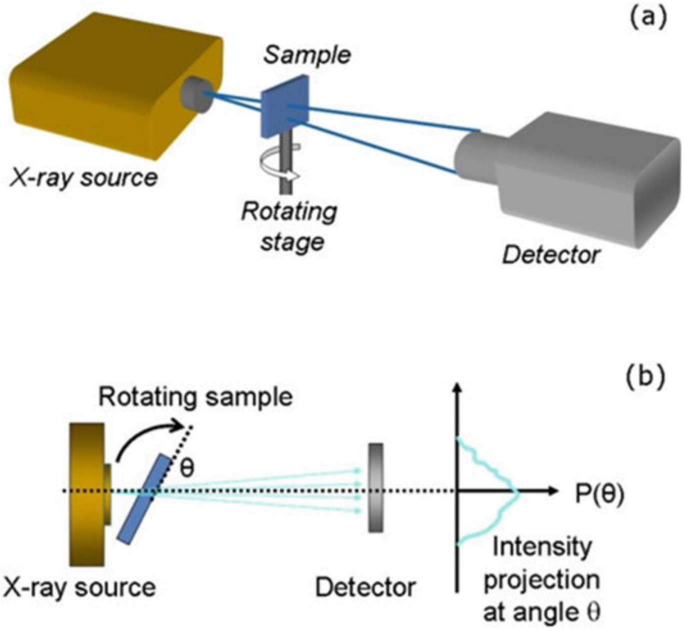


Fig. 33 **a** Schematic of the 3D X-ray CT setup consisting of X-ray source, rotating stage, and detector. **b** The intensity projection at each angle are superimposed and mathematically processed to generate the three-dimensional image (Adapted from ref. [3])

optics is designed. The unique setup allows 3D X-ray high-resolution scanning of intact samples, regardless of sample size, without sample-source collision [75].

3D X-ray CT has been successfully applied in the semiconductor package assembly process to monitor micrometer size defects in devices, for example voids in TSV, microbump integrity, partial non-wet and voids (<5 μm) in solder joints, and voids or cracking in substrate Cu vias. The 3D X-ray CT of TSVs and microbumps in packages with multi-die stacking configuration is shown in Fig. 34. Figure 34a displays a virtual cross-sectional view presenting TSVs and microbumps in a package having multiple memory dice stacked on top of the base die or logic die. Microbump voids are found by the 3D X-ray CT inspection and illustrated in Fig. 34b, the zoom-in image of the white frame area in (a). The TSV voids can be distinctively captured by 3D X-ray CT, as demonstrated in Fig. 34c [76, 77]. The progressive studies of device failures during reliability tests, such as consistent current flow at elevated temperatures, temperature cycling, and reflow, can be enabled by the 3D X-ray CT technique. Because it can non-destructively provide high-resolution information of semiconductor devices. These progressive studies have been employed to resolve key reliability issues in semiconductor devices [69, 75].

The current 3D X-ray CT technique has been applied in the field of kinetic study, process control, and failure analysis of semiconductor devices. However, the through

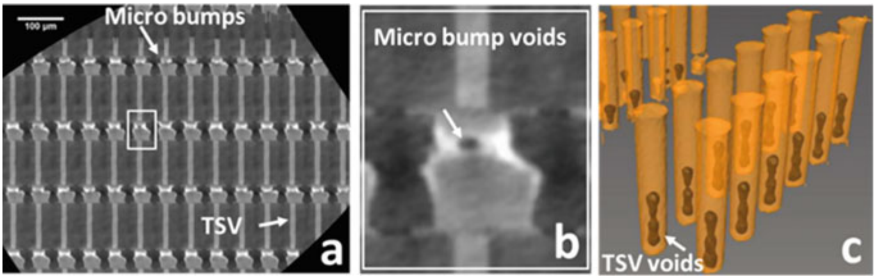


Fig. 34 **a**, **b** 3D X-ray CT of microbumps and TSVs in a 3D package with stacked die configuration (adapted from ref. [75]). **c** 3D X-ray CT of TSVs showing distinctly the TSV voids (Adapted from ref. [76])

put time (TPT) for high-resolution scan is usually long. Because the flux and brightness of lab-scale X-ray sources are relatively low, which causes the long exposure or image capture time, for high-resolution imaging. To cope with highly absorbing Cu or solder components in semiconductor devices, X-ray beam with energy higher than 100 kV is usually used during the scan. This reduces the image resolution due to the beam spot size blooming at higher energy. Additionally, high energy X-ray imaging makes the organic packaging materials "invisible," as it sacrifices the phase contrast. The applications of 3D X-ray CT are limited due to these factors. Defects in materials with low Z numbers, for instance, voids, delamination, and cracks in solder resist, molding compound, underfill, and other dielectric materials in semiconductor devices, are hard to image by current lab-scale 3D X-ray CT. One way to overcome the challenges is to extract a very small piece (~100 µm in size) containing the possible failures out of an intact device by advanced sample prep techniques. The technique is also called nanoscale phase contrast 3D X-ray CT, which has been used to reveal intermetallic compound details in microbumps or subtle defects in Si layer routing. However, very precise three-dimensional fault isolation and superior sample preparation is needed to integrate the nanoscale 3D X-ray CT into daily fault isolation and failure analysis flow of semiconductor devices. Further development of lab-scale 3D X-ray CT technique to enable non-destructive imaging of all types of defects in semiconductor devices with short TPT would be more preferential. The synchrotron 3D X-ray CT study of semiconductor devices discloses the feasibility of expanding applications of 3D X-ray CT in semiconductor industry, also lists some suggestions to the development of next generation lab-scale 3D X-ray CT systems [75, 78].

6 Summary

Advanced fault isolation and failure analysis techniques for ADAS are reviewed as well as efficient FA flow and FA strategies to illustrate their applications. It has been demonstrated that I-V curve tracing, TDR, EOTPR, LIT, and MFI are effective fault isolation tools for package or board level failure analysis. Optical fault isolation tools, such as IREM (or PEM), TIVA, and SEI are typically used for die level fault isolation. Nanoprobing and e-beam imaging techniques are applied during PFA to further localize the subtle defects in die level FA. E-beam probing is an alternative approach for die level nanometer scale defect localization.

Sample preparation techniques, like broad-beam ion milling, ultra-short pulse laser ablation along with conventional FIB and SEM, and plasma-FIB are very effective to provide artifact free cross sections of interconnects in die, packages, and boards or systems. Ion etching in plasma-FIB or ion milling tools is commonly used for Si de-layering in die level failure analysis.

Optical and SEM are routinely used in ADAS FA for defect imaging. For subtle defects in die level FA, TEM is employed to achieve atomic-level resolution. EDX, EBSD, XPS, TOF SIMIS, FTIR, and AFM-IR have been demonstrated to be very useful material analysis techniques for root cause understanding of failures by providing information about crystal orientation, chemical composition, and interface contamination of defects.

Comprehensive understanding of failure rate distribution, various reliability tests, and manufactory process details is the first key step of ADAS failure analysis. An efficient, systematic, and artifact free FA flow well designed to save both cost and time is crucial to quickly disclose defects led to electrical test failures. In-depth root cause and failure mechanism understanding to offer solution paths of failures is the goal of failure analysis. Direct observation of the origination and growth of semiconductor device defects by non-destructive and in situ techniques are proven to be valid for thorough root cause studies.

Acknowledgements The editors would like to thank Dr. Tom Tong from Intel cooperation for his critical review of this chapter.

References

1. J. Fischer and G. Meyer, Advanced Microsystems for Automotive Applications, Berlin: Springer, 2014.
2. T. Gandhi, Microelectronics Failure Analysis Desk Reference 7th edtion, ASM International, 2019.
3. Y. Li and D. Goyal, 3D Microelectronic Packaging From Architectures to Applications Second Edition, Springer, 2020, p. 607.
4. P. Nowakowski, M. Ray and P. Fischione, "Solder bump joint failure investigation: from sample preparation to advanced stuctural characterizations and strain measurements," in *Conference Proceedings from the 45th International Symposium for Testing and Failure Analysis*, Portland, 2019.
5. E. Cattey, A. V. Vianen, M. Curiel, G. Huo, K. Dickson and Y. Meiche, "The Development of an EVB Socket Solution for Automotive Mixed-Signal ICs," in *Conference Proceedings from the 45th International Symposium for Testing and Failure Analysis*, Portland, 2019.
6. L. Cao, M. Venkata, J. Huynh, J. Tan, M.-Y. Tay and W. Qiu, "Lock-in Thermography for Flip-chip Package Failure Analysis," in *Conference Proceedings from the 38th International Symposium for Testing and Failure Analysis*, Phoenix, 2012.

7. I. D. Wolf, A. Khaled, M. Herms, M. Wagner, T. Djuric, P. Czurratis and S. Brand, “Failure and Stress Analysis of Cu TSVs using GHz-Scanning Acoustic Microscopy and Scanning Infrared Polariscopy,” in *Proceedings from the 41st International Symposium for Testing and Failure Analysis (ISTFA, ASM International)*, Portland, 2015.
8. A. Meyer, G. Grimm, M. Hecker, M. Weisheit and E. Langer, “Challenges for Physical Failure Analysis of 3D-Integrated Devices—Sample Preparation and Analysis to Support Process Development of TSVs,” in *Proceedings from the 38th In-ternational Symposium for Testing and Failure Analysis (ISTFA, ASM International)*, San Jose, 2013.
9. S. Mishra and D. R. Bockelman, “Case Study of ESD-induced Pin Leakage Failure Solved using Novel Powered TIVA Technique,” in *Conference Proceedings from the 44th International Symposium for Testing and Failure Analysis*, Phoenix, 2018.
10. T. Tong, H. J. Ryu, Y. Wang, W.-H. Chuang, J. Huening, P. Joshi and Z. Ma, “Electron beam probing of active advanced FinFET circuit with fin level resolution,” in *Conference Proceedings from the 44th International Symposium for Testing and Failure Analysis*, Phoenix, 2018.
11. Y. Li, Y. Cai, M. Pacheco, R. C. Dias and D. Goyal, “Non-destructive failure analysis of 3D electronic packages using both Electro Optical Terahertz Pulse Reflectometry and 3D X-ray Computed Tomography,” in *Proceedings from the 38th In-ternational Symposium for Testing and Failure Analysis*, Phoenix, 2012.
12. D. Smolyansky, *Printed Circuit Design,* p. 20, 2002.
13. D. Abessolo-Bidzo, P. Poirier, P. Descamps and B. Domenges, “Failure localization in IC packages using time domain reflectometry: technique limitations and possible improvements,” in *Proceedings from the 12th International Symposium on the Physical and Failure Analysis of Integrated Cir-cuits (IPFA)*, Singapore, 2005.
14. Y. Cai, Z. Wang, R. C. Dias and D. Goyal, “Electro Optical Terahertz Pulse Reflectometry - an Innovative Fault Isolation Tool,” in *Proceedings of the 60th Electronic Components and Technology Conference (ECTC)*, Las Vega, 2010.
15. [15] S. H. Hall, G. W. Hal and J. A. McCall, High-Speed Digital System Design: A Handbook of Interconnect Theory and Design Practices, NY: John Wiley & Sons, Inc., 2000.
16. R. Schlangen, S. Motegi, T. Nagatomo, C. Schmidt, F. Altmann, H. Murakami, S. Hollingshead and J. West, “Use of Lock-In Thermography for Non-Destructive 3D Defect Localization on System in Package and Stacked-Die Technology,” in *Proceedings from the 37th International Symposium for Testing and Failure Analysis*, San Jose, 2011.
17. F. Naumann, F. Altmann, C.Grosse and R.Herold, “Efficient non-destructive 3D Defect Localization by Lock-in Thermography utilizing Multi Harmonics Analysis,” in *Proceedings from the 40th Interna-tional Symposium for Testing and Failure Analysis*, Houston, 2014.
18. R. Dias, L. Skoglund, Z. Wang and D. Smith, “Integration of SQUID Microscopy into FA Flow,” in *Proceedings from the 27th International Symposium for Testing and Failure Analysis*, Santa Clara, 2001.
19. M. Xie, Z. Qian, M. Pacheco, Z. Wang, R. Dias and V. Talanov, “Fault Isolation of Open Defects Using Space Domain Reflectometry,” in *Proceedings from the 38th International Symposium for Testing and Failure Analysis*, Phoenix, 2012.
20. J. Gaudestad, D. Nuez and P. Tan, “Short Localization in 2.5D Microchip with Interposer Using Magnetic Current Imaging,” in *Proceedings from the 40th International Symposi-um for Testing and Failure Analysis*, Houston, 2014.
21. [21] M. Kögel, F. Altmann, S. Tismer and S. Brand, “Magnetic field and current density imaging using off-line lock-in analysis,” *Microelectronics Reliability,* vol. 64, p. 346, 2016.
22. [22] M. J. Turner, N. Langellier, R. Bainbridge, D. Walters, S. Meesala, T. M. Babinec, P. Kehayias, A. Yacoby, E. Hu, M. Loncar, R. L. Walsworth and E. V. Levine, “Magnetic Field Fingerprinting of Integrated-Circuit Activity with a Quantum Diamond Microscope,” *PHYSICAL REVIEW APPLIED,* vol. 14, no. 1, p. 014097, 2020.
23. E. V. Levine, M. J. Turner, N. Langellier, T. M. Babinec, M. Lončar and R. L. Walsworth, “Backside Integrated Circuit Magnetic Field Imaging with a Quantum Diamond Microscope,” in *46th International Symposium for Testing and Failure Analysis*, 2020.

24. A. Orozco, N. E. Gagliolo, C. Rowlett, E. Wong, A. Moghe, J. Gaudestad, V. Ta-lanov, A. Jeffers, K. Torkashvan, F. C. Wellstood, S. Dobritz, M. Boettcher, A. B. Cawthorne and F. Infante, "3D IC/Stacked Device Fault Isolation Using 3D Magnetic Field Imaging," in *Conference Proceedings from the 40th International Symposium for Testing and Failure Analysis*, Houston, 2014.
25. J. Phang, D. Chan, M. Palaniappanl, J. M. Chin, B. Davis, M. Bruce, J. Wilcox, G. Gilfeather, C. M. Chua, L. Koh, H. Ng and S. Tan, "A Review of Laser Induced Techniques for Microelectronic Failure Analysis," in *The 11th edition of the IEEE International Symposium on the Physical and Failure Analysis of Integrated Circuits (IPFA)*, Taiwan, 2004.
26. [26] Z. Song and L. Safran, "Diagnostic Technique Selection For SRAM Logic Type Failures," *Electronic Device Failure Analysis,* vol. 20, no. 2, pp. 18-24, 2018.
27. N. K. a. C. L. Chiang, "Analysis of Product Hot Electron Problem by Gated Emission Microscopy," in *IEEE 24th Ann. Int. Reliab.Phys. Symp. (IRPS)*, 1986.
28. J. Phang, D. Chan, S. Tan, W. Len, K. Yim, L. Koh, C.M.Chua and L. Balk, "A Review of Near Infrared Photon Emission Microscopy and Spectroscopy," in *Proc. Int. Symp. Phys. Fail. Anal. Integr.Circuits (IPFA)*, 2005.
29. D. R. Bockelman, S. Chen and B. Obradovic, "Infrared Emission-based Static Logic State Imaging on Advanced Silicon Technologies," in *Conference Proceedings from the 28th International Symposium for Testing and Failure Analysis*, Phoenix, 2002.
30. Y. Y. Chew, K. H. Siek and W. M. Yee, "Novel Backside Sample Preparation Processes for Advanced CMOS Integrated Circuits Failure Analysis," in *Proceedings of 7th IPFA*, Singapore, 1999.
31. E. I., C. Jr., P. Tangyunyong and D. L. Barton, "Backside Localization of Open and Shorted IC Interconnections," in *IEEE 98CH36173. 36th Annual International Reliability Physics Symposium*, Reno, 1998.
32. E. Cole, P. T. Jr., D. Benson and D. Barton, "TIVA and SEI developments for enhanced front and backside interconnection failure analysis," *Microelectronics Reliability,* vol. 39, pp. 991–996, 1999.
33. K. Dickson, G. Lange, K. Erington and J. Ybarra, "Electron Beam Absorbed Current as a Means of Locating Metal Defectivity on 45nm SOI Technology," in *Conference Proceedings from the 36th International Symposium for Testing and Failure Analysis*, Dallas, 2010.
34. R.-L. Chiu, H. Zhang, W.-S. Chung, M. Cherng and X. Liu, "Beam-Based Localization Techniques for 0.18um IC Failure Analysis after Reliability Test," in *Conference Proceedings from the 28th International Symposium for Testing and Failure Analysis*, Phoenix, 2002.
35. [35] C. Boit, R. Schlangen, A. Glowacki, U. Kindereit, T. Kiyan, U. Kerst, T. Lundquist, S. Kasapi and H. Suzuki, "Physical IC debug – backside approach and nanoscale challenge," *Adv. Radio Sci.,* vol. 6, pp. 265-272, 2008.
36. C. Boit, U.Kerst, R. Schlangen, A. Kabakow, E. L. Roya, T. Lundquista and S. Pauthner, "Impact of back side circuit edit on active device performance in bulk silicon ICs," in *International Test Conference*, 2005.
37. J. J. Huening, P. Joshi, H. J. Ryu, W.-h. Chuang, D. Xu, M. L. Oh, S. Zhao and T. Tong, "High spatial and energy resolution fault isolation by electron beam probing for advanced technology nodes," in *46th International Symposium for Testing and Failure Analysis*, 2020.
38. F. Altmann, J. Beyersdorfer, J. Schischka, M. Krause, G. Franz and L. Kwakman, "Cross section analysis of Cu filled TSVs based on high throughput plasma-FIB milling," in *Conference Proceedings from the 38th International Symposium for Testing and Failure Analysis (ISTFA, ASM International)*, Phoenix, 2012.
39. L. Kwakman, M. Straw, G. Coustillier, M. Sentis, J. Beyersdorfer, J. Schischka, F. Naumann and F. Altmann, "Sample Preparation Strategies for Fast and Effective Failure Analysis of 3D Devices," in *Proceedings from the 39th International Symposium for Testing and Failure Analysis (ISTFA, ASM International)*, San Jose, 2013.
40. [40] E. Principe, N. Asadizanjani, D. Forte, M. Tehranipoor, R. Chivas, M. DiBattista and S. Silverman, "Plasma FIB deprocessing of integrated circuits from the backside," *Electronic Device Failure Analysis,* vol. 19, no. 4, pp. 36-44, 2017.

41. H. Z. a. M. X. H. Xu, “A case study of package delamination with combination of EDX and MiniSIMS,” in *Conference Proceedings from the 15th Electronic Components and Technology Conference (ECTC)*, 2013.
42. [42] A. Centrone, “Infrared Imaging and Spectroscopy Beyond the Diffraction Limit,” *Annual Review of Analytical Chemistry,* vol. 8, p. 101, 2015.
43. [43] C. B. P. A. Dazzi, “AFM-IR: Technology and Applications in Nanoscale Infrared Spectroscopy and Chemical Imaging,” *Chemical Reviews,* vol. 117, no. 7, p. 5146, 2017.
44. P. V. d. Heide, X-ray Photoelectron Spectroscopy: An introduction to Principles and Practices 1st edition, Wiley, 2011.
45. P. V. d. Heide, Secondary Ion Mass Spectrometry: An Introduction to Principles and Practices 1st edition, Wiley, 2014.
46. [46] M. H. Lu, D. Y. Shih, P. Lauro, C. Goldsmith and D. W. Henderson, “Effect of Sn grain orientation on electromigration degradation mechanism in high Sn-based Pb-free solders,” *Applied Physics Letter,* vol. 92, p. 211909, 2008.
47. [47] Y. Wang and P. S. Ho, “Mode II electromigration failure mechanism in Sn-based Pb-free solder joints with Ni under-bump metallization,” *Applied Physics Letter,* vol. 103, p. 121909, 2013.
48. P. Liu, A. Overson and D. Goyal, “Key parameters for fast Ni dissolution during electromigration of Sn0.7Cu solder joint,” in *Conference Proceedings from the 65th Electronic Components and Technology Conference (ECTC)*, 2015.
49. C. Wu, T. Jiang, J. IM, K. M. Liechti, R. Huang and a. P. S. Ho, “Material characterization and failure analysis of through-silicon vias.,” in *Proceedings from the 21st International Symposium on the Physical and Failure Analysis of Integrated Cir-cuits (IPFA)*, 2014.
50. [50] S. F. Lin, C. H. Chen and C. Y. Lo, “Near-infrared imaging system for nondestructive inspection of micro-crack in wafer through dicing tape,” *Applied Optics,* vol. 54, no. 28, pp. E123-8, 2015.
51. M. H. Ettenberg and D. Malchow, “InGaAs SWIR Imagers Optimize Semiconductor Inspection,” *Photonic Tech Briefs,* vol. July 1, 2007.
52. [52] M. Matysiak, J. P. Parry, F. Albri, J. G. Crowder, N. Jones, K. Jonas, N. Weston, D. P. Hand and J. D. Shephard, “Infrared confocal imaging for inspection of flaws in yttria-stabilized tetragonal zirconia polycrystal (Y-TZP),” *Measurement Science and Technology,* vol. 22, p. 125502, 2011.
53. [53] W. Lawton and J. Barrett, “Characterisation of chip-on-board and flip chip packaging technologies by acoustic microscopy,” *Microelectronics. Reliability,* vol. 36, p. 1803, 1996.
54. [54] J. E. Semmens, “Flip chips and acoustic micro imaging: An overview of past applications, present status, and roadmap for the future,” *Microelectronics. Reliability,* vol. 40, p. 1539, 2000.
55. D. A. Hutt, D. P. Webb, K. C. Hung, C. W. Tang, P. P. Conway, D. C. Whalley and Y. C. Chan, “Scanning acoustic microscopy investigation of engineered flip-chip delamination,” in *Proceedings from the 2000 IEEE/CPMT Int’l Electronics Manufacturing Technology Symposium*, 2000.
56. [56] J. E. Semmens and L. W. Kessler, “Application of Acoustic Frequency Domain Imaging for the Evaluation of Advanced Micro Electronic Packages,” *Microelectronics. Reliability,* vol. 42, p. 1735, 2002.
57. [57] B. T. Khuri-Yakub, “Scanning acoustic microscopy,” *Ultrasonics,* vol. 31, no. 5, p. 361, 1993.
58. G. A. D. Briggs and O. V. Kolosov, Acoustic Microscopy 2nd edition, Oxford University, 2009.
59. T. M. Moore and C. D. Hartfield, “Trends in nondestructive imaging of IC packages,” in *Proceedings from AIP*, 1998.
60. A. J. Komrowski, L. A. Curiel, Q. Nguyen, D. J. D. Sullivan and L. Logan-Willams, “Backside Application of Acoustic Micro Imaging (AMI) on Plastic Ball Grid Array (PBGA) and Plastic Quad Flat Pack (PQFP) Packages,” in *Proceedings from the 30th International Symposium for Testing and Failure Analysis*, 2004.
61. [61] L. Angrisani, L. Bechou, D. Dallet, P. Daponte and Y. Ousten, “Detection and location of defects in electronic devices by means of scanning ultrasonic microscopy and the wavelet transform,” *Measurement,* vol. 31, p. 77, 2002.

62. J. Sigmund and M. Kearney, "TAMI Analysis of Flip Chip Packages," *Advanced Packaging,* no. July/August, 1998.
63. [63] S. Haque, G. Lu, J. Goings and J. Sigmund, "Characterization of interfacial thermal resistance by acoustic micrography imaging," *Microelectronics. Reliability,* vol. 40, p. 465, 2000.
64. T. S. Leng, J. C. P. McKeon and H. S. Jang, "Correlation study of delamination gap detection capability of SAM and cross section/SEM analysis," in *Proceedings from the 32th International Symposium for Testing and Failure Analysis*, Austin, 2006.
65. Y. Li, L. Hu, G. Li, R. Dias and D. Goyal, "High Resolution C-Mode Scanning Acoustic Microscope Techniques for the Failure Analysis of Microelectronic Packages," in *Proceedings from the 39th International Symposium for Testing and Failure Analysis (ISTFA, ASM International)*, San Jose, 2013.
66. [66] S. Brand, A. Lapadatu, T. Djuric, P. Czurratis, J. Schischka and M. Petzold, "Scanning acoustic gigahertz microscopy for metrology applications in three-dimensional integration technologies," *Journal of Micro/Nanolithography, MEMS, and MOEMS,* vol. 13, no. 1, p. 011207, 2014.
67. J. Perraud, S. Enouz-Vedrenne, J. Clement and A. Grivon, "Development of Advanced Non Destructive Techniques for Failure Analysis of PCBs and PCBAs," in *Proceedings from the 38th In-ternational Symposium for Testing and Failure Analysis (ISTFA, ASM International),* Phoenix, 2012.
68. [68] M. Feser, J. Gelb, H. Chang, H. Cui, F. Duewer, S. H. Lau, A. Tkachuk and W. Yun, "Submicron resolution CT for failure analysis and process development," *Measurement Science and Technology,* vol. 19, p. 094001, 2008.
69. [69] Y. Li, J. S. Moore, B. Pathangey, R. C. Dias and D. Goyal, "Lead-Free Solder Joint Void Evolution During Multiple Subsequent High-Temperature Reflows," *IEEE Transactions on Device and Materials Reliability,* vol. 12, no. 2, p. 494, 2012.
70. [70] Y. Li, P. K. M. Srinath and D. Goyal, "A Review of Failure Analysis Methods for Advanced 3D Microelectronic Packages," *Journal of Electronic Materials,* vol. 45, p. 116, 2016.
71. [71] Y. Li, R. Panat, B. Li, R. Mulligan, P. K. M. Srinath and A. Raman, "The Application of Two-Dimensional X-ray Hot Stage in Flip Chip Package Failure Analysis," *IEEE Transac-tions on Device and Materials Reliability,* vol. 11, no. 1, p. 141, 2011.
72. Y. Li, Y.Cai, M.Pacheco, R. C. Dias and D. Goyal, "Non Destructive Failure Analysis of 3D Electronic Packages Using Both Electro Optical Terahertz Pulse Reflectometry and 3D X-ray Computed Tomography," in *Proceedings from the 38th In-ternational Symposium for Testing and Failure Analysis (ISTFA, ASM International)*, Phoenix, 2012.
73. M. Pacheco and D. Goyal, "Detection and characterization of defects in microelectronic packages and boards by means of high-resolution X-ray computed tomography (CT)," in *IEEE 61st Electronic Components and Technology Conference (ECTC),* Lake Buena Vista, 2011.
74. C. Schmidt, C. Hartfield, S. T. Kelly, L. England and S. Kannan, "Nanoscale 3D X-ray Microscopy for High Density Multi-Chip Packaging FA," in *Conference Proceed-ings from the 44th International Symposium for Testing and Failure Analysis (ISTFA, ASM International),* 2018.
75. Y. Li, M. Pacheco, D. Goyal, J. W. Elmer, H. D. Barth and D. Parkinson, "High Resolution and Fast Throughput-time X-ray Computed Tomography for Semiconductor Packaging Applications," in *Conference Proceedings from the 64th Electronic Components and Technology Conference (ECTC)*, 2014.
76. E. Zschech, M. Löffler, J. Gluch and M. J. Wolf, "Micoro and Nano X-ray Tomorgraphy of 3D IC Stacks," in *Materials Research Society (MRS) Spring*, Phoenix, 2016.
77. W. Yun, M. Feser, J. Gelb and L. Hunter, "Multiscale 3-D X-ray Imaging for 3-D IC," in *International Conference on Frontiers of Characterization and Metrology for Nanoelectronics*, Gaithersburg, 2013.

78. J. W. Elmer, Y. Li, H. D. Barth, D. Y. Parkinson, M. Pacheco and D. Goyal, “Synchrotron radiation microtomography for large area 3D imaging of multilevel microelectronic packages,” *Journal of Electronic Materials,* p. 4421, 2014.

Corrosion Mechanisms of Copper and Gold Ball Bonds in Semiconductor Packages

A Unification of Structure-Based Inference and Electrochemical Investigation

Wentao Qin

Abstract Ball bonding is the most widely used interconnection method in microelectronic packages. It has enabled many modern technologies including medical implants, aerospace, automobiles and Internet of Things. In the automotive industry, driving automation and advanced driver assistance systems motivated mainly by safety enhancement are gaining traction. The reliabilities of these technologies necessitate those of the underlying ball bonds. This chapter provides interpretation of the mechanisms of corrosion that causes reliability failures of Cu and Au ball bonds, by unifying approaches based on microstructure characterization and electrochemical investigation. The corrosion of Cu ball bonds starts with pitting of the most Cu-rich layer (MCRL) under the chlorinated water layer, evolves into crevice corrosion, and can be assisted by stress corrosion cracking. In the MCRL, Al is preferentially oxidized, while the Cu atoms remain largely immune and coalesce to form nanoparticles. Four methods to address the corrosion are presented. Limited data indicate the same corrosion mechanisms for Au ball bonds.

Keywords Automotive · Driving automation · Advanced driver assistance systems (ADAS) · Wire bonding · Copper ball bond · Gold ball bond · Interface · Intermetallic · Microstructures · Chlorine · Chloride ions · Corrosion · Electrochemical analyses · Corrosion potential · Pourbaix diagram · Crevice corrosion · Hydrogen embrittlement · Stress corrosion cracking (SCC) · Dealloying

1 Background and Motivation

1.1 *The Importance of Ball Bonds*

In microelectronics, wire bonding provides an electrical connection between the semiconductor chip and the metal leadframe, and between semiconductor chips.

W. Qin (🖂)
Phoenix Product Analysis Lab, Onsemi, Phoenix, AZ 85008, USA
e-mail: Wentao.Qin@onsemi.com

Y. Li and H. Shi (eds.), *Advanced Driver Assistance Systems and Autonomous Vehicles*,
https://doi.org/10.1007/978-981-19-5053-7_15

Ball bonding and wedge bonding are the two major variations of wire bonding, with ball bonding being the larger portion. The bond is formed with the free air ball (FAB) pressed against the bond pad and ultrasonic energy applied to the FAB [1, 2]. The ultrasonic energy causes the FAB to break through processing residue on the pad and the surface oxides of the Cu FAB and the pad, so that intimate Cu-to-Al contact is achieved. Figure 1 presents a cross-sectional scanning electron microscopy (SEM) image of a Cu ball bond on an Al alloy pad. Structures related to the corrosion, namely Cu ball, Al alloy pad, the epoxy mold compound (EMC), and Al smear, are labeled in the image. After the ball bond is formed, the wire is looped to a bond pad of either a leadframe or another chip where the wedge bond is made.

The wire bonding technology was developed in the 1950s. Over the years, the technology has been constantly reinvented, especially for Cu wire bonding, and has remained the workhorse in the packaging industry. The global semiconductor wire bonding market had reached $12.15 billion in 2021 [3]. Wire bonding accounts for about 80% of the entire semiconductor products assembled in 2014 [4]. About 70% of wire-bond packages use Cu wires as of 2017 [5], as Cu has taken the place of Au as the number one bond wire material, due to its cost competitiveness [1, 6–8], while the remaining portion is shared by Au or Ag wires.

The advancements of integrated circuits and packaging have enabled many modern technologies. As the speeds to acquire and process complex information and to control machines are dramatically improved, increasingly more devices and machines are being made "smart" and "connected". The dawn of a new era of Internet of Things has already begun. The automotive sector is no exception. As summarized by Car and Driver, a new white paper from Deloitte notes that as of 2017, electronics systems powered by semiconductor chips comprised 40% of the cost of a new car. That was up from 18% in 2000, 20% in 2007, and is projected to reach 45% in 2030 [9]. Currently, the average number of packaged semiconductor chips per vehicle is about 2000 for electric vehicles, and 1000 for non-electric vehicles as of November 2021 [10]. The number of packaged semiconductor chips reached 6000 to 8000 for an Audi A8, and 75% to 90% of innovations in modern cars rely

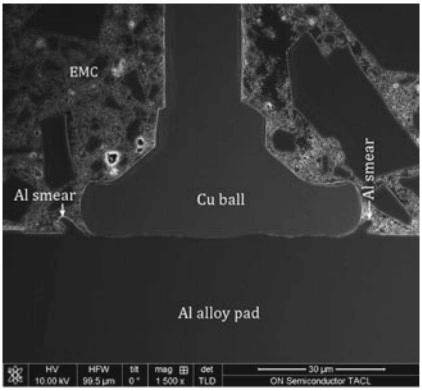


Fig. 1 Cross-sectional SEM image of a Cu ball bond on Al alloy pad. Structures related to the corrosion, namely Cu ball, Al alloy pad, EMC, and Al smear, are labeled in the image

on packaged semiconductor chips as of 2012 [11]. Among the innovations, those for safety improvement are of the utmost importance. Government data showed driver errors accounted for 94% of crashes as of 2021 [12]. The correlation calls for automation of the vehicle for safety enhancement. Other motivations for vehicle automation include economic benefits, productivity improvements, traffic congestion reduction, environmental gains, and greater independence for people with disabilities.

There are six levels of automation, from 0 to 5, with the independence of the vehicle on operation control progressively increases higher with the level. The advanced driver assistance system (ADAS) enters the automation hierarchy from level 1. At level 5, the vehicle is fully capable of operation itself without any human intervention, and the vehicle becomes an autonomous vehicle (AV). Overall the vehicle automation involves dynamically sensing and processing information, and subsequent implementation of vehicle control. The process is enabled by packaged semiconductor chips, with the sensing portion accomplished with one type of chips that serve as sensors.

1.2 Corrosion of Cu Ball Bonds and Its Impact

Reliability failures of the packaged chips can cause the failure of the automation, and the criticality is further escalated when safety is concerned. For example, ADAS applications such as driver drowsiness detection, blind spot detection, lane departure warning/correction, pedestrian detection/avoidance, automatic emergency braking, and traffic sign recognition are all safety–critical [13]. It is imperative to understand the mechanisms of reliability failures in order to address the issue. Corrosion is known to be a cause of reliability failures of mechanical, electric, and electronic parts, including Cu and Au ball bonds, for automotives. Cu ball bonds were first introduced to consumer products where the reliability was less demanding and later used in automotive electronics which have imposed more stringent reliability specifications in harsh environments with high relative humidity and high temperature [14, 15]. A major reliability failure of Cu ball bonds is due to corrosion along the bond interface, across which intermetallic compounds (IMC) form [16, 17].

In order to achieve sufficient adhesion and electrical conduction between the Cu wire and the pad, intimate Cu-to-Al contact must be made and chemical bonds must form across the interface, which, however, inevitably causes the formation of IMCs. Subsequently, a preferential corrosion of the most Cu-rich layer (MCRL) takes place in the presence of a thin-film electrolyte that contains mobile Cl^- ions from the EMC [18–20]. Sources of mobile Cl^- ions include the following.

1. Cl-containing impurities from the synthetic components for the EMC [21],
2. chlorinated fire-retardants in the EMC [22], and
3. particles of carbon-chlorine polymer parts of the equipment in the assembly process. These particles can be embedded in the EMC [21].

Moisture absorbed in the EMC and sufficiently high temperature can trigger the degradation of the Cl-containing materials in the EMC and the EMC, which can lead to the generation and enhanced transport of mobile Cl^- ions.

The corrosion proceeds along the MCRL/Cu interface, which elevates the interfacial resistance. A crack can propagate along this interface. A large enough increase of the electrical resistance is manifested as a reliability failure of the ball bond [16, 17]. Figure 2 presents TEM brightfield (BF) images of Cu ball bonds on Al(1%Si) alloy pads that had passed and failed an unbiased highly accelerated stress test (uHAST) [23]. Figure 2a and b reveal the IMCs across the passing bond, and the corrosion products across the failing bond, respectively. The stack across the passing bond is (in the stack specification, q and d should have been the Greek letters of theta and eta, respectively)

$$\text{AlSi bond pad}/CuAl_2(q)/CuAl(h)/Cu_3Al_2(d)/\text{Cu ball.}$$

Hereon "Cu ball" is used to refer to the deformed Cu FAB on the bond pad after the ball bonding. The Greek letter for each IMC is written only in the first occurrence of the IMC. The four interfaces in the stack are labeled I – IV progressively upward as shown in Fig. 2a, for brevity. Figure 2 reveals the MCRL of Cu_3Al_2 had been corroded in the failing bond. The MCRL can be Cu_3Al_2 in some cases such as this one, and $Cu_9Al_4(\gamma_2)$ in other cases which will be shown.

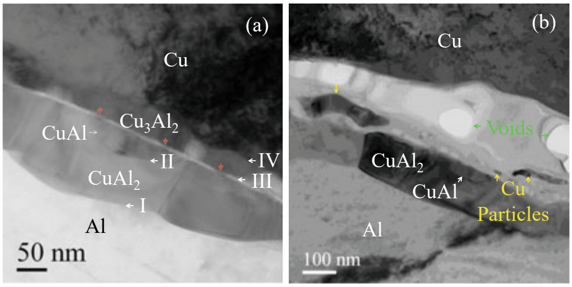


Fig. 2 TEM brightfield (BF) images of Cu ball bond interfaces that had passed (**a**) and failed (**b**) the uHAST [23]. The IMCs across the interfaces, and the corrosion products are both revealed. The four interfaces in the stack of AlSi bond pad/$CuAl_2$/CuAl/Cu_3Al_2/Cu ball across the bond are labeled I–IV progressively upward. Three Cu particles are marked with yellow arrows, and two voids with green arrows. The vertical semitransparent red arrows in **a** mark diffusion paths across interface III where chemical bonds between CuAl and Cu_3Al_2 have formed

1.3 Structure-Based Inference of Corrosion Mechanisms and Its Limitation

The corrosion products were always found under the Cu ball [18, 20, 23], which indicated that the MCRL had already been corroded ahead of other IMCs. This is consistent with an earlier study of Al-Cu model alloys that showed the susceptibility to pitting increased with the Cu content of the alloy [24]. Without sufficient perspective of electrochemistry, however, the structure-based inference of the corrosion mechanisms encountered difficulties. For example, regarding the observation of $CuAl_2$, instead of the MCRL of either Cu_3Al_2 or Cu_9Al_4, surviving the reliability test, it was postulated that $CuAl_2$ with the highest Al content should possess the highest corrosion resistance that would be attributed to the passivation of Al [18, 20]; on the other hand, however, a counterargument was presented that $CuAl_2$ should be the most reactive IMC since Al is more reactive than Cu [18]. In the work described in reference [23], the corrosion products were characterized, and the corrosion starting at the MCRL/Cu interface inferred; however, the following information was absent: the pitting mechanisms, thermodynamics of the corrosion at room temperature and high temperature, and the form of the corrosion in its advanced stage. To overcome the difficulty, electrochemical investigation is needed. It was realized that the Cu wire, the IMCs, and the Al bond pads could be treated as the basic unit components of an electrochemical system with each of them having its own corrosion potential, which, if known, could help explain the preferential corrosion of the MCRL [20]. In the corrosion of a metal immersed in an aqueous environment, two electrodes, a cathode and an anode, are formed spontaneously. Oxidation and reduction reactions proceed on the anode and the cathode, respectively, until equilibrium is reached. The net current, which is the difference between the anodic current and the cathodic current, equals zero. This potential is the corrosion potential of the metal under this specific set of conditions. Equivalently stated, in equilibrium with the environment, the metal acquires a potential at which the net current is zero, and the potential is the corrosion potential under the current set of conditions. The corrosion potentials of the Cu-Al IMCs were hypothesized to be between those of Cu and Al, but not determined [25].

Overall with structure-based inferences, the microstructures of the electrodes and the location of the corrosion can be well-characterized; however, the electrochemical fundamentals of the corrosion, including thermodynamics and kinetics, such as stable species under different conditions, galvanic corrosion of the coupled phases in the thin-film stack, anodic current density dependence on the cathode-to-anode area ratio, pitting, crevice corrosion, and stress corrosion cracking (SCC), which are all beyond, but still underlying, the structure information, are missing. Therefore, electrochemical investigation needs to be incorporated in the study of the corrosion mechanisms.

1.4 Electrochemical Investigation of Corrosion Mechanisms and Its Limitation

The long-anticipated determination of corrosion potentials of the Cu-Al IMCs in chlorinated-electrolytes has recently been completed [26, 27]. The study by Wu involved stacking discrete bulk ingots of Cu-Al IMCs and subsequent immersion of the stack in electrolyte for a mass loss experiment [27]. The stack, however, differed from the IMC thin-film stack in a Cu ball bond in a few aspects that concern the corrosion. For example, the ingot stack was built by physically stacking the ingots; in contrast, the thin-film stack in the Cu ball bond forms through solid-state reactions that take place during and after the ball bonding. The solid-state reactions create significantly unequal interfacial resistances. Since interface III has captured the processing residue and surface oxides of the Cu FAB and the Al alloy pad, it is more resistive than the other three interfaces. The MCRL above interface III is therefore less cathodically protected by the underlying IMC layer; on the other hand, the MCRL is anodically corroded due to its galvanic coupling with the superjacent phase, which is the Cu ball. Such a preferential corrosion of one component layer/phase due to an elevated interfacial resistance is absent in the ingot stack. Secondly, the electrolyte is present in the form of a thin water-film covering the Cu ball bond. The Ohmic resistance of the thin-film electrolyte is high, which limits the spatial range of the galvanic corrosion of the MCRL-Cu couple to the vicinity of the interface. This makes the corrosion hard to be detected until the bond resistance has either exceeded a preset limit or caused device failure. The corrosion products are generally visualized with SEM or TEM at magnifications on the orders of 10kx and higher [23, 28]. In contrast, the electrolyte for the mass loss experiment was in the form of bulk solution to immerse the entirety of the ingot stack. The corrosion was not limited to the size scale of hundreds of nanometers from the Al/$CuAl_2$ and $CuAl_2$/MCRL interfaces; instead, it took place over a much larger size range on the order of millimeters from the interface and was visible under optical microscope at a magnification on the order of 10x [27]. Third, the Cu ball bond is encapsulated by EMC. The material heterogeneity of the package can cause significant stress at the bond interface. This stress and the lattice strain along the thin-film stack interfaces facilitate the corrosion, but neither of the factors is present in the ingot stack. The differences will be further elaborated in Sect. 3.2. “Ingot stack to simulate Cu ball bonds and the limitations.” As a result, in the ingot stack it was the ingots of Al and $CuAl_2$ that exhibited the most severe corrosion, an outcome that is entirely different from that seen in Cu ball bonds, where it is the MCRL, which can be Cu_3Al_2 or Cu_9Al_4, that is corroded the most. Overall the use of bulk samples immersed in electrolyte could not incorporate sufficient elements pertaining to the corrosion of Cu ball bonds, such as the interfacial resistances, the stress, the chemical bonding between adjacent layers, lattice strain, and the geometry of the Cu ball bond system. With these elements influencing the corrosion, the electrochemical investigation needs to incorporate microstructure information for amendments.

1.5 The Need to Unify the Two Approaches

These two approaches, each with its value and limitation, are complementary. Their unification is therefore necessary, to gain new insights and develop a more comprehensive understanding of the corrosion mechanisms. For example, in the initial stage of corrosion, the passivation of the MCRL is mainly due to that of Al, and the passivity is lower than that of Al because the passivation film is more defective than that of Al. Both water and Cl^- ions are Lewis acids, tending to bond with Al^{3+} ions, which drives the oxidation of the Al in the MCRL. Secondly, pitting is initiated with Cl^- ions from the EMC, and followed by crevice corrosion along the MCRL/Cu ball interface, and the crevice corrosion can be assisted by SCC. The corrosion has been addressed accordingly for reliability improvement. Effective methods will be described in Sect. 7, "Addressing the corrosion," and they include decreasing the concentration of the extractable Cl^- ions in the EMC, and post-bond heating to relieve strain along the bond interface created by the ball bonding and to increase the areal fraction of the MCRL.

1.6 Beyond Cu Ball Bonds—Au Ball Bonds

The understanding of the mechanisms very likely benefit beyond Cu ball bonds, since a similar reliability issue is known to exist in Au ball bonds. Both types of ball bonds exhibit a preferential oxidation of the most Au-rich layer (MARL) [29] and MCRL [23]. The knowing of whether these two types of ball bonds are subjected to the same corrosion mechanisms is of practical value.

1.7 Sources of Data

As stated in the preceding Sect. 1.6, this chapter involves unification of the two approaches that are to a great extent independently based on microstructure characterization and electrochemical analyses. The data used in this chapter therefore include those acquired via these two approaches, from experiments conducted for this chapter, as well as from literature, including papers published by the author of this chapter. All the figures containing electrochemical data are from literature. All the figures generated for this chapter are to present microstructure data. The figures from other authors' papers and containing microstructure data represent their types that are consistent with the microstructure data acquired for this chapter. Sources of all the figures are listed in Table 1. Since only reliability tests and microstructure characterization were performed for this chapter, their materials and methods are presented in the following section.

Table 1 Sources of figures in this paper

Figures used in this paper		Sources
Microstructure data	Electrochemical data	
Figures 1,7, 8, 10, 11, 12 and 19		Figures generated for this chapter
Figures 2, 6, 9, 23 and 24		Figures from the author's papers
Figures 5, 13, 14, 20, 21, 22, 25 and 26	Figures 3, 4, 15, 16, 17 and 18	Figures from other authors' papers

"The author" refers to the author of this chapter

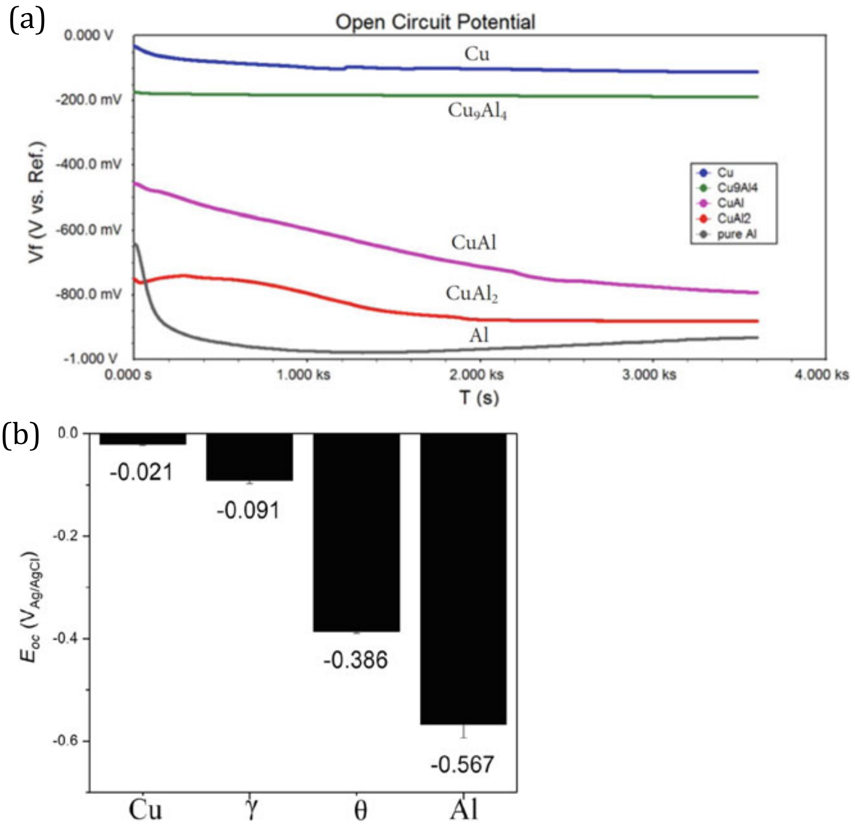


Fig. 3 **a** Open-circuit potentials versus time of Cu, Al, and Cu–Al IMCs in a 25 ppm chloride solution at pH = 6 [26], and **b** open-circuit potentials of Cu, Al, and Cu–Al IMCs in a 20 ppm chloride solution at pH = 6 [27]. Open-circuit potential is also called corrosion potential. In an electrochemical cell, the conduction electron flow is from a lower corrosion potential to a higher corrosion potential

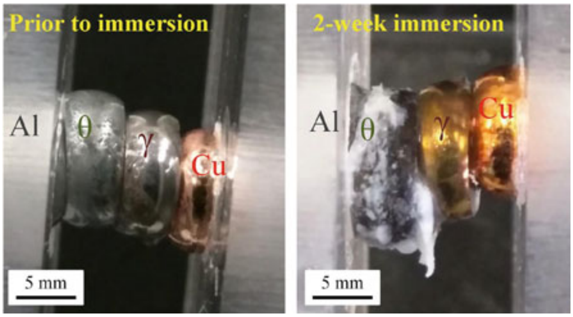


Fig. 4 Ingot stack of, Al/$CuAl_2$/Cu_9Al_4/Cu before and after 2 weeks of immersion in a near-neutral 20 ppm NaCl solution [27]. Being the most anodic in the stack, $CuAl_2$ and Al exhibited the highest corrosion rates, with visible corrosion product covering the entirety of the $CuAl_2$ ingot and likely that of Al ingot

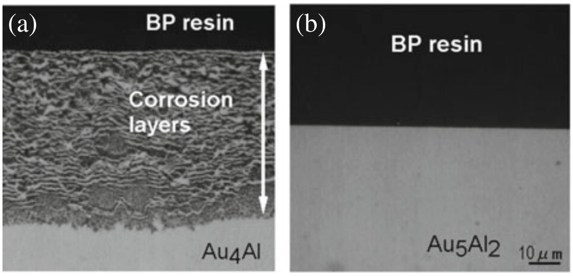


Fig. 5 Micrographs of Au-Al IMC heated at 498 K for 500 h in Bi-phenyl epoxy resin [29]. **a** Au_4Al, and **b** Au_5Al_2. The figure indicates the corrosion rate of Au_4Al, which is Au-richer compared with Au_5Al_2, is higher than that of Au_5Al_2

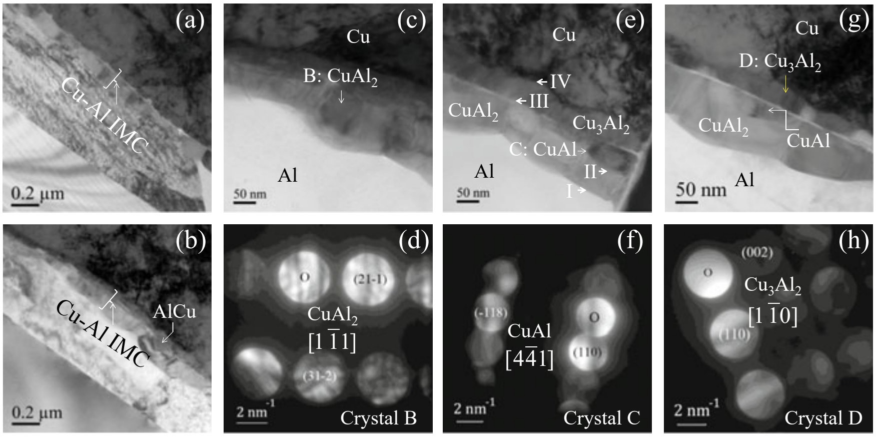


Fig. 6 TEM BF images (**a**–**c**, **e**, **g**) of IMC crystals across the bond that passed the uHAST, and electron diffraction patterns (**d**, **f**, **h**) of the IMC crystals labeled "B," "C," and "D" in the TEM images [23]

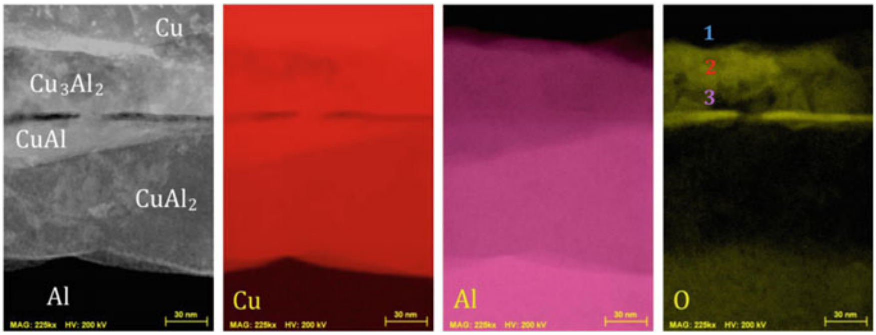


Fig. 7 STEM DF image of the bond that passed the uHAST, and the associated elemental maps. The figure shows the IMC stack and elemental distributions across the bond

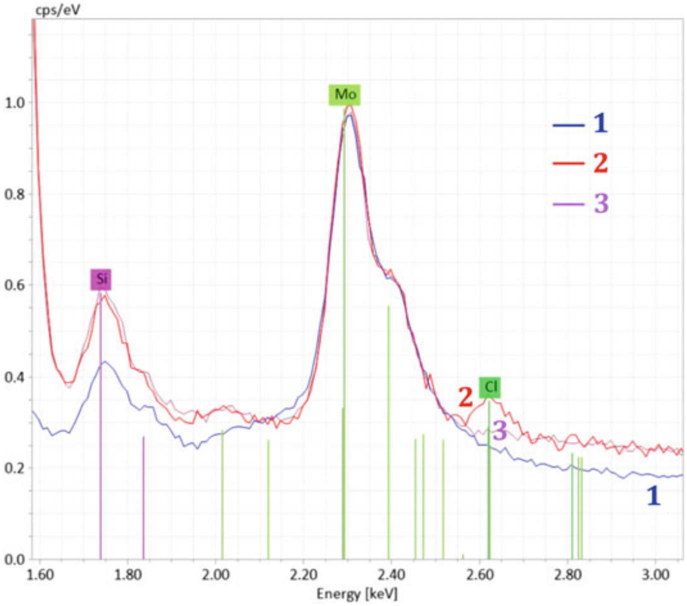


Fig. 8 EDS spectra of areas 1–3 marked in Fig. 7. The figure indicates Cl is present in the top portion of Cu_3Al_2 and suggests corrosion started along the Cu_3Al_2/Cu interface and propagated down to consume Cu_3Al_2, a subject that will be further explored in the forthcoming discussion. The Si and Mo peaks are artifact from the fluorescence of the Si epi and Si-oxide in the sample, and Mo sample grid, respectively

2 Materials and Methods

Corrosion of the ball bond is usually discovered from the reliability failure that either is detected through a reliability test or occurs while the packaged electronic device

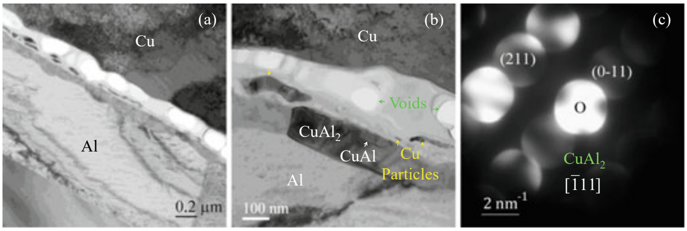


Fig. 9 TEM BF images **a**, **b** of the first bond interface that failed the uHAST, and electron diffraction pattern **c** of the $CuAl_2$ crystal labeled in **b** [23]. The figure shows the corrosion products across the bond

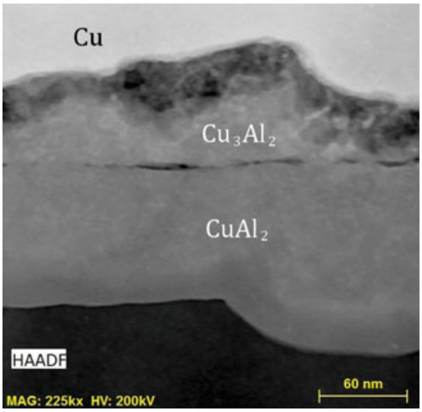


Fig. 10 HAADF image of the second bond interface that had failed the uHAST. The crevice corrosion had taken place in Cu_3Al_2 along the Cu_3Al_2/Cu interface. There is no air gap shown in the image, which indicates voiding in the corrosion products is less severe than that shown in Fig. 9

is in use in a customer application. The failure can be manifested as an increase of electrical resistance of the bond interface that is large enough to either have failed the device or be considered a failing value [23]. The failure can also exhibit as a significant decrease of the shear strength of the bond [30]. The reliability test to be performed was an unbiased highly accelerated stress test (uHAST) at 130 °C and 85% relative humidity, and lasting for 288 h. Each of the packaged die contains a pair of Cu wires that were ball-bonded to two AlSi (1%) bond pads. The bond pads were connected to the base and emitter of a bipolar junction transistor (BJT), while the other end of each of the wires was wedge-bonded to an external lead. The chip with the pair of bond wires was plastic-molded into a package. The extractable Cl^- ion concentration of the EMC was specified to be lower than 20 ppm. The electrical resistance between the pair of external leads of the package was measured immediately after assembly and recorded as the resistance at time zero. The package was next subjected to the uHAST. After the uHAST, the electrical resistance between the same pair of external leads of the package was measured. Compared with the measurement at time zero, a resistance measurement after the uHAST that had increased by more than 20% (of the resistance at time zero) is classified as a failing value, and a passing value otherwise. One package that passed and four packages that failed the uHAST were chosen for

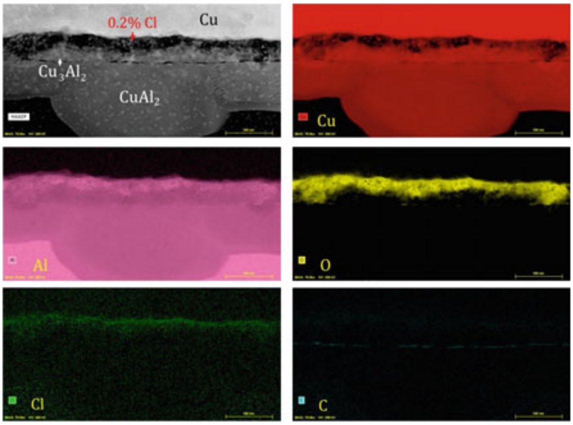


Fig. 11 HAADF image of another segment of the second bond interface that had failed the uHAST, and the associated elemental maps. The figure shows that the corrosion was in the form of crevice corrosion that took place in Cu_3Al_2, along the Cu_3Al_2/Cu ball interface, and simultaneously proceeded downward to consume Cu_3Al_2. The figure also reveals the separation of the Cu nanoparticles, three of which are marked by the green arrows) from the residual Cu_3Al_2, which caused the passivation to be weakened during the crevice corrosion of the Cu_3Al_2. The small white particles (three of which are marked by the white arrows) in the $CuAl_2$ area are Cu particles re-deposited by the FIB, and they are therefore artifact

microstructure analysis. In each of the packages, a scanning/transmission electron microscopy (S/TEM) sample was prepared, by FIB liftout, to contain the bond on the pad that was connected to the emitter. The bond interface could also be analyzed with SEM after the interface was exposed by focused ion beam (FIB), which would provide fields of view larger than those acquired with S/TEM. The composition analysis was performed based on energy dispersive spectroscopy (EDS), and the composition was presented in the unit of atomic percent.

3 Electrochemical Investigation

3.1 Corrosion Potentials of Cu, Al, and Cu-Al IMCs

Lim et al. performed electrochemical studies of phases of Cu ball bonds in bulk form and immersed in DI water with 25 ppm chloride and pH = 6. The $[Cl^-]$ (hereon the square brackets, [], with a dissolved species inside, are used to represent the concentration of the species) and pH were chosen to simulate conditions for Cu ball bonds embedded in EMCs. The corrosion potentials of these phases were found to increase with the increase of Cu composition, in the order of

$$Al < CuAl_2 < CuAl < Cu_9Al_4 < Cu$$

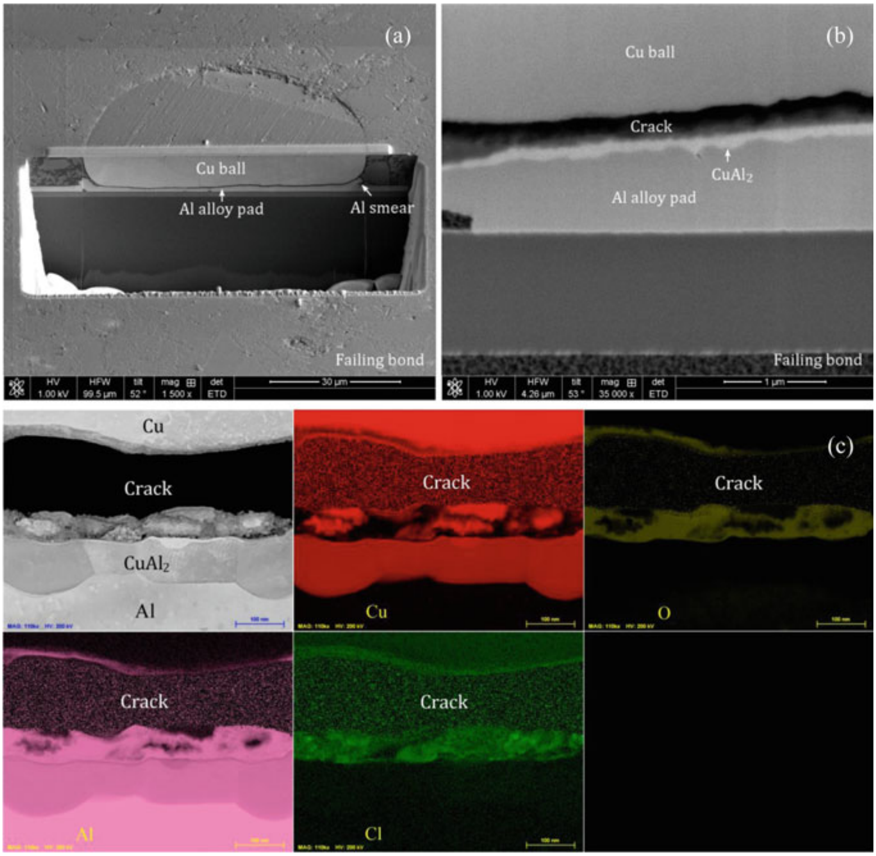


Fig. 12 Cross-sectional SEM images of the third bond (**a**, **b**) and STEM DF image and the associated elemental maps of the fourth bond **c** that failed the uHAST. A crack extended throughout the diameter of the bond with continuous air gap between the corrosion products and the Cu ball. **a**, **b** reveal the large ratio of (the length of the crack)/(the thickness of $CuAl_2$), which indicates the rapid propagation of the crack throughout the bond. The volume fraction of the Cu nanoparticles in the corrosion products is higher than those shown in Figs. 2, 9, 10 and 11, because the air gap is continuous and had blocked the movement of the Cu nanoparticles to coalesce with the Cu ball. This figure has revealed the presence of both the crack, and the corrosion products therefore indicate the SCC nature of the corrosion

as shown in Fig. 3a. Wu and Lee revealed the same sequence of corrosion potentials of

$$Al < CuAl_2 < Cu_9Al_4 < Cu$$

in aerated DI water with 25 ppm NaCl and pH = 6 [27], as shown in Fig. 3b. Corrosion potential is also called open-circuit potential.

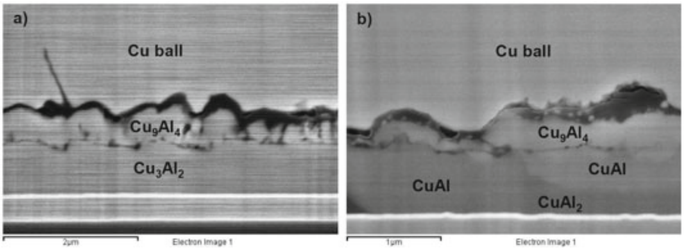
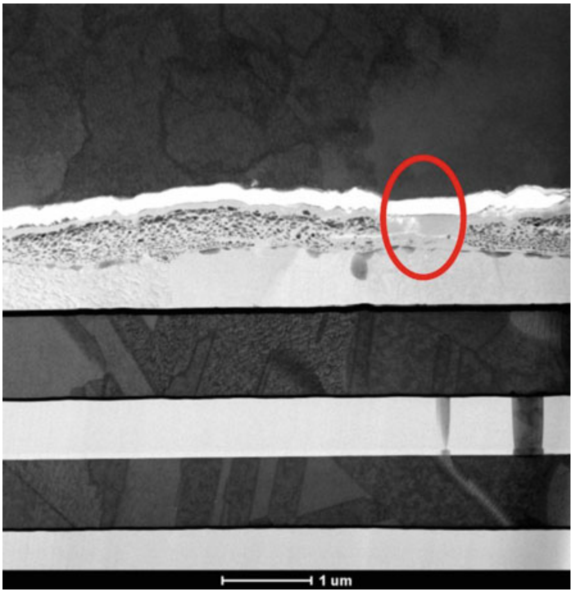


Fig. 13 SEM images of Cu ball bond interface after **a** baking at 300 °C for 24 h; **b** baking at 300 °C for 24 h, followed by autoclave for 96 h [28]. The figure shows that the corrosion took place in the MCRL of Cu_9Al_4, along the MCRL/Cu ball interface and simultaneously proceeds downward to consume MCRL. After the autoclave, as the volume fraction of air gaps under the Cu ball became larger, the Cu particles became large enough to be revealed clearly in **b**

Fig. 14 TEM image of a bond that failed after exposure to 288 h of uHAST [20]. The white band under the Cu ball is a crack with continuous air gap. The dark particles in the low-z matrix are Cu particles. The figure shows the microstructures across the failing bond, and their configuration is the same as those shown in Fig. 12. The cracks and the corrosion products are an indication of SCC



Part of the trend of corrosion potential increasing with the increase of Cu composition has been reported in the literature. For example, the corrosion potential of $CuAl_2$ was observed to be higher than that of the Al in 0.1 M Na_2SO_4 solution with pH = 4.3 [31]. A $CuAl_2$ particle was determined to be cathodic relative to Al [32], and the Al matrix behaved anodically and suffered a local corrosion along its interface with the $CuAl_2$ precipitates [33].

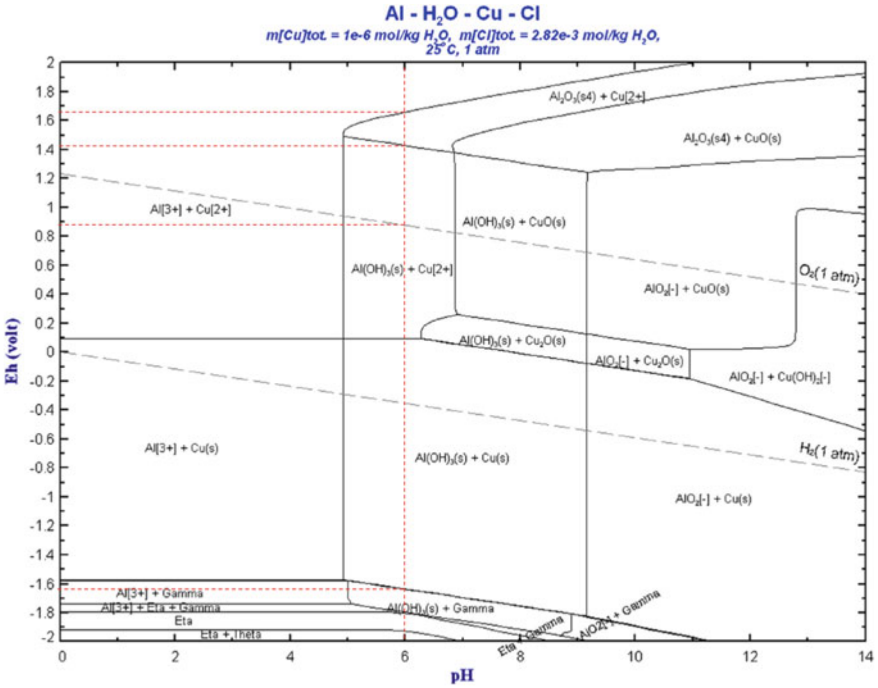


Fig. 15 Pourbaix diagram of Al–Cu–Cl–H_2O at 25 °C with the coexistence of Cu-Al IMCs [46]. Gamma, Eta, and Theta denote the phases of Cu_9Al_4, CuAl, and Cu_2Al, respectively. The red dashed lines are for the reading of electrode potential boundaries at pH = 6

3.2 Ingot Stack to Simulate Cu Ball Bonds and the Limitations

After the corrosion potentials of Al, $CuAl_2$, Cu_9Al_4, and Cu were measured, Wu & Lee stacked ingots of the four phases in the sequence of Al/$CuAl_2$/Cu_9Al_4/Cu and immersed the stack for weight loss measurement. The $CuAl_2$ and Al ingots were covered with the largest amount of corrosion products as shown in Fig. 3 of reference [27], which is presented in Fig. 4. Being the most anodic in the stack, $CuAl_2$ and Al exhibited the highest corrosion rates.

The ball bond stack, however, possesses some differences from the ingot stack. Such differences change the corrosion behaviors and therefore shall be noted, as follows.

1. The contact between adjacent layers is more intimate and more conductive in the Cu ball bond than in the ingot stack, since all the interfaces formed through solid-state reactions; in contrast, all the interfaces in the ingot stack resulted from physical stacking of the ingots.

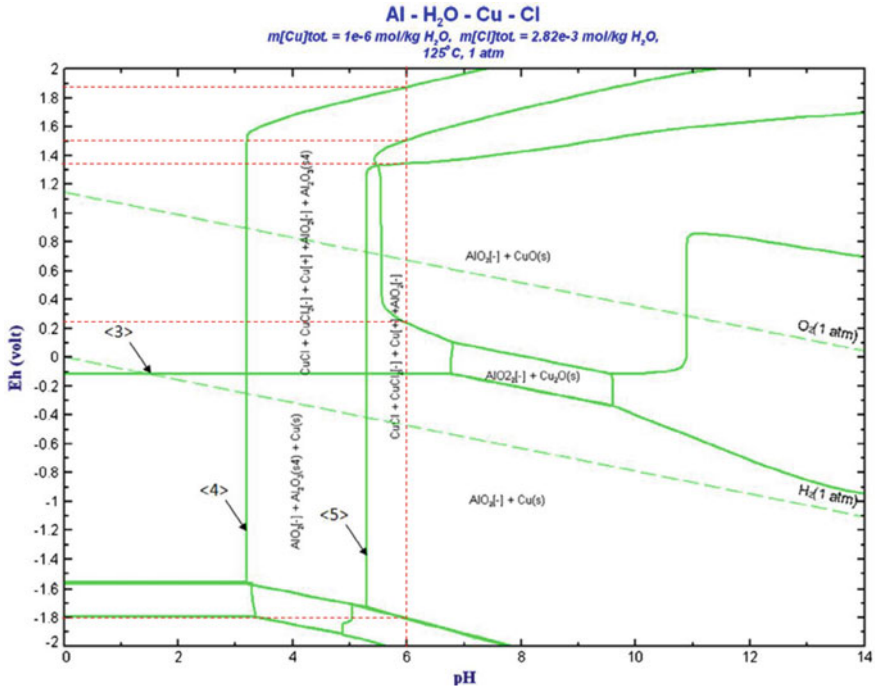


Fig. 16 Pourbaix diagram of Al–Cu–Cl–H_2O at 125 °C [46]. The figure indicates the corrosions of both Cu_9Al_4 and Cu are enhanced at the higher temperature of 125 °C compared with the lower temperature of 25 °C. The red dashed lines are for the reading of electrode potential boundaries at pH = 6

2. Interface III started as the bonding interface between the Cu ball and the bond pad and had captured non-bonding materials including residue from the bond pad opening process, the surface oxides of the Cu FAB and the bond pad, and voids from the ball bonding process. Interface III is therefore the most resistive among the four interfaces I–IV. The presence of voids and surface oxides in interface III was reported in reference [19]. On the other hand, the wet electrolyte on Cu ball bonds was present in the form of a thin water film that covered the Cu ball bonds. The formation of such a layer of water on surfaces of solids, including those for electronics, and more specifically Al oxide and hydroxide, has been well-documented in literature such as references [34–37]. Due to its small thickness and high resistivity, the thin-film electrolyte imparted a high Ohmic resistance to the closed circuit of the galvanic corrosion and substantially restricted the spatial range of the galvanic effect to the vicinity of the interface of a pair of joining phases [27, 38]. We can therefore consider galvanic effect of only pairs of joining phases. Across the bond, there are four pairs of such galvanic couples, which are Al-$CuAl_2$, $CuAl_2$-CuAl, (CuAl + $CuAl_2$)-MCRL, and MCRL-Cu. Interface III being the most resistive made the corrosion current

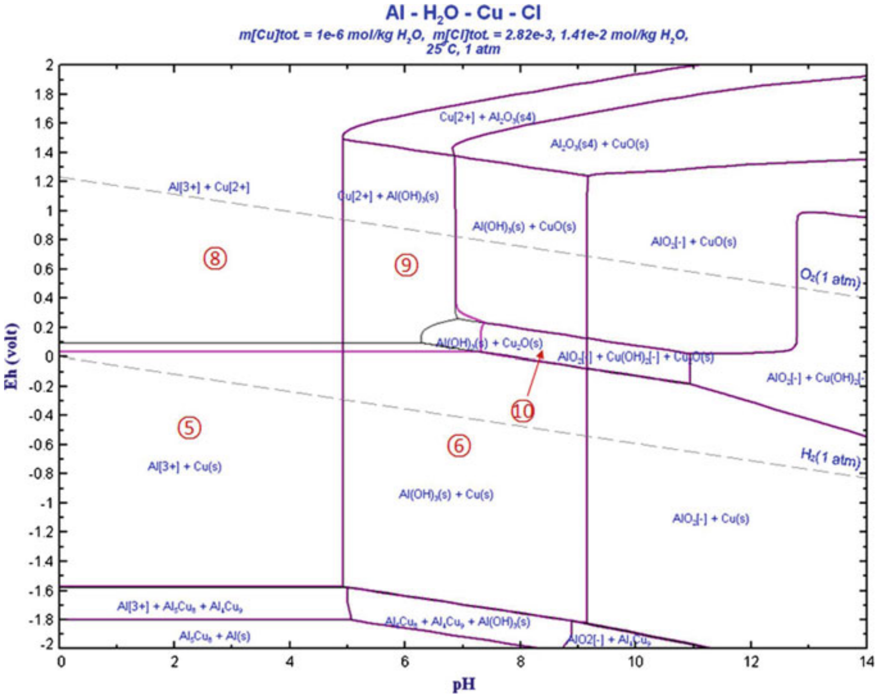


Fig. 17 Effect of [Cl^-] on Pourbaix diagrams of Al–Cu–Cl–H_2O at **25 °C** (black fine lines—100 ppm, magenta bold lines – 500 ppm) [46]. The figure shows that all the regions remain nearly unchanged compared with those in the diagram with [Cl^-] = 100 ppm, except those of Cu^{2+} expanding by about 0.1 V at the expenses of those of Cu or Cu_2O

in the (CuAl + $CuAl_2$)-MCRL couple lower than those of the other three galvanic couples. In the (CuAl + $CuAl_2$)-MCRL couple, the MCRL was less cathodically protected by $CuAl_2$, and meanwhile, $CuAl_2$ was less anodically corroded since its coupling with the MCRL was weak.

3. There was lattice strain along each of interfaces I–IV, due to lattice mismatch between the two joining phases. In addition, there was residual strain across the bond interface introduced by the ball bonding (in Sect. 7, “Addressing the corrosion,” a method of post-bond heating to relieve the strain and subsequently suppress the corrosion will be described). The strain elevated the energy of the local valence electrons and subsequently contributed to the corrosion of the anodic phase along the interface. The same principle of strain-induced preferential oxidation has been applied in the semiconductor industry to delineate the Si defect with a wet-etch, where the defect area is etched faster due to the elevated local lattice strain [39]. Across interface III, diffusion took place in localized areas where chemical bonds subsequently formed between CuAl and Cu_3Al_2. Three of such areas are marked by the vertical semitransparent red arrows in Fig. 2a. Outside of such areas, there was no chemical bond across interface III.

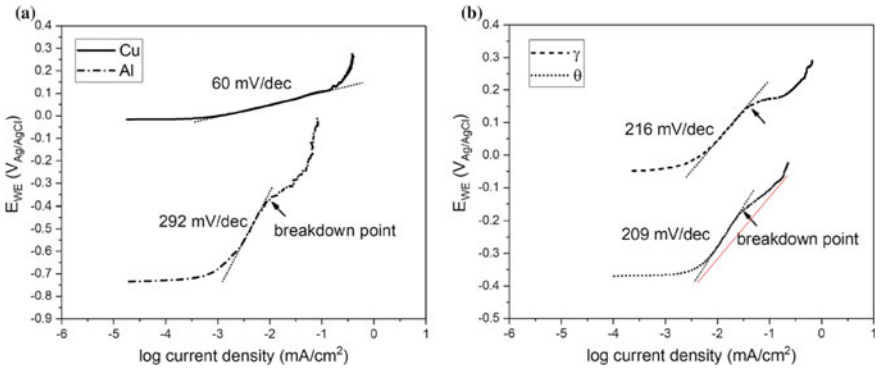


Fig. 18 Anodic polarization curves of Cu and Al (**a**), and Cu_9Al_4(γ) and $CuAl_2$(θ) **b** [27]. The Tafel slopes of Cu_9Al_4 and $CuAl_2$ are closer to that of Al than to that of Cu, suggesting the passivations of the two IMCs are due mainly to that of Al. The Tafel slopes of the two IMCs are smaller than that of Al, which indicates that the passivities of the two IMCs are lower than that of Al. The lower passivities result from the passivations being more defective, and likely thinner as well since the Al percentages of the two IMCs are lower than that of Al. The red line in **b** was drawn with the same Tafel slope of Cu_9Al_4 and positioned near the curve of $CuAl_2$, to show that the Tafel slope of Cu_9Al_4 is actually lower than that of $CuAl_2$. Therefore, it is possible that the Tafel slope values of Cu_9Al_4 and $CuAl_2$ labeled in **b** have been swapped. This similarly indicates the passivity of Cu_9Al_4 is lower than that of $CuAl_2$. The lower passivity results from the passivation being more defective, and likely thinner as well

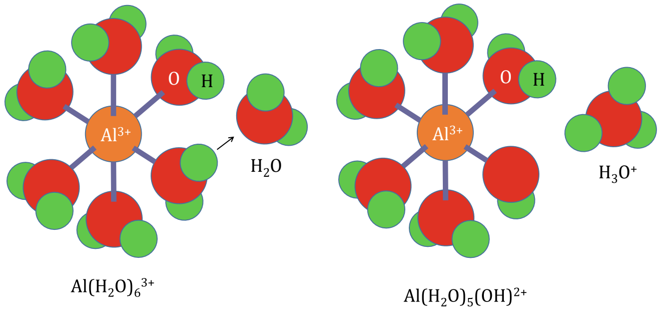


Fig. 19 Schematic to illustrate the deprotonation of the $[Al(H_2O)_6]^{3+}$(aq) complex ion

Interface III had the largest fraction of the total area that was free of chemical bond, and subsequently free of lattice strain along the interface. Therefore, interface III was the least impacted by lattice strain-assisted corrosion. As for the ingot stack, there was simply no lattice strain along any of the interfaces.

4. In Cu ball bonds, the perimeter of the bond interface is the boundary of multiple phases of Cu, MCRL, possibly CuAl, $CuAl_2$, Al alloy, and EMC. The local stress due to an external load is magnified along the perimeter of the bond, which can lead to SCC when pitting or crevice corrosion has taken place along interface IV.

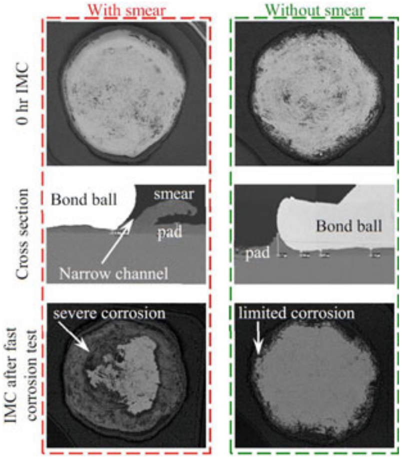


Fig. 20 Top-down and cross-sectional views of Cu ball bond on Al alloy pads with and without Al smear, before and after high-temperature storage (HTS) at 200 °C for 96 h [67]. The figure indicates a decrease of the Al bond pad smear had abated the corrosion

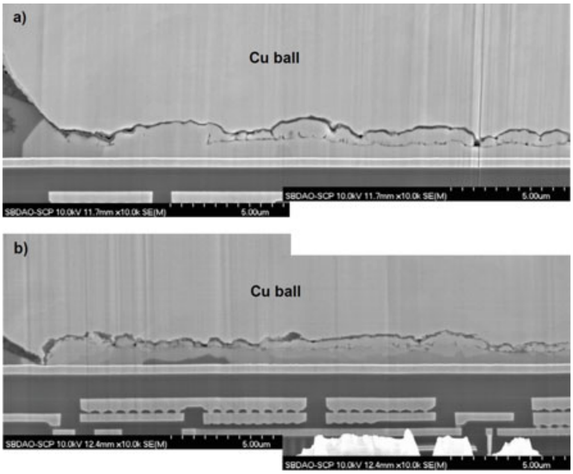

a)
Cu ball
SBDAO-SCP 10.0kV 11.7mm x10.0k SE(M) 5.00um
SBDAO-SCP 10.0kV 11.7mm x10.0k SE(M) 5.00um
b)
Cu ball
SBDAO-SCP 10.0kV 12.4mm x10.0k SE(M) 5.00um
SBDAO-SCP 10.0kV 12.4mm x10.0k SE(M) 5.00um

Fig. 21 Cross-sectional SEM view of Cu ball bonds in the samples **a** after baking at 300 °C for 24 h and **b** after the sequence of baking at 300 °C for 24 h/autoclave test [28]. The figure reveals the large ratio of (the lateral depth of the corrosion crevice/crack)/(the thickness of Cu_9Al_4), which indicates the rapid propagation of the crevice/crack throughout the bond, with SCC being a likely contributor

5. Within the range of significant galvanic current (determined by the resistance of the electrolyte), the surface area ratio of

[(exposed Cu ball)/(exposed MCRL)]

is larger than either of those of

[(exposed CuAl)/(exposed $CuAl_2$)] and

[(exposed $CuAl_2$)/(exposed bond pad)].

The difference elevated the anodic current density of the MCRL-Cu galvanic couple relative to those of the Al-$CuAl_2$ and $CuAl_2$-CuAl couples.

6. The Cu/Al atomic ratio was the highest at interface IV, and lowest at interface I. The anodic current of the individual Al atoms was the highest in the MCRL

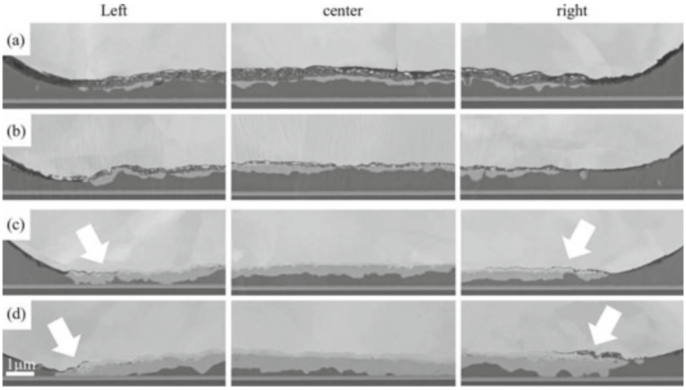


Fig. 22 Cross-sectional SEM images of Cu ball bond after bHAST at 130 °C, 85% RH, and 4 V bias for 96 h [30]. **a**–**d** are images of samples **a**–**d**, respectively. The processing conditions of the samples are as follows. a: no post-bond heating, b: post-bond heating at 180 °C for 48 h, **c** post-bond heating at 200 °C for 24 h, and **d** post-bond heating at 200 °C for 48 h. **a**, **b** each reveals not only the crevice/crack propagation along interface IV across the bond, but also an oxide matrix embedding Cu nanoparticles that resulted from the corrosion of the MCRL

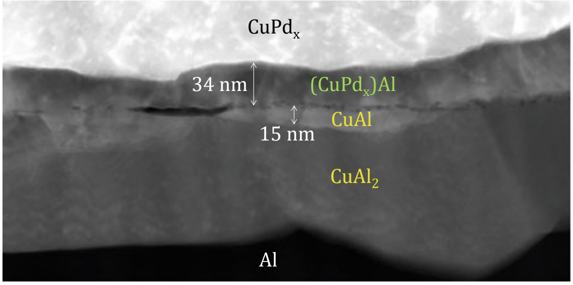
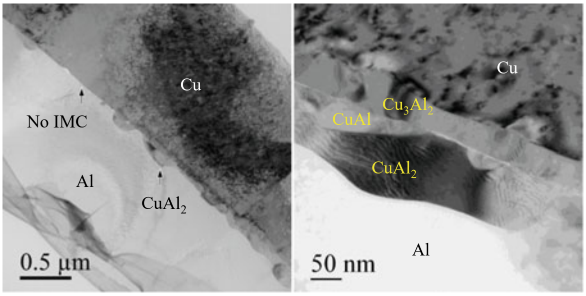


Fig. 23 HAADF image to show the presence of (CuPd$_x$)Al under the PCC ball bond along about 80% of the bond interface after uHAST for 384 h [76]. The Pd on one hand decreases the (Cu + Pd)/Al atomic ratio at interface IV and likely increases the areal fraction of interface IV in the bond area, but on the other hand enhances the cathodic reactions of ORR and HER. Nonetheless, the net result is the enhancement of the corrosion resistance

Fig. 24 TEM BF images to show the presence of Cu_3Al_2 under the Cu ball bond after uHAST for 384 h [76]



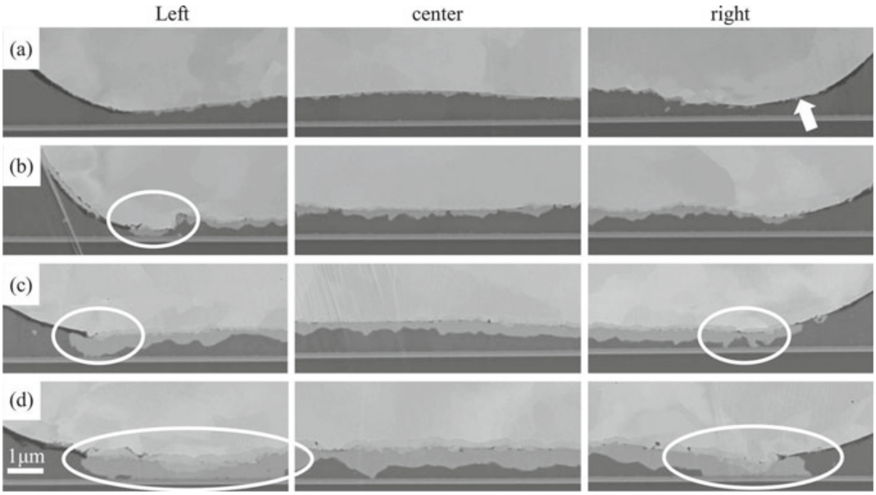


Fig. 25 Cross-sectional SEM images of bond interfaces after post-bond heating [30]. **a**–**d** are images of samples **a**–**d**. The processing conditions are as follows. **a**:No post-bond heating, **b** post-bond heating at 180 °C for 48 h, **c** post-bond heating at 200 °C for 24 h, and d:post-bond heating at 200 °C for 48 h. The figure shows the areal fraction of MCRL had increased with the post-bond heating, which led to a higher corrosion resistance

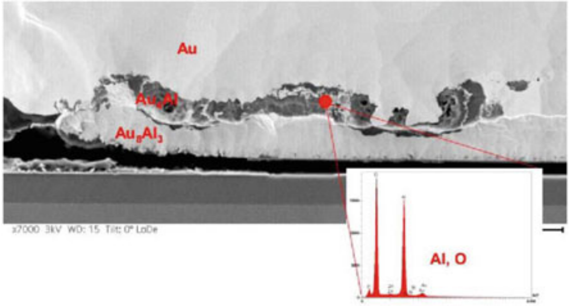


Fig. 26 HAADF image of a Au ball bond interface after pressure cooker test [85]. The figure shows that the corrosion in the Au ball bond had similarly propagated along the Au_4Al/Au ball interface and proceeded downward to consume Au_4Al

at interface IV and lowest in $CuAl_2$ at interface I; in addition, the highest Cu/Al atomic ratio at interface IV also caused the lowest passivity in the near-neutral electrolyte in the initial stage of the corrosion (in Section 6.1.1. "$Al(OH)_3/Al_2O_3$ dominates IMC passivation and the defectivity of IMC passivation," it will be elaborated that the passivity of a Cu-Al IMC in a near-neutral electrolyte is mainly due to that of Al in the IMC, i.e., the $Al(OH)_3/Al_2O_3$ precipitating on the surface of the IMC that results from the oxidation of the Al in the IMC). Both factors make the corrosion of the MCRL at interface IV the most aggressive [23]. The following observations have provided substantiation of the conclusion: (1) In a study of galvanic coupling between different model Al-Cu alloys, the higher the Cu content of a phase, the greater its susceptibility to pitting corrosion was [24].

(2) The weight loss rate of Al-Cu alloys in NaCl solution increased with the Cu content [40]. We extend the investigation to Au ball bonds on Al alloy pads, to look for further accreditation. After exposure of Au_4Al and Au_5Al_2 to Bi-phenyl epoxy resin, corrosion was visible only on Au_4Al (i.e., Au-richer than Au_5Al_2) but not Au_5Al_2, as shown in Fig. 8 of reference [29] which is presented in Fig. 5.

Here it was the Br^- ions that caused the corrosion of Au_4Al. In general, Br^- ions, along with Cl^- ions, can cause pitting corrosion of Al and passive metal alloys [41]. The preferential corrosion of Au_4Al over Au_5Al_2 induced by Br^- in Au ball bonds was also reported in reference [42].

Because of these geometric, structural, and electrochemical characteristics of Cu ball bonds, the corrosion rate of the exposed MCRL along interface IV should be the highest, while that of the exposed Al along interface I the lowest, in the initial stage of the corrosion. The corrosion should start at the exposed MCRL along interface IV, in the form of pitting corrosion [43], propagate along interface IV, and evolve into a crevice corrosion. The inference has been verified with microstructure characterization that is to be described.

4 Microstructures at the Bond Interfaces

4.1 *The Passing Interface*

TEM BF images of Cu-Al IMC layers across the bond that had passed the uHAST are presented in Fig. 6a–c, e, g [23]. $CuAl_2$, CuAl, and Cu_3Al_2 are located progressively upward. Each of these IMCs was identified with both electron diffraction and EDS quantification. Figure 6d, f, h displays the electron diffraction patterns of crystals of the three IMCs. The Cu compositions of the IMC crystals are 33% for $CuAl_2$, 49% for CuAl, and 63% for Cu_3Al_2 (in Cu_3Al_2, the Cu atomic percentage range is 60.7 to 61.8 at.%, per the Cu-Al phase diagram in reference [44]).

Figure 7 presents a STEM DF (darkfield) image of the bond that passed the uHAST and the associated elemental maps. The IMC stack is revealed. The top portion of Cu_3Al_2 in the area labeled “2” has been partly corroded (as indicated by the locally elevated brightness, or pixel values, in the O map). There is 22% of O in area 2, which is much lower than that of 64% in a completely corroded area that does not exhibit diffraction contrast and is to be shown in Fig. 9 in the following section. Area 2 also contains 0.2% of Cl; in contrast, Cl is below detection in both areas 1 and 3, as shown in Fig. 8. The data suggest the corrosion started along the Cu_3Al_2/Cu interface and propagated down to consume Cu_3Al_2, a subject that will be further explored in the forthcoming discussion.

The stacking sequence of the IMCs is consistent with the overall concentration gradients of Cu and Al between the Cu ball and the pad. The most Al-rich phase, $CuAl_2$, is on the Al alloy pad. The more Cu-rich phases of CuAl and Cu_3Al_2 are further away from Al and located closer to Cu. The MCRL is Cu_3Al_2 in some cases, such as those shown here and in Fig. 7d of reference [45], and Cu_9Al_4 in other cases, such as that shown in Fig. 5 of reference [28] and will be presented as Fig. 13 in this chapter. Therefore, interface IV is either the Cu_3Al_2/Cu ball interface or the Cu_9Al_4/Cu ball interface.

4.2 The Failing Interfaces

Figure 9 presents TEM BF images of the center of the first failing bond shown in Fig. 2 and an electron diffraction pattern of the $CuAl_2$ crystal [23]. The MCRL, which is Cu_3Al_2 per Figs. 6 and 7, and the forthcoming Figs. 10 and 11, had been entirely corroded. The Cu-Al interface has the following four components: (1) low-z (z: atomic number) matrix containing mainly Al (35%) and O (64%), and trace amounts of Cl (0.5%) and Cu (~0.5%), (2) voids, (3) Cu nanocrystals embedded in the matrix, (4) a CuAl nanocrystal inlaid in a $CuAl_2$ nanocrystal, and (5) The $CuAl_2$ nanocrystals on the pad. Phase identification of the $CuAl_2$ crystal in Fig. 9b was conducted with both EDS and electron diffraction. The Cu composition of the crystal is 32%. Figure 9c shows the $[-111]$ zone axis pattern of the crystal. The identification of these five structure components has been similarly reported by other researchers [17, 18, 20, 30]. No Cu-Al IMC was found along the perimeter of the Cu ball.

The evidences listed below form the base for an inference that will follow.

1. Cu_3Al_2 that was present under the Cu ball in the passing bond is absent in the failing bond. Under the Cu ball of the failing bond, there are the low-z matrix, voids, and Cu nanocrystals instead.
2. The low-z matrix contains mainly Al and O (with a sum of atomic percents of 99%), and trace amounts of Cl and Cu. This observation is consistent with the very small molar ratio of the oxidized Cu over the oxidized Al in Cu_3Al_2 stated in Sect. 5, "Thermodynamics of Cu ball bond corrosion from published Pourbaix diagrams." The matrix embeds the Cu nanoparticles and voids.
3. The voids exhibit smoothly curved perimeters that closely resemble circles.

The above evidences indicate the following.

1. The low-z matrix, Cu particles, and voids are products of a wet corrosion of Cu_3Al_2. The wet corrosion produced soluble and insoluble species. In the drying process, after the uHAST until the time when the TEM sample was positioned in the TEM, water in the electrolyte had evaporated, and voids were created consequently. The perimeters of the voids are smooth, which indicates surface tension of the water to keep the surfaces of the voids smooth, and the minimization of free energy to keep the voids close to spheres. The electrolyte for the wet corrosion was present as a thin water layer on Cu ball bonds [34–37], as has been stated in Sect. 3.2. "Ingot stack to simulate Cu ball bonds and the limitations."

2. The wet corrosion of Cu_3Al_2 was in the form of dealloying and involved the oxidation of mainly the more active component of Al since Al and O account for 99% of the matrix; in contrast, the exposed Cu atoms of Cu_3Al_2 were cathodically coupled to the exposed Al atoms and therefore largely protected. The immunity of the Cu atoms led to the coalescence of most of them to form Cu nanoparticles, which tended to further coalesce with the Cu ball above, before a continuous air gap emerged as a result of SCC and blocked their upward movement, The coalescence was driven by the minimization of the surface energies of the Cu atoms and Cu nanoparticles. The corrosion in the sample was in the advanced stage of crevice corrosion/SCC. Per the Pourbaix diagrams of the Al–Cu–Cl–H_2O system presented in reference [46], and the discussion in Sect. 6.2, "Advanced stage of corrosion—crevice corrosion/SCC", in the crevice the electrolyte was acidic, and the exposed Al atoms of Cu_3Al_2 were oxidized to Al^{3+} ions. The Al^{3+} ions were subsequently hydrated to form Al hexahydrate $[Al(H_2O)_6]^{3+}$(aq). The low-z matrix resulted from the dehydration of the Al hexahydrate $[Al(H_2O)_6]^{3+}$(aq) to a certain degree. On the other hand, the Cu atoms in Cu_3Al_2 largely resisted corrosion in an acidic electrolyte with a very small fraction oxidized to Cu^{2+}(aq). Therefore, the thermodynamic information is consistent with the microstructure characterization.

Figures 10 and 11 present high angle annular dark field (HAADF) images of the second failing bond with the MCRL of Cu_3Al_2 being not completed corroded. The crevice corrosion had taken place in Cu_3Al_2 along interface IV, the Cu_3Al_2/Cu interface; i.e., Cu_3Al_2 was corroded from interface IV. There is no air gap shown in the image Voiding in the corrosion products is less severe than that shown in Fig. 9.

Figure 11 includes elemental maps associated with the HAADF image. The corrosion products similarly contain mainly Al, O, and small amounts of Cl (0.2%) and Cu (2%). The small white particles, a few of which are marked by the white arrows, in the $CuAl_2$ area are Cu particles re-deposited by the FIB and they are therefore artifact. The C detected along the Cu_3Al_2/$CuAl_2$ interface could result from a combination of photoresist residue for opening the bond pad and C contamination built up by the electron beam. The following are revealed in the figure: (1) The surface of the residual Cu_3Al_2 is covered mainly by Al oxide/hydroxide instead of Cu, as indicated by the high pixel values of Al and O and low pixel values of Cu on the residual Cu_3Al_2, in the Al, O, and Cu maps, respectively. There is not a Cu layer on the residual Cu_3Al_2 layer (such a Cu layer, had it been formed, would show up with dark pixels on the residual Cu_3Al_2 layer in both the Al and O maps). (2) The Cu nanoparticles are separated from the surface of the residual Cu_3Al_2. The three Cu nanoparticles marked by the green arrows in each of the Cu, Al, and O maps are likely such particles. These three Cu nanoparticles exhibit high pixel values in the Cu map, but low pixel values in the Al and O maps. These low pixel values suggest that the Cu nanoparticles are not nanoparticles re-deposited by FIB, which would otherwise have these pixel values being high from the underlying Al oxide/hydroxide (in the Al and O maps). The self-coalescence of the Cu nanoparticles and their coalescence with the Cu ball are supported by the very small volume fraction of the Cu nanoparticles in the corrosion products, and the three marked Cu nanoparticles are located further away from the residual Cu_3Al_2 layer and closer to the Cu ball. After an air gap emerges under the Cu ball as a result of SCC, the movement of the Cu nanoparticles toward the Cu ball will be stopped. The Cu nanoparticles will start to coalesce with themselves instead, which, nonetheless, will still cause the Cu nanoparticles to be separated from the residual Cu_3Al_2 layer. The separation of Cu nanoparticles from the Cu ball can be seen in Figs. 2 and 9 where the air gap is discontinuous and in the form of discrete voids, and in the Cu map of Fig. 12(c), where the air gap is continuous. Figure 12 displays a STEM DF image of the center of the third bond that failed the uHAST, and the associated elemental maps. A crack had extended throughout the diameter of the bond with a continuous air gap between the corrosion products and the Cu ball (the difference between a crack and an air gap is emphasized here since in Figs. 2 and 9 there is likely a crack that is continuous, but the air gap is discontinuous). The corrosion products are located under the crack. The matrix similarly contains mainly Al and O, and trace amounts of Cl (0.6%) and Cu (0.4%) and embeds Cu nanoparticles. Figure 12 has revealed the presence of both the corrosion products and the crack with a continuous air gap, therefore indicated the SCC nature of the corrosion.

With the formation of air gap as shown in Figs. 9 and 12, the volume fraction of the Cu nanoparticles in the corrosion products shall become higher than its counterpart with no air gap as shown in Fig. 11. This turns out to be true, since the volume fraction of Cu nanoparticles in the corrosion products increases in the order of "No air gap" (Fig. 11) < "Discrete air gap" (Fig. 9) < "Continuous air gap" (Fig. 12c). The separation of the Cu nanoparticles from the residual Cu_3Al_2 caused the passivation to be weakened during the crevice corrosion of the Cu_3Al_2 in Cu ball bonds.

An observation of crevice corrosion with no air gap and similar to that shown in Figs. 10 and 11 was reported for a bond with Cu_9Al_4 as the MCRL, after the processing sequence of preconditioning (125 °C/24 h, 30 °C/60RH/192 h, 3 × Pb-free reflow at 260 °C), baking at 300 °C for 24 h, and autoclave for 96 h [28]. As shown in Fig. 5 of reference [28], presented in Fig. 13, the Cu_9Al_4 had not been entirely corroded, and the corrosion products are located in the top portion of Cu_9Al_4 under the Cu ball. After the baking, there had already been a crevice corrosion along interface IV as shown in part (a) of the figure. After the autoclave, as the volume fraction of air gaps under the Cu ball became larger, the Cu particles became large enough to be revealed clearly in part (b) of the figure.

Corrosion products consisting of a low-z matrix embedding Cu nanoparticles and under a crack have also been reported and shown in Fig. 5a of reference [30], Fig. 6 of reference [17], and Fig. 3 of reference [20]. Figure 3 of reference [20] is presented in Fig. 14.

The data presented in Figs. 2, 6, 7, 8, 9, 10, 11, 12, 13 and 14, indicate that the Cu ball bond failure resulted from crevice corrosion of the MCRL along interface IV, under an electrolyte of water containing mobile Cl^- ions from the EMC. In each of the cases shown in Figs. 12 and 14, there is a component of SCC.

5 Thermodynamics of Cu Ball Bond Corrosion—Pourbaix Diagrams

The thermodynamics of the corrosion of Cu ball bonds, and more specifically of Cu_9Al_4 along interface IV, can be assessed with Pourbaix diagrams [47, 48]. A Pourbaix diagram, also known as a potential/pH diagram, EH–pH diagram, or equilibrium diagram, is a graphical representation of the corrosion thermodynamic information, in a potential–pH diagram, for a metal in equilibrium with its aqueous environment. The diagram shows the immunity, passivation, and corrosion regions of the metal. Stable species are marked in the corresponding regions. Boundaries between regions are represented by lines. A Pourbaix diagram is therefore analogous to a standard phase diagram, but with a different pair of axes.

Two ambient factors of the corrosion of Cu ball bonds are humidity and temperature. When the humidity is sufficient for a water layer to form on the Cu ball bond, wet corrosion takes place. The ambient temperature varies and can reach much higher than room temperature, such as that of about 150 °C for automotive underhood components. Pourbaix diagrams of the Al–H_2O–Cu-Cl system at 25 and 125 °C will be used for the discussion since they are available [46]. The discussion is separated into sections about the initial and advanced stages of the corrosion. In the initial stage of the corrosion, pH = 6, and in the advanced stage of the corrosion, pH is below 6 since the electrolyte has been acidified, and the $[Cl^-]$ also becomes higher than that in the initial stage of the corrosion.

5.1 Initial Stage of the Corrosion at 25 °C

Pourbaix diagram in Fig. 1 of reference [46], under the conditions of 25 °C, $[Cl^-]$ = 100 ppm, and pH = 6, is presented in Fig. 15. Stable species under the conditions are listed in Table 2.

The data indicate that Cu_9Al_4 starts to be oxidized at $E = -1.64$ V. In range b, Cu_9Al_4 is corroded to $Al(OH)_3(s)$ + Cu(s). In the ranges b–d, Cu_9Al_4 possesses

Table 2 Stable species under the conditions of 25 °C, $[Cl^-]$ = 100 ppm, and pH = 6. E is the electrode potential

E (V)	E range identifier	Stable species	Region
[− 2.00, − 1.64]	a	Cu_9Al_4(s)	Immunity
[− 1.64, 0.09]	b	$Al(OH)_3$(s) + Cu(s)	Passivity
[0.09, 1.42]	c	$Al(OH)_3$(s) + Cu^{2+}	Passivity
[1.42, 1.66]	d	Al_2O_3(s) + Cu^{2+}	Passivity
[1.66, 2.00]	e	Al^{3+} + Cu^{2+}	Corrosion

a passivity imparted by the three solid species of $Al(OH)_3(s)$, Cu(s), and $Al_2O_3(s)$, while in the topmost potential range e there is no longer any passivation species, and Cu_9Al_4 is in active dissolution, i.e., Cu_9Al_4, together with Cu of the Cu ball, enters the corrosion region above $E = 1.66$ V. In ranges c and those higher, the solid species Cu(s) is oxidized to the soluble species $Cu^{2+}(aq)$. Throughout the entire pH range, the boundary lines between the immunity region and oxidation (including corrosion and passivation) region of Cu are much higher than those of any of the three IMCs, including Cu_9Al_4. With the electrode potential not too far above these boundary lines of Cu, the molar ratio of the oxidized Cu over the oxidized Al in Cu_9Al_4 is very small. The same conclusion is expected for Cu_3Al_2, i.e., a very small molar ratio of the oxidized Cu over the oxidized Al in Cu_3Al_2. Overall the trend is the progressive corrosion enhancement toward anodic dissolution of Cu, Al, and IMCs with the increase of electric bias. On the other hand, the passivation film is defective due to the inhomogeneity of the film which lowers the passivity of the film. In addition, above the slanted line for the anodic evolution of $O_2(g)$ at $E = 1.23 - 0.0592\ \text{pH} = 1.23 - 0.0592 \times 6 = 0.88$ V, water molecules dissociate to produce $O_2(g)$ via the following reaction of

$$2H_2O(l) = O_2(g) + 4H^+(aq) + 4e^-.$$

The production of O_2 and H^+ ions enhances the corrosion of the MCRL. Below the bottom slanted line for cathodic hydrogen evolution reaction (HER) at $E = -\ 0.0592\ \text{pH} = -\ 0.0592 \times 6 = \text{-}0.355$ V, $H_2(g)$ is evolved via the reaction of

$$2H^+(aq) + 2e^- = H_2(g).$$

A description of the two lines can be found in Chap. 4.2 POURBAIX DIAGRAM FOR WATER of reference [41]. The production of H atoms in the intermediate step of HER increases the likelihood of hydrogen embrittlement and subsequently SCC.

The small molar ratio of the oxidized Cu over the oxidized Al in the MCRL, the defectivity of the passivation film on the MCRL, anodic evolution of $O_2(g)$, and cathodic HER are valid in the following two stages of corrosion, and their description will not be repeated.

5.2 Initial Stage of the Corrosion at 125 °C

The Pourbaix diagram in Fig. 2 of reference [46], i.e., 125 °C, $[Cl^-] = 100$ ppm, and pH = 6, is presented in Fig. 16. Stable species under the conditions are listed in Table 3.

The data reveal that Cu_9Al_4 starts to be oxidized at $E = -\ 1.80$ V, which is more negative than that of − 1.64 V at 25 °C as described in the previous section. The difference indicates Cu_9Al_4 is less stable at the higher temperature of 125 °C. The presence of soluble species of Al^{3+} in the range of pH < 3.2, and $AlO_2{}^-$ elsewhere

Table 3 Stable species under the conditions of 125 °C, [Cl$^-$] = 100 ppm, and pH = 6

E (V)	E range identifier	Stable species	Region
[− 2.00, − 1.80]	f	$Cu_9Al_4(s)$	Immunity
[− 1.80, 0.09]	g	$AlO_2^- + Cu(s)$	Passivity
[0.09, 0.24]	h	$CuCl + CuCl_2^- + Cu^+ + AlO_2^-$	Corrosion
[0.24, 1.34]	i	$AlO_2^- + CuO(s)$	Passivity
[1.34, 1.50]	j	Not specified	
[1.50, 1.85]	k	$CuCl + CuCl_2^- + Cu^+ + AlO_2^- + Al_2O_3(s4)$	Passivity
[1.85, 2.00]	l	$Al^{3+} + Cu^{2+}$	Corrosion

indicates a partial passivity breakdown of Al and Cu_9Al_4 throughout the entire pH range, including pH = 6. This thermodynamic enhancement of the corrosion across the entire pH range at the high temperature of 125 °C is unique among the three corrosion stages discussed Sects. 5.1 to 5.3.

5.3 *Advanced Stage of the Corrosion at 25 °C*

In the advanced stage, the corrosion has evolved into crevice corrosion. Crevice corrosion is characteristic of high [Cl$^-$] and low pH, which will be elaborated in Sect. 6.2.1. "Corrosion enhancement in the crevice." The Pourbaix diagram in Fig. 3 of reference [46], under the conditions of 25 °C and [Cl$^-$] = 500 ppm is presented in Fig. 17. Stable species under the conditions are listed in Table 4.

The data reveal that Cu_9Al_4 starts to be oxidized at $E = -1.57$ V, which is slightly more positive than those of − 1.64 V and − 1.80 V at pH = 6 described in the previous two sections. The difference arises because Al is more stable at low pH than high pH as indicated by the Al Pourbaix diagram; however, Cu_9Al_4, together with Cu, enters the corrosion region o at $E = 0.03$ V, which is much lower than that of 1.66 V in the initial stage of the corrosion **at** 25 °C with [Cl$^-$] = 100 ppm. The difference indicates the thermodynamic enhancement of the corrosion/dissolution of Al caused by the decrease of pH. From range m to n, $Cu_9Al_4(s)$ is corroded to Al^{3+} + Cu(s). From range n to o, the solid species Cu(s) is oxidized to the soluble species Cu^{2+}. In range n, Cu_9Al_4 possesses a passivity imparted by the solid species

Table 4 Stable species of the under the conditions of 25 °C, [Cl$^-$] = 500 ppm, and pH ∈ [0, 5]

E (V)	Electrode potential range identifier	Stable species	Region
[−2.00, − 1.57]	m	$Cu_9Al_4(s)$	Immunity
[−1.57, 0.03]	n	$Al^{3+} + Cu(s)$	Passivity
[0.03, 2.00]	o	$Al^{3+} + Cu^{2+}$	Corrosion

of Cu(s), while in the topmost potential range o there is no longer any passivation species, and both Cu_9Al_4 and Cu are in active dissolution, Overall the trend is the progressive enhancement of the corrosion toward dissolutions of Cu, Al, and IMCs with the increase of electric bias.

To be complete, a Pourbaix diagram at a high temperature, such as 125 °C, and a high Cl^- ion concentration, such as $[Cl^-] = 500$ ppm, is needed for the thermodynamic assessment of the crevice corrosion during the uHAST.

6 The Corrosion Process

The corrosion process can be considered to consist of the initial stage and the advanced stage. The initial stage spans up to the formation of pits, during which the pH stays close to the starting value of 6. The advanced stage includes the pit formation and the subsequent crevice corrosion that can be accompanied by SCC. Inside the crevice the pH decreases to significantly below the starting value of 6.

6.1 Initial Stage of Corrosion

6.1.1 $Al(OH)_3/Al_2O_3$ Dominates IMC Passivation and the Defectivity of IMC Passivation

Anodic polarization curves of Cu, Cu_9Al_4, $CuAl_2$, and Al reported in reference [27] are presented in Fig. 18. The electrolyte temperature was not specified, but was likely room temperature. Therefore, it is assumed to be room temperature. In the figure, there is a linear segment of every polarization curve. The slope of this linear segment is the Tafel slope, which indicates for a given increase in the corrosion current density, the amount of the electrode potential increase needs to be. The Tafel slope is therefore an important parameter of the passivation. The Tafel slopes of the four phases are labeled in the figure. The slopes decrease with the Cu/Al atomic ratio, i.e., in the order of

$$Al > CuAl_2 > Cu_9Al_4 > Cu.$$

The figure shows the following.

- the Tafel slopes of Cu_9Al_4 and $CuAl_2$ are closer to that of Al than to that of Cu. This suggests the passivations of the two IMCs are due mainly to that of Al. This is consistent with the Cu in the IMCs largely remaining immune to corrosion under the specified conditions. The Cu atoms' self-coalescence and their further coalescence with the Cu ball lead to the separation of Cu nanoparticles from the passivation layer of the IMC, which greatly decreases the Cu atoms' contribution to the passivation of the IMC.

- the Tafel slopes of the two IMCs are smaller than that of Al, which indicates that for the same amount of potential increase, the current densities of the IMCs increase more than that of Al. Therefore, the passivities of the two IMCs are lower than that of Al. The lower passivities result from the passivations being more inhomogeneous or defective, and likely thinner as well since the Al percentages of both IMCs are less than that of Al which equals one. The higher defectivity increases both the electronic and ionic conductivities of the passivation layer and enhances the corrosion of the MCRL. This bullet point, and the next, responds to the correlation between the high Cu/Al atomic ratio and low passivity stated in bullet point #6 in Sect. 3.2. "Ingot stack to simulate Cu ball bonds and the limitations."
- the Tafel slope of Cu_9Al_4 is actually lower than that of $CuAl_2$ (this is illustrated with the red line in (b) being drawn with the same Tafel slope of Cu_9Al_4 and positioned near the linear segment of the curve of $CuAl_2$. Therefore, it is possible that the Tafel slope values of Cu_9Al_4 and $CuAl_2$ labeled in (b) have been swapped). This similarly indicates the passivity of Cu_9Al_4 is lower than that of $CuAl_2$. The lower passivity results from the passivation being more defective, and likely thinner as well since the Al percentage of Cu_9Al_4 is lower than that of $CuAl_2$.

The above interpretation of the figure is consistent with the thermodynamic information retrieved from the Pourbaix diagrams presented in Sect. 5, "Thermodynamics of Cu ball bond corrosion from published Pourbaix diagrams". Per Tables 2, 3 and 4, soluble and insoluble species are produced in the initial stage of the corrosion of the MCRL. The Cu-containing solid species produced are Cu(s) and CuO(s). The following processes involving Cu(s) create broken bonds and open space in the passivation layer and contribute to the defectivity of the passivation layer.

- The incorporation of soluble species in the passivation layer as the $Al(OH)_3$ or Al_2O_3 precipitates on the MCRL.
- For the species of Cu(s), the Cu atoms' self-coalescence and their further coalescence with the Cu ball and the resultant separation of Cu nanoparticles from the passivation layer of the IMC.

As for CuO(s), it is produced together with AlO_2^-(aq) in potential range i per Table 3 at 125 °C and pH = 6. It will soon be shown that CuO(s) will no longer be a stable species as the corrosion evolves, and the passivity of the MCRL will be due to Al_2O_3(s) as the pH enters the range of [3.2, 5.3].

At the two temperatures of 25 and 125 °C and with pH = 6, and outside of the immunity region, the stable Al-containing species are listed below, per Tables 2 and 3.

- $Al(OH)_3$(s), Al_2O_3(s) and Al^{3+}(aq) at 25 °C,
- AlO_2^-(aq), Al_2O_3(s4) and Al^{3+}(aq) at 125 °C.

With either $Al(OH)_3(s)$ or $Al_2O_3(s)$ precipitating on the surface of the MCRL at 25 °C, it imparts the MCRL passivity. The following describes at 125 °C, how the initial production of $AlO_2^-(aq)$ and CuO(s) leads to the precipitation of $Al_2O_3(s)$ on the MCRL to impart passivity.

At 125 °C, AlO_2^- ions are produced via the reaction of

$$Cu_9Al_4 + 8H_2O - 12e = 4AlO_2^- + 16H^+ + 9Cu,$$

which is essentially [41, 49]

$$Al + 2H_2O - 3e = AlO_2^- + 4H^+. \tag{1}$$

Reaction (1) proceeds with the products of AlO_2^- and H^+ at a molar ratio of AlO_2^-: H^+ = 1:4; meanwhile AlO_2^- and H^+ react with a 1:1 molar ratio to form Al_2O_3 and H_2O, which can be described by the following reaction [41, 49]

$$2AlO_2^- + 2H^+ = Al_2O_3 + H_2O \tag{2}$$

The forward reaction rate of reaction (2) is higher than the back reaction rate, until the equilibrium of reaction (2) is reached, when $\frac{1}{[AlO_2^-]^2[H^+]^2} = K$ is satisfied, where K is the equilibrium constant of reaction (2).

Since the AlO_2^- and H^+ are produced at the molar ratio of 1:4 per reaction (1), but consumed at the ratio of 1:1 per reaction (2), the $[H^+]$ increases in the vicinity of the corrosion surface. The local pH on the corrosion surface decreases. When the pH is in the range of [3.2, 5.3], CuO(s) is no longer stable, and the only stable solid species other than Cu(s) is $Al_2O_3(s)$. This is how the $Al_2O_3(s)$ is produced on MCRL to impart passivity to the MCRL.

Based on the preceding analysis, at 25 °C and pH = 6, the corrosion of the MCRL produces $Al_2O_3(s)$ or $Al(OH)_3(s)$, which causes the passivation of the MCRL. At 125 °C and pH = 6, the corrosion of the MCRL acidifies the surface of the MCRL in the process to produce AlO_2^-, which leads to the formation of $Al_2O_3(s)$ that imparts passivation to the MCRL. To summarize, $Al_2O_3(s)$ or $Al(OH)_3(s)$ forms on the surface of the MCRL in the initial stage of the corrosion, which imparts passivity to the MCRL.

A similar observation was reported from a study of the effect of Al content on the anodic behavior of Cu-Al alloys in 3.5 wt% NaCl solution, In the study, pure Cu and single α-phase Cu-Al alloys (which are solid solutions with Cu as the solvent and Al the solute) developed a surface layer of CuO or $Cu(OH)_2$ at the passivation potential [50]. However, as the Al content was increased to 10 wt%, the Cu-Al alloy was passivated to $Al(OH)_3$. Since the Al contents in $CuAl_2$ (Al 45.9 wt.%), Cu_3Al_2 (Al 22.0 wt.%), and Cu_9Al_4 (Al 15.9 wt.%) are all higher than 10 wt%, the passivation layers of the three IMCs are expected to be predominantly $Al(OH)_3$ or Al_2O_3. Similarly, it was reported that the anodic dissolution of Cu_9Al_4 and $CuAl_2$

started with a dealloying step leading to the ionization of Al to Al^{3+}, whose hydrolysis led to the formation of $Al(OH)_3$ [51].

6.1.2 The Pitting

Cl^- Adsorption to the IMC Surface

Along the perimeter of the bond interface, and likewise elsewhere, the electrolyte is near-neutral to neutral before the corrosion. The isoelectric points (IEP) are 9.0–9.4 for Al_2O_3 [52], 7.5 for Cu_2O [53], and 8.4–9.5 for CuO [54]. As a result, the passivation film of the MCRL acquires a positive surface charge, with no electric bias applied to the Cu ball bond. Cl^- ions therefore diffuse and migrate from the EMC to the MCRL along the perimeter of the bond and become adsorbed on the MCRL, driven by concentration gradient and attracted by the positive charge on the passivation film of the MCRL.

The polarity of the electric bias on Cu ball bonds plays an important role in the $[Cl^-]$ along the bond perimeter. With a positive bias, the attraction of Cl^- ions toward the bond and the tendency of anodic dissolution of the MCRL both become stronger, which enhances the corrosion. Quantitatively, when the electric field was greater than 2 V/cm, the diffusion coefficient of Cl^- ions in EMC increased by one order of magnitude [21]. With a negative electric bias, the impact is the opposite; i.e., the transport of Cl^- ions toward the bond perimeter is decelerated and the tendency of cathodic protection of the MCRL both becomes stronger, which abates the corrosion. However, qualitative assessment of the impact of negative electric bias on the reliability of Cu ball bonds has not been reported.

The Pitting Mechanisms

Adsorption of Cl^- ions on the MCRL will cause pitting corrosion of the MCRL along interface IV. The pitting corrosion breaks down the passivation of the MCRL. Since the passivation of the MCRL is attributed mainly to that of Al, mechanisms of the pitting corrosion and crevice corrosion of Al are active in Cu ball bonds and are described as follows.

The pitting corrosion of Al can be divided into an initiation stage and a propagation stage [55]. The initiation stage includes the passive film breakdown and the onset of an anodic current. These initiation mechanisms involve the transport of Cl^- ions through the passivation film to reach the underlying MCRL, and the dissolution of the passivation film. The exact mechanism of pit initiation is not known with great certainty, but it is generally understood that the following three main mechanisms are possible:

1. The penetration mechanism. The adsorbed Cl^- ions are transported, by diffusion and migration (due to the attraction by the Al^{3+} ions at the MCRL/passivation

interface and the positive surface charge on the passivation film on the MCRL), across the passivation film, possibly through oxygen vacancies or water channels, to the substrate surface where they participate in local dissolution of Al at the MCRL/passivation interface [56]. The ionic radii are 1.84 Å for Cl^- and 1.40 Å for O^{2-}, with the former being slightly larger than the latter, which makes it possible for the Cl^- ions to be transported through oxygen vacancies.

2. The film thinning mechanism. The adsorbed Cl^- ions form complexes with the oxide film which leads to local dissolution and thinning of the passivation film. Al_2O_3 being dissolved to some extent in the presence of NaCl solutions provided support for this mechanism [57]. High electric field in the passivation film can also lead to electrochemical breakdown of the film [58].
3. Film rupture mechanism. The adsorbed Cl^- ions penetrate the passivation film through cracks or flaws in the film to reach the MCRL/passivation interface.

These mechanisms do not mutually exclude each other necessarily; instead, there can be variable combinations of them under different circumstances. The defectivity of the passivation film of the MCRL enhances all three mechanisms, as the defectivity enhances the Cl^- ions' transport through and the dissolution of the film, as well as the transport of water molecules through the film.

In addition to the Cl^- ions, molecular water also reaches the MCRL/passivation interface through structural defects or pores. Lastly, there are dissolved O_2 molecules adsorbed on the passivation film of the MCRL. The adsorbed O_2 dissociates into two adsorbed O atoms.

With both the Cl^- ions and molecular water at the MCRL/passivation interface, and adsorbed O atoms on the passivation film, the Al at the MCRL/passivation interface becomes oxidized. There are three factors to drive the Al oxidation of

$$Al = Al^{3+} + 3e.$$

- The O_{ads} attracts the valence electrons of the Al under the passive film [59].
- The Cl^- is a Lewis base, tending to bond with Al^{3+} which is a strong Lewis acid (since the difference in electronegativity of $\Delta\chi_{\text{Al-Cl}} = 1.5 < 1.7$, The Al-Cl bond is mainly covalent and not as strong).
- H_2O is also a Lewis base tending to bond with Al^{3+}, and with an affinity to Al^{3+} being higher than that of Cl^- to Al^{3+}. An Al^{3+} ion bonds to six water molecules to form an aluminum hexahydrate $Al(H_2O)_6{}^{3+}(aq)$ which is a complex ion, through the following reaction of

$$Al^{3+}(aq) + 6H_2O(l) = Al(H_2O)_6^{3+}(aq)$$

Due to the small size and high electric charge of Al^{3+}, the hydration of Al^{3+} is highly exothermic. The Al^{3+} polarizes the attached H_2O ligands, weakening the O-H bonds and enabling the solvent H_2O molecule to act as a Lewis base to extract H^+

from $[Al(H_2O)_6]^{3+}$. The H_2O ligand is more acidic than a free H_2O molecule for two reasons [60]. First, there is a repulsion between the Al^{3+} and the partially positively charged $H^{\delta+}$ of a H_2O molecule. Secondly, the Al^{3+} attracts valence electrons from the O atoms of the H_2O molecules, which decreases the electron density in the O–H bonds and weakens the O-H bonds. Both tend to make the H_2O molecule lose an H^+. The complex ion is consequently deprotonated, which can be described by the following reaction of

$$\left[Al(H_2O)_6\right]^{3+}(aq) + H_2O(l) \rightarrow \left[Al(H_2O)_5(OH)\right]^{2+}(aq) + H_3O^+(aq)$$

The reaction is illustrated by the schematic presented in Fig. 19.

In a strongly acidic electrolyte, $[Al(H_2O)_6]^{3+}(aq)$ is the dominant species [61]. With the increase of pH, $[Al(H_2O)_5(OH)]^{2+}(aq)$ further undergoes dissociation to produce $H_3O^+(aq)$ via additional deprotonation, until the pH enters the neutral range, when the insoluble species of $[Al(H_2O)_3(OH)_3](s)$ precipitates out via the reaction of

$$\left[Al(H_2O)_4(OH)_2\right]^{1+}(aq) + H_2O(l) \rightarrow \left[Al(H_2O)_3(OH)_3\right](s) + H_3O^+(aq).$$

This sequence of hydrolysis by Al^{3+} is frequently simplified as

$$Al^{3+}(aq) + 3H_2O\ (l) = Al(OH)_3(s) + 3H^+(aq),$$

or

$$Al^{3+}(aq) + 6H_2O(l) = Al(H_2O)_3(OH)_3(s) + 3H^+(aq).$$

On the other hand, in an alkaline electrolyte, the dominant species is the soluble $[Al(OH)_4(\mathbf{H_2}O)_\mathbf{2}]^-$ (aq) that is produced via the following reaction of

$$\left[Al(H_2O)_6\right]^{3+}(aq) + 4OH^-(aq) \rightarrow \left[Al(H_2O)_2(OH)_4\right]^-(aq) + 4H_2O(l).$$

$[Al(H_2O)_2(OH)_4]^-$ is also written as $(AlO_2{}^-)(H_2O)_4$ [62, 63], which is labeled as $AlO_2{}^-$ for brevity in the Pourbaix diagrams presented in Figs. 15, 16 and 17. The stable species being $[Al(H_2O)_6]^{3+}(aq)$, $Al(H_2O)_3(OH)_3(s)$, $(AlO_2{}^-)(H_2O)_4(aq)$ in acidic, neutral, and alkaline electrolytes, respectively, is consistent with the Pourbaix diagrams presented in Figs. 15, 16 and 17.

Al^{3+} forms stronger bonds with H_2O than with Cl^-. This is indicated by the anhydrous $AlCl_3$ being hygroscopic, and the hydration of $AlCl_3$ is highly exothermic. When $AlCl_3$ reacts with water, the Cl^- ligands are replaced with H_2O molecules to form the aluminum hexahydrate $[Al(H_2O)_6]^{3+}$. The reaction can be described below.

$$AlCl_3(s) + 6H_2O\ (l) \rightarrow \left[Al(H_2O)_6\right]^{3+}(aq) + 3Cl^-(aq)$$

The three Cl^- ions are weakly bonded to the hexahydrate. The hexahydrate has an octahedral molecular geometry, with the central Al^{3+} surrounded by six H_2O ligand molecules as shown in Fig. 19, and further away from the hexahydrate center are the Cl^- counterions which are located beyond the first hydration shell of Al^{3+}, i.e., between the first and second hydration shells of Al^{3+}, and between the second shell and the bulk [64]. Since the Cl^- ions are loosely bonded to the hexahydrate, it is anticipated that the Cl^- ions possess significant probability to dissociate with the hexahydrates and re-participate in the oxidation of new Al atoms, and the repetition of the cycle underlies the catalysis of the Cl^- ions for the Cu ball bond corrosion.

In summary, pitting of MCRL starts with Cl^- ions adsorption on the passivation layer of MCRL, and the transport of both the Cl^- ions and molecular water through the passivation layer to reach the underlying MCRL. In addition, the adsorbed Cl^- ions lead to local dissolution and thinning of the passive film. There are also O atoms adsorbed on the passivation film. The Cl^- ions and molecular water are both Lewis base and tend to bond with Al^{3+}, and the adsorbed O atoms attract valence electrons from the Al atoms of MCRL, both of which cause the Al ionization. Aluminum hexahydrate $[Al(H_2O)_6]^{3+}$ subsequently forms, and three Cl^- ions are weakly bonded to the hexahydrate. The dominant species are $[Al(H_2O)_6]^{3+}(aq)$, $[Al(H_2O)_3(OH)_3](s)$ and $[Al(H_2O)_2(OH)_4]^-(aq)$ in the acidic, neutral, and alkaline ranges, respectively. The low bond energy between the aluminum hexahydrate $[Al(H_2O)_6]^{3+}$ and Cl^- ions make the Cl^- ions catalytic for the corrosion.

The microstructure data presented in Figs. 2, 6, 7, 8, 9, 10, 11, 12, 13 and 14, together with the geometric, structural, and electrochemical characteristics of Cu ball bonds described in Sect. 3.2. "Ingot stack to simulate Cu ball bonds and the limitations" indicate the pitting started at interface IV. A similar observation in a study of AA 2024-T3 Al alloy (with weight percent of 90.7–94.7 of Al, < 0.1 of Cr, 3.8–4.9 of Cu, < 0.5 of Fe, 1.2–1.8 of Mg, 0.3–0.9 of Mn, < 0.5 of Si, < 1.5 of Ti, and < 0.25 of Zn) immersed in 0.1 M NaCl at $pH = 4.2$. The S-phase (Al_2CuMg) was dealloyed, leaving Cu-rich remnants that were cathodic to the matrix. Pit formation took place along the anodic peripheral around the Cu-rich cathodic remnants [65]. In the galvanic corrosion, a potential difference is developed between the cathode and the anode to drive the galvanic current. The resistance of the electrolyte segment across the interface of the galvanic couple in the corrosion current circuit is the smallest, which makes the local galvanic current higher than those further away from the interface. In other words, the corrosion attack on the anode decreases with increasing distance from the interface. As the electrolyte resistivity becomes higher, this maximum galvanic current at the interface of the galvanic couple is preferentially elevated higher [66]. The pitting being located along the anodic peripheral reported in reference [65] and the further increase of the maximum galvanic current density at the interface of the galvanic couple caused by the increase of electrolyte resistivity are both consistent with our observation of the corrosion of the MCRL along interface IV.

6.2 Advanced Stage of Corrosion—Crevice Corrosion/SCC

6.2.1 The Corrosion Enhancement in the Crevice

As the pit grows, the corroded MCRL along interface IV becomes a crevice, and the pitting corrosion transitions to a crevice corrosion [43]. The dissolved Al^{3+} ions are largely confined within the crevice and accumulate. The increased $[Al^{3+}]$ and hydrolysis of the Al^{3+} ions lead to the production of more H^+ ions, which elevates the local acidity. The higher acidity enhances the corrosion since Al^{3+}, instead of $Al(OH)_3$ or Al_2O_3, becomes the stable species, as shown in Figs. 15, 16 and 17.

The accumulation of Al^{3+} ions and H^+ ions within the crevice attracts Cl^- ions and OH^- ions from outside of the crevice to maintain charge neutrality, which also increases the inward transport speeds of Cl^- and OH^- anions (the transport of OH^- is, however, slower than that of Cl^-, since the size of OH^- is larger than that of Cl^-). The subsequent increase of $[Cl^-]$ in the crevice enhances not only the oxidation of the Al atoms in the MCRL kinetically, but also the corrosion of Cu thermodynamically which is indicated by downshift of the Cu immunity region boundary line by 0.08 V as shown in Fig. 17.

The $[O_2]$ in the crevice decreases due to the oxygen reduction reaction (ORR) in the crevice. As a result, some of the electrons generated from the anodic reaction are consumed by the ORR outside of the crevice while the rest by the HER in the crevice, both of which sustain the corrosion of the MCRL.

6.2.2 Crevice Corrosion Enhancement by Al Smears

As a ball bonding defect, the Al bond pad smear, written as Al smear for brevity, once created, exacerbates the crevice corrosion. A different phrase of “Al splash” has been used for the Al smear in literature. The Al smear and the Cu ball form a crevice with the tip of the crevice at the bond perimeter. Crevice corrosion can take place in this crevice, as shown in Fig. 2 of reference [67], which is now presented in Fig. 20. The Al smear is labeled “Narrow channel” in the figure. The figure also shows that a decrease of the Al smear had indeed abated the corrosion.

6.2.3 SCC and the Causes

Before the corrosion, the weakest interface was interface III, since it contained non-bonding materials, including surface oxides on the bond pad and the Cu ball, and processing residue. Interface III therefore had the lowest adhesion density among the four interfaces. After the reliability test, however, the crack was not along interface III; instead, it was along interface IV, as shown Fig. 6 of reference [17], Figs. 12 and 14. The observation indicates the SCC nature of the corrosion along interface IV. The causes of SCC are listed as follows.

Geometry and Materials Around the Bond Perimeter

As stated in Sect. 3.2. "Ingot stack to simulate Cu ball bonds and the limitations," the perimeter of the bond interface is a boundary line of different phases of the EMC, the Cu ball, the IMCs, and the bond pad. More broadly the packaged IC is a highly inhomogeneous heterostructure that consists of many different phases such as the Si substrate/epitaxial Si, various dielectrics, silicide, alloys, metallic films, and adhesive. Due to the geometric, structural, and electrochemical characteristics of Cu ball bonds listed in Sect. 3.2, pitting started, followed by crevice corrosion, in the MCRL along the MCRL/Cu interface. An external load caused by the interactions between some of these different phases can lead to the magnification of a local stress at the pit/crevice, and result in SCC.

Hydrogen Embrittlement

Inside the pit and crevice, the [H^+] is high. En route to the HER, hydrogen atoms adsorbed on the cathode surface are produced via

$$H^+ + e = H_{ads}.$$

The adsorbed H atoms can enter the cathode lattice and cause hydrogen embrittlement of the cathode [43], which assists the SCC along interface IV. Hydrogen gas evolution due to corrosion with moisture on IMCs has been extensively documented and is one of the known causes of IMC embrittlement [68]. Hydrogen embrittlement and the resulted SCC in Cu ball bonds were hypothesized [69], and hydrogen gas evolution had been detected from the corrosion of Cu ball bonds immersed in a 20 ppm Cl^-/pH = 5 solution [70].

Stress-Sorption Cracking

The chemisorption of the Cl^- ions on the surfaces of the crack decreases the surface energy, which favors the crack formation. The threshold stress for cracking is consequently decreased [71]. This is because cracking requires energy input to create new surfaces; on the other hand, the cracking releases strain energy from the bulk that surrounds the crack. The total energy change for the crack formation equals the (positive) energy input to create the new surfaces minus the strain energy released from the bulk surrounding the crack. A decrease of the energy to create the fresh surfaces favors the cracking. The reader can refer to Sect. 8.3.4 Stress Sorption of reference [72] for more information.

6.2.4 Propagation of the Crevice/Crack Across the Bond

At the tip of the corrosion crevice where the Cu ball is still in contact with the MCRL, the corrosion of the MCRL along interface IV contains the following components.

- the galvanic corrosion of the MCRL-Cu galvanic couple,
- lattice strain-assisted corrosion, and
- likely SCC contributed by hydrogen embrittlement and stress-sorption cracking.
- the self-corrosion in the form of dealloying,

When the Cu ball is no longer galvanically coupled to the MCRL, only the self-corrosion component is active for the MCRL. It is obvious that the lateral corrosion along interface IV propagates faster than the self-corrosion of the MCRL.

In literature, such crevices/cracks are frequently observed to have propagated laterally across the bond over distances of tens of microns as shown in Fig. 4 of reference [28], which is presented in Fig. 21.

The lateral depth of the crevice/crack into the bond is at least ~25 μm, and the maximum thickness of the crevice is about 0.5 μm as shown in the figure. The ratio of these two measurements is 25 μm/0.5 μm= 50. Our observation indicates an even large ratio. The maximum thickness of Cu_3Al_2 as shown in Fig. 6e is about 50 nm, while the length of the crack shown in Fig. 12a is 56 μm. The ratio is 56 μm/50 nm $\approx$1100. The fast propagation of the crevice aggravates the severity of reliability failure of the Cu ball bond.

A similar observation was reported in Fig. 3 of reference [30], which is shown in Fig. 22. Figure 22a, b each reveals not only the crevice/crack propagation along interface IV across the bond, but also an oxide matrix embedding Cu nanoparticles that resulted from the corrosion of the MCRL.

6.3 Cathodic Reactions

The O_2 dissolved in the electrolyte enables the cathodic reaction of ORR.

$$O_2 + 2H_2O + 4e^- = 4OH^-,$$

which is the dominant cathodic reaction in the initial stage of pitting. As the pitting proceeds, H^+ ions produced by the hydrolysis of Al^{3+} in the electrolyte enable HER of

$$2H^+ + 2e^- = H_2$$

inside the pit or the crevice when the pit or crevice has formed, while ORR still takes place outside of the pit or crevice. The HER in the corrosion of Al had been verified by the observation of H_2 blisters beneath the oxide film on Al [73].

7 Addressing the Corrosion

7.1 Decreasing Extractable Cl^- Ion Concentration of EMC

Since Cl^- ions cause the pitting corrosion and contribute to the crevice corrosion/SCC of Cu ball bonds, decreasing the extractable Cl^- ion concentration of the EMC can suppress the corrosion. The use of EMC with low concentration of extractable Cl^- ions of less than 20 ppm, and further less than 10 ppm, was reported to be successful to reduce failure rates of Cu ball bonds for automotive products [15]. With optimization, Cu ball bonds have exceeded the Joint Electron Device Engineering Council (JEDEC) reliability requirements by four to six times and achieve assembly yields close to those of Au ball bonds [14].

7.2 Pd-Coating of Cu Wires

Pd-coating of Cu bond wires has been deployed in the industry to address the corrosion issue [17, 74, 75]. The Pd-coated Cu (PCC) wires were demonstrated to have achieved a reliability that is higher than that of Cu wires and at the same level of Au wires [74]. The coating leads to the formation of a solid solution of Pd and Cu on the wire. This is because Pd and Cu obey the Hume-Rothery rules for solid solutions: (1) their atomic radii, 138 pm for Pd and 128 pm for Cu, differ by less than 15%, (2) their lattices are both face-center cubic, (3) their electronegativities, 2.2 for Pd and 1.9 for Cu differ by less than 30% (too large a difference in electronegativity will make the bond more ionic instead of metallic and result in the formation of an intermetallic compound instead of a solid solution), and (4) Pd and Cu are in two adjacent columns of the periodic table. The following are three mechanisms for Pd-coating to improve the resistance to the corrosion of Cu ball bonds.

1. To decrease the (Cu + Pd)/Al atomic ratio at interface IV. The Pd enhances atomic bonds in both the Cu-Pd solid solution and $(CuPd_x)Al$ [76]. The atomic bond enhancement for $(CuPd_x)Al$ increases its thermodynamic stability, which impedes the phase transformation of $(CuPd_x)Al$ to the next Cu-richer phase of $(CuPd_x)_3Al_2$ [76]. Consequently, with the PCC the MCRL was $(CuPd_x)Al$ along about 80% of the bond interface. This contrasted the MCRL being Cu_3Al_2 with the Cu wire. Images of the bond interfaces with PCC wire and Cu wire were shown in Figs. 2 and 1 of reference [76] and presented in Figs. 23 and 24, respectively. Simulation consistently showed Pd had inhibited the formation of Cu_3Al_2 [75]. The decrease of the (Cu + Pd)/Al atomic ratio not only decreases the anodic current of the exposed individual Al atoms, but also enhances the passivation of the MCRL over that of the MCRL under the Cu ball. The Al atomic percent is 50% in $(CuPd_x)Al$, which is higher than that of 40% in Cu_3Al_2. Since the passivity of the IMC is mainly from that of Al in the initial stage of the

corrosion, the passivity of ($CuPd_x$)Al is higher than that of Cu_3Al_2. The higher passivity causes the decrease of the corrosion rate.

2. To likely increase the areal fraction of interface IV in the bond area. Pd is nobler than Cu. Pd resists oxidation in ambient air [77]; in contrast, Cu is oxidized to Cu_2O in ambient air [78]. Consequently, the surface oxide on the FAB of the PCC is likely thinner than that on the Cu wire. While there has been no direct evidence of the thinner surface oxide on the FAB of the PCC wire, the suppressed oxidation can be inferred from the relaxed demand of the shielding environment for the FAB formation; i.e., N_2 can be used for the PCC, which contrasts the need of forming gas with 5% H_2 and 95% N_2 for Cu wires [79, 80]. The thinner surface oxide of the FAB can increase the areal fraction of intimate contact between the FAB and the bond pad, subsequently increase the areal fraction of ($CuPd_x$)Al formation in the bond area, and equivalently that of interface IV. Since it takes longer for the corrosion to propagate through the larger interface IV, the PCC ball bond is more resistant to the corrosion.

On the other hand, Pd catalyzes ORR [81, 82] and exhibited an exchange current density for the HER that is higher than that on Cu [83], which indicate that the Pd presence in the PCC enhances both ORR in the initial stage of the corrosion with the neutral electrolyte, and the HER in the advanced stage of the corrosion with the acidic electrolyte. The enhancement of ORR by the Pd addition to bulk Cu in neutral electrolyte has been verified [84]. The net result of improved corrosion resistance for the ball bonds has been reported in references of [17, 74, 75].

.

7.3 Post-bond Heating

In addition to the mature methods described in Sects. 7.1 and 7.2, a new approach that involves post-bond heating has been assessed [30]. The post-bond heating was carried out after the ball bonding and before the encapsulation with EMC, to increase the areal fraction of MCRL, and relieve the residual strain introduced by the ball bonding, both of which enhanced the resistance to the corrosion. Figure 25 presents SEM images of the bond interfaces after the post-bond heating, which shows the areal fraction of MCRL had increased with the post-bond heating. Consistently, the ball shear strength after biased-HAST (bHAST) had been improved.

7.4 Application of Corrosion Inhibitor

Ross et al. have demonstrated the decrease of the corrosion rate of exposed Cu ball bonds (that had not been encapsulated with EMC) with an inhibitor to suppress the

HER on Cu [70]. The inhibitor was introduced to a 20 ppm Cl^- solution in which the Cu ball bonds were immersed. The authors reported the application of the inhibitor for encapsulated Cu ball bonds was in development.

8 Similar Pitting and Crevice Corrosion in Au Ball Bonds

Literature on the corrosion of Au ball bonds has been less than that of Cu ball bonds. Muller reported the corrosion to be similarly along the Au_4Al/Au interface and proceeded downward to consume Au_4Al, as shown in Fig. 10 of [85] which is presented in Fig. 26.

The characteristics #2 to #6 listed for Cu ball bonds are also possessed by the Au ball bonds shown in Fig. 26. These characteristics are

1. The Au_8Al_3/Au_4Al interface is similar to interface III in the Cu ball bond in that it contains voids, the surface oxide of the bond pad, and the processing residue from the bond pad opening. For brevity, this interface is also labeled interface III. Interface III is the most resistive interface across the bond. The Au_8Al_3-Au_4Al galvanic couple is the weakest couple, where Au_4Al is less cathodically protected by Au_8Al_3, and Au_8Al_3 less anodically corroded. The high Ohmic resistance of the thin-film electrolyte limits the spatial range of galvanic corrosion. Therefore, the following three galvanic couples of Al-$AuAl_2$, $AuAl_2$-Au_8Al_3, and Au_4Al-Au have the highest galvanic corrosion currents. Since Au_4Al was less cathodically protected, it was more vulnerable to corrosion.
2. The lattice strain in the Au_4Al layer along the interface of the Au_4Al-Au couple, and residual strain introduced by the ball bonding, contributed to the corrosion of Au_4Al.
3. The perimeter of the bond interface is the boundary of multiple phases of the Au ball, the Au-Al IMCs, the Al alloy pad, and the EMC. The local stress due to an external load is magnified locally at the perimeter, which can lead to SCC when pitting has taken place along the Au_4Al/Au interface.
4. The surface area ratio of cathode/anode of the Au_4Al-Au couple is higher than those of the Al-$AuAl_2$ and $AuAl_2$-Au_8Al_3 couples, which makes the corrosion current density of Au_4Al the highest.
5. The Au/Al atomic ratio at the Au_4Al/Au interface is the highest.
6. The passivity of a Au-Al IMC in a near-neutral electrolyte is mainly due to that of Al in the IMC, i.e., the $Al(OH)_3/Al_2O_3$ precipitated on the surface of the IMC that results from the oxidation of the Al. The Al content is the lowest in Au_4Al, which makes the passivity of Au_4Al the lowest among the Au-Al IMCs, and subsequently, the Au_4Al the most vulnerable to the corrosion in the initial stage of the corrosion in which the electrolyte is near-neutral.

These five characteristics determined pitting corrosion, and subsequently crevice corrosion, took place in Au_4Al along the Au_4Al/Au interface. SCC along the same interface could take place after the pit or the crevice had formed. Nonetheless, Au ball bonds possess some differences from Cu ball bonds. One lies in the IMC growth rate in Au ball bonds being much higher than that in Cu ball bonds [86]. Differences in the corrosion mechanisms can arise. Therefore, an analysis of a larger set of data is needed to ensure comprehensiveness of the understanding.

9 Conclusions

Wire bonding is the most widely used interconnection method in semiconductor packaging. It has enabled many modern technologies. In the automotive industry, driving automation and advanced driver assistance systems mainly for safety enhancement are gaining momentum. The reliabilities of these technologies can be achieved only when the underlying ball bonds are reliable. The corrosion of Cu ball bonds can cause reliability failures. The corrosion mechanisms have been studied quite independently based on microstructure characterization and electrochemical analyses. This chapter is aimed to unify these two approaches, to develop a more comprehensive understanding of the mechanisms.

The corrosion potentials of the phases of Cu, Cu_9Al_4, CuAl, $CuAl_{2,}$ and Al decrease with the decrease of Cu content in the phase, in the order of Cu > Cu_9Al_4 > CuAl > $CuAl_2$ > Al. The most Cu-rich layer (MCRL) is corroded along the MCRL/Cu ball interface. The corrosion begins with pitting and evolves into crevice corrosion that produces hydrated Al^{3+} ions. The Cu in the MCRL is mostly transformed into Cu nanoparticles, which both self-coalesce and coalesce with the Cu ball to decrease the surface energy. Consequently, the Cu nanoparticles are separated from the residual MCRL. Stress corrosion cracking (SCC) can accompany the corrosion.

Pourbaix diagrams of the Al–Cu–Cl–H_2O system indicate that the increase of electric bias enhances the corrosion, from immunity toward corrosion. The molar ratio of the oxidized Cu over the oxidized Al in the MCRL is very small. The corrosion is also enhanced at the high temperature of 125 °C with a partial passivity breakdown of Al and the IMCs in the pH of 0–14. With pH < 3, the Al is oxidized to $Al^{3+}(aq)$. With near-neutral to neutral electrolyte, $Al(OH)_3(s)$ or $Al_2O_3(s)$ forms. The high Cl^- concentration enhances the corrosion kinetically, and expands the corrosion region of Cu thermodynamically.

In the initial stage of the corrosion, the passivity of the MCRL is mainly due to that created by the oxidation of the Al in the MCRL. This passivity of the MCRL is lower than that of Al metal since the passivation film of the MCRL is more defective than that of Al. In neutral solutions, the surface of the passivation film of the MCRL acquires a positive charge. Cl^- ions are therefore attracted to and adsorbed on the MCRL. Pitting corrosion ensues. Three main mechanisms are possible to cause the pitting. They are (i) the penetration mechanism, (ii) the film thinning mechanism, and (iii) the film rupture mechanism. The higher defectivity of the passivation film of the MCRL, compared with that of Al, enhances all three mechanisms. With pitting to break down the passivation film on the MCRL, oxidation of the Al in the MCRL follows.

Three factors to drive the Al oxidation of the MCRL are: (1) adsorbed oxygen atoms attract the valence electrons of the Al of the MCRL, (2 & 3) The Cl^- and H_2O are both Lewis bases tending to bond with Al^{3+}. The subsequent H_2O-complexation of Al^{3+} leads to the formation of complex ion of $[Al(H_2O)_6]^{3+}(aq)$. The Cl^- counterions are located beyond the first hydration shell of the Al^{3+} ion.

In the advanced stage, the corrosion transitions into crevice corrosion, which is characteristic of low pH and high $[Cl^-]$, both of which enhance the corrosion. Al smear created as a ball bonding defect forms a crevice with the Cu ball, which exacerbates the corrosion. The perimeter of the bond is a multiphase boundary where an external load can lead to a local stress magnification, and subsequently SCC when either the pit or the crevice has formed. The adsorbed hydrogen atoms generated from hydrolysis of Al^{3+} ions and chemisorbed Cl^- ions can lead to hydrogen embrittlement and stress-sorption cracking, respectively, both of which contribute to the SCC.

The dominant cathodic reaction is oxygen reduction reaction (ORR) in the initial stage of pitting. As the pit grows and develops into a crevice, hydrogen evolution reaction (HER) becomes the dominant cathodic reaction inside the pit and the crevice, while ORR still takes place outside of the pit and crevice.

In the industry, decreasing the extractable Cl^- ion concentration of the EMC to below 20 ppm and Pd-coating of the Cu wires have been used to address the corrosion. Post-bond heating to increase the areal fraction of the MCRL and relieve stress introduced by the ball bonding process was reported to have improved the corrosion resistance. The use of an inhibitor to suppress the HER on Cu also decreased the corrosion rate of Cu ball bonds that had not been encapsulated with EMC.

Au ball bonds are similar to Cu ball bonds in terms of the structural, geometric, and electrochemical characteristics. Crevice corrosion in a Au ball bond was similarly along the Au_4Al/Au interface and proceeding downward to consume Au_4Al, which suggests the corrosion mechanisms to be similar to those of Cu ball bonds.

Acknowledgements The author thanks Denise Barrientos for the SEM images of the Cu wire bonds, Jeff Gambino, Jim Mohr, Eric Barbian and Y.Y. Tan for their helpful review of the manuscript.

References

1. G. G. Harman, “Wire Bonding in Microelectronics, Third Edition,” p. 426, 2010, Accessed: Dec. 03, 2021. [Online]. Available: https://www.accessengineeringlibrary.com/content/book/9780071476232.

2. C. D. Breach and F. W. Wulff, "A brief review of selected aspects of the materials science of ball bonding," *Microelectronics Reliability*, vol. 50, no. 1, pp. 1–20, Jan. 2010, doi: https://doi.org/10.1016/J.MICROREL.2009.08.003.
3. "Wire Bonding Market - Global Industry Analysis, Size, Share, Growth, Trends, and Forecast, 2021–2031," Nov. 2021. Accessed: Jan. 10, 2022. [Online]. Available: https://www.researchandmarkets.com/reports/5510909/wire-bonding-market-global-industry-analysis?utm_source=GNOM&utm_medium=PressRelease&utm_code=z8frbj&utm_campaign=1642757+-+Global+Wire+Bonding+Market+(2021+to+2031)+-+Industry+Analysis%2c+Size%2c+Share%2c+Growth%2c+Trends%2c+and+Forecasts&utm_exec=jamu273prd.
4. D. Y. JUNG, S. R. CAIN, and W. T. CHEN, "Introduction to Wire Bond Technology," in *Encyclopedia of Packaging Materials, Processes, and Mechanics*, 2019, pp. 1–21. doi: https://doi.org/10.1142/9789811209666_0001.
5. M. Lapedus, "Wirebond Technology Rolls On," *semiengineering.com*, May 2107. https://semiengineering.com/wirebond-technology-rolls-on/ (accessed Jan. 10, 2022).
6. C. L. Gan and U. Hashim, "Evolutions of bonding wires used in semiconductor electronics: perspective over 25 years," *Journal of Materials Science: Materials in Electronics 2015 26:7*, vol. 26, no. 7, pp. 4412–4424, Mar. 2015, doi: https://doi.org/10.1007/S10854-015-2892-8.
7. S. Murali, N. Srikanth, Y. M. Wong, and C. J. Vath, "Fundamentals of thermo-sonic copper wire bonding in microelectronics packaging," *Journal of Materials Science 2006 42:2*, vol. 42, no. 2, pp. 615–623, Nov. 2006, doi: https://doi.org/10.1007/S10853-006-1148-7.
8. C. D. Breach, "What is the future of bonding wire? Will copper entirely replace gold?," *Gold Bulletin 2010 43:3*, vol. 43, no. 3, pp. 150–168, 2010, doi: https://doi.org/10.1007/BF03214983.
9. J. Ramsey, "Semiconductor-based electronics make up 40% of a car's cost," *autoblog.com*, May 11, 2020. https://www.autoblog.com/2020/05/11/car-electronics-cost-semiconductor-chips/ (accessed Jan. 10, 2022).
10. D. Ferris, "Chip Shortage Threatens Biden's Electric Vehicle Plans, Commerce Secretary Says - Scientific American," www.scientificamerica.com, Nov. 2021. https://www.scientificamerican.com/article/chip-shortage-threatens-bidens-electric-vehicle-plans-commerce-secretary-says/ (accessed Jan. 10, 2022).
11. B. Hellenthal, "Power electronics - Key to the next level of automotive electrification," in *Proceedings of the International Symposium on Power Semiconductor Devices and ICs*, 2012, pp. 13–16. doi: https://doi.org/10.1109/ISPSD.2012.6229011.
12. "Automated Vehicles for Safety | NHTSA," www.nhtsa.gov. https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety (accessed Jan. 15, 2022).
13. M. Lu, K. Wevers, and R. van der Heijden, "Technical Feasibility of Advanced Driver Assistance Systems (ADAS) for Road Traffic Safety," *Transportation Planning and Technology*, vol. 28, no. 3, pp. 167–187, Jun. 2007, doi: https://doi.org/10.1080/03081060500120282.
14. B. K. Appelt, A. Tseng, L. Huang, and S. Chen, "Is copper wire bonding ready for automotive applications?," in *2011 IEEE 13th Electronics Packaging Technology Conference, EPTC 2011*, 2011, pp. 387–390. doi: https://doi.org/10.1109/EPTC.2011.6184451.
15. J. McLeish and R. Schueller, "Ensuring Suitability of Cu Wire Bonded ICs for Automotive Applications," *International Symposium on Microelectronics*, vol. 2015, no. 1, pp. 000751–000756, Oct. 2015, doi: https://doi.org/10.4071/ISOM-2015-POSTER8.
16. S. H. Kim, J. W. Park, S. J. Hong, and J. T. Moon, "The interface behavior of the Cu-Al bond system in high humidity conditions," in *2010 12th Electronics Packaging Technology Conference, EPTC 2010*, 2010, pp. 545–549. doi: https://doi.org/10.1109/EPTC.2010.5702699.
17. T. Uno and T. Yamada, "Improving humidity bond reliability of copper bonding wires," in *Proceedings - Electronic Components and Technology Conference*, 2010, pp. 1725–1732. doi: https://doi.org/10.1109/ECTC.2010.5490741.
18. T. Boettcher *et al.*, "On the intermetallic corrosion of Cu-Al wire bonds," in *2010 12th Electronics Packaging Technology Conference, EPTC 2010*, 2010, pp. 585–590. doi: https://doi.org/10.1109/EPTC.2010.5702706.

19. H. Xu *et al.*, "Behavior of aluminum oxide, intermetallics and voids in Cu–Al wire bonds," *Acta Materialia*, vol. 59, no. 14, pp. 5661–5673, Aug. 2011, doi: https://doi.org/10.1016/J.ACTAMAT.2011.05.041.
20. J. Osenbach, B. Q. Wang, S. Emerich, J. Delucca, and D. Meng, "Corrosion of the Cu/Al interface in Cu-Wire-bonded integrated circuits," in *Proceedings - Electronic Components and Technology Conference*, 2013, pp. 1574–1586. doi: https://doi.org/10.1109/ECTC.2013.6575782.
21. V. Mathew, E. Wikramanayake, and S. F. Chopin, "Corrosion of Copper Wire bonded Packages by Chlorine Containing Foreign Particles," in *Proceedings - Electronic Components and Technology Conference*, Jun. 2020, vol. 2020-June, pp. 504–511. doi: https://doi.org/10.1109/ECTC32862.2020.00086.
22. Y. Luo, "Identification and Predictive Modeling of High Propensity of Defects and Field Failure in Copper-aluminum Wire Bond Interconnect under Exposure to High Temperature and Humidity," 2018. Accessed: Dec. 05, 2021. [Online]. Available: https://etd.auburn.edu//handle/10415/6382.
23. W. Qin, T. Anderson, H. Anderson, G. Chang, and D. Barrientos, "Corrosion mechanisms of Cu Wire Bonding on Al Pads," in *Proceedings - Electronic Components and Technology Conference*, 2018, vol. 2018-May, pp. 1446–1454. doi: https://doi.org/10.1109/ECTC.2018.00221.
24. J. Idrac, G. Mankowski, G. Thompson, P. Skeldon, Y. Kihn, and C. Blanc, "Galvanic corrosion of aluminium–copper model alloys," *Electrochimica Acta*, vol. 52, no. 27, pp. 7626–7633, Oct. 2007, doi: https://doi.org/10.1016/J.ELECTACTA.2007.05.056.
25. A. v. Benedeti, P. T. A. Sumodjo, K. Nobe, P. L. Cabot, and W. G. Proud, "Electrochemical studies of copper, copper-aluminium and copper-aluminium-silver alloys: Impedance results in 0.5M NaCl," *Electrochimica Acta*, vol. 40, no. 16, pp. 2657–2668, Nov. 1995, doi: https://doi.org/10.1016/0013-4686(95)00108-Q.
26. A. B. Y. Lim, W. J. Neo, O. Yauw, B. Chylak, C. L. Gan, and Z. Chen, "Evaluation of the corrosion performance of Cu–Al intermetallic compounds and the effect of Pd addition," *Microelectronics Reliability*, vol. 56, pp. 155–161, Jan. 2016, doi: https://doi.org/10.1016/J.MICROREL.2015.10.012.
27. Y. Wu and A. Lee, "Corrosion-Induced Mass Loss of Cu9Al4 at the Cu-Al Ball–Bond Interface: Explained Based on Full Immersion of Cu, Al, and Cu-Al Intermetallic Galvanic Couples," *Journal of Electronic Materials 2018 48:1*, vol. 48, no. 1, pp. 44–52, Sep. 2018, doi: https://doi.org/10.1007/S11664-018-6625-7.
28. K. Zeng and A. Nangia, "Effect of high temperature bake on evolution of interfacial structure in Cu wire bonds and its impact on Cu/Al interfacial corrosion," in *Proceedings - Electronic Components and Technology Conference*, Jul. 2015, vol. 2015-July, pp. 1586–1593. doi: https://doi.org/10.1109/ECTC.2015.7159808.
29. T. Uno and K. Tatsumi, "Thermal reliability of gold–aluminum bonds encapsulated in bi-phenyl epoxy resin," *Microelectronics Reliability*, vol. 40, no. 1, pp. 145–153, Jan. 2000, doi: https://doi.org/10.1016/S0026-2714(99)00087-6.
30. M. Eto, N. Araki, T. Yamada, M. Sugiyama, and S. Fujimoto, "Influence of post-bonding heating process on the long-term reliability of Cu/Al contact," *Microelectronics Reliability*, vol. 118, p. 114058, Mar. 2021, doi: https://doi.org/10.1016/J.MICROREL.2021.114058.
31. H. Zhou, W. Yao, C. Du, S. Song, and R. Wu, "Corrosion Behavior of the Al 2 Cu Intermetallic Compound and Coupled Al 2 Cu/Al," *International Journal of Electrochemical Science*, vol. 12, pp. 9542–9554, 2017, doi: https://doi.org/10.20964/2017.10.32.
32. J. R. Scully, T. O. Knight, R. G. Buchheit, and D. E. Peebles, "Electrochemical characteristics of the Al2Cu, Al3Ta and Al3Zr intermetallic phases and their relevancy to the localized corrosion of Al alloys," *Corrosion Science*, vol. 35, no. 1–4, pp. 185–195, Jan. 1993, doi: https://doi.org/10.1016/0010-938X(93)90148-A.
33. S. Thomas and H. M. Berg, "Micro-Corrosion of Al–Cu Bonding Pads," *IEEE Transactions on Components, Hybrids, and Manufacturing Technology*, vol. 10, no. 2, pp. 252–257, 1987, doi: https://doi.org/10.1109/TCHMT.1987.1134741.

34. E. J. W. Wensink, A. C. Hoffmann, M. E. F. Apol, and H. J. C. Berendsen, “Properties of Adsorbed Water Layers and the Effect of Adsorbed Layers on Interparticle Forces by Liquid Bridging,” *Langmuir*, vol. 16, no. 19, pp. 7392–7400, Sep. 2000, doi: https://doi.org/10.1021/LA000009E.
35. R. Ambat and K. Piotrowska, *HUMIDITY AND ELECTRONICS corrosion reliability issues and preventive measures.* WOODHEAD PUBLISHING UK, 2021.
36. G. W. Warren, P. Wynblatt, and M. Zamanzadeh, “The role of electrochemical migration and moisture adsorption on the reliability of metallized ceramic substrates,” *Journal of Electronic Materials 1989 18:2*, vol. 18, no. 2, pp. 339–353, Mar. 1989, doi: https://doi.org/10.1007/BF02657426.
37. S. Goldberg, J. A. Davis, and J. D. Hem, “The Surface Chemistry of Aluminum Oxides and Hydroxides,” in *The Environmental Chemistry of Aluminum*, G. Sposito, Ed. CRC Press, 1996, pp. 271–331. doi: https://doi.org/10.1201/9780138736781-7.
38. G. L. Song, “Potential and current distributions of one-dimensional galvanic corrosion systems,” *Corrosion Science*, vol. 52, no. 2, pp. 455–480, Feb. 2010, doi: https://doi.org/10.1016/J.CORSCI.2009.10.003.
39. A. Abbadie, J.-M. Hartmann, F. Brunier, A. Abbadie, J.-M. Hartmann, and F. Brunier, “A Review of Different and Promising Defect Etching Techniques: from Si to Ge,” *ECSTr*, vol. 10, no. 1, p. 3, Dec. 2007, doi: https://doi.org/10.1149/1.2773972.
40. O. Seri, S. Furuya, and N. Soga, “Effect of copper content on corrosion of aluminumアルミニウムの腐食に及ぼす銅含有量の効果,” *Journal of Japan Institute of Light Metals軽金属*, vol. 39, no. 10, pp. 724–729, Oct. 1989, doi: https://doi.org/10.2464/JILM.39.724.
41. D. E. J. Talbot and J. D. R. Tabolt, *Corrosion Science and Technology, Third Edition.* CRC Press, 2018. doi: https://doi.org/10.1201/9781351259910/CORROSION-SCIENCE-TECHNOLOGY-DAVID-TALBOT-JAMES-TALBOT.
42. M. H. Lue, C. T. Huang, S. T. Huang, and K. C. Hsieh, “Bromine- and chlorine-induced degradation of gold-aluminum bonds,” *Journal of Electronic Materials 2004 33:10*, vol. 33, no. 10, pp. 1111–1117, 2004, doi: https://doi.org/10.1007/S11664-004-0112-Z.
43. Z. Ahmad, “Principles of Corrosion Engineering and Corrosion Control,” *Principles of Corrosion Engineering and Corrosion Control*, 2006, doi: https://doi.org/10.1016/B978-0-7506-5924-6.X5000-4.
44. J. L. Murray, “The aluminium-copper system,” *International Metals Reviews*, vol. 30, no. 1, pp. 211–234, Jan. 2013, doi: https://doi.org/10.1179/IMTR.1985.30.1.211.
45. Y. H. Lu, Y. W. Wang, B. K. Appelt, Y. S. Lai, and C. R. Kao, “Growth of CuAl intermetallic compounds in Cu and Cu(Pd) wire bonding,” in *Proceedings - Electronic Components and Technology Conference*, 2011, pp. 1481–1488. doi: https://doi.org/10.1109/ECTC.2011.5898706.
46. Y. Zeng, K. Bai, and H. Jin, “Thermodynamic study on the corrosion mechanism of copper wire bonding,” *Microelectronics Reliability*, vol. 53, no. 7, pp. 985–1001, Jul. 2013, doi: https://doi.org/10.1016/J.MICROREL.2013.03.006.
47. M. Pourbaix, *Atlas of electrochemical equilibria in aqueous solutions.* Houston, Tex. : National Association of Corrosion Engineers, 1974.
48. James D.R. Talbot. and David E.J. Talbot., “Pourbaix (Potential-pH) Diagrams,” in *Corrosion Science and Technology*, 1998.
49. Corrosion-Doctor.org, “Aluminum E-pH (Pourbaix) diagram.” https://corrosion-doctors.org/Corrosion-Thermodynamics/Potential-pH-diagram-aluminum.htm (accessed Dec. 03, 2021).
50. C. Young-Gab, P. Su-Il, and K. Chang-Ha, “Effect of aluminium content on the anodic behaviour of copper-aluminium alloys in 3.5 wt% NaCl solution,” *Materials Letters*, vol. 20, no. 5–6, pp. 265–270, Aug. 1994, doi: https://doi.org/10.1016/0167-577X(94)90027-2.
51. B. Mazurkiewicz and A. Piotrowski, “The electrochemical behaviour of the Al_2Cu intermetallic compound,” *Corrosion Science*, vol. 23, no. 7, pp. 697–707, Jan. 1983, doi: https://doi.org/10.1016/0010-938X(83)90034-3.
52. P. M. Natishan, E. McCafferty, and G. K. Hubler, “Surface Charge Considerations in the Pitting of Ion-Implanted Aluminum,” *Journal of The Electrochemical Society*, vol. 135, no. 2, pp. 321–327, Feb. 1988, doi: https://doi.org/10.1149/1.2095608/XML.

53. G. Salek, C. Tenailleau, P. Dufour, and S. Guillemet-Fritsch, “Room temperature inorganic polycondensation of oxide (Cu_2O and ZnO) nanoparticles and thin films preparation by the dip-coating technique,” *Thin Solid Films*, vol. 589, pp. 872–876, Aug. 2015, doi: https://doi.org/10.1016/J.TSF.2015.04.082.
54. S. Kittaka and T. Morimoto, “Isoelectric point of metal oxides and binary metal oxides having spinel structure,” *Journal of Colloid and Interface Science*, vol. 75, no. 2, pp. 398–403, Jun. 1980, doi: https://doi.org/10.1016/0021-9797(80)90464-6.
55. E. McCafferty, “Sequence of steps in the pitting of aluminum by chloride ions,” *Corrosion Science*, vol. 45, no. 7, pp. 1421–1438, Jul. 2003, doi: https://doi.org/10.1016/S0010-938X(02)00231-7.
56. P. M. Natishan, W. E. O’Grady, E. McCafferty, D. E. Ramaker, K. Pandya, and A. Russell, “Chloride Uptake by Oxide Covered Aluminum as Determined by X-Ray Photoelectron and X-Ray Absorption Spectroscopy,” *Journal of The Electrochemical Society*, vol. 146, no. 5, pp. 1737–1740, May 1999, doi: https://doi.org/10.1149/1.1391835/XML.
57. T. H. Nguyen and R. T. Foley, “The Chemical Nature of Aluminum Corrosion: III. The Dissolution Mechanism of Aluminum Oxide and Aluminum Powder in Various Electrolytes,” *Journal of The Electrochemical Society*, vol. 127, no. 12, pp. 2563–2566, Dec. 1980, doi: https://doi.org/10.1149/1.2129520/XML.
58. N. Sato, K. Kudo, and T. Noda, “The anodic oxide film on iron in neutral solution,” *Electrochimica Acta*, vol. 16, no. 11, pp. 1909–1921, Nov. 1971, doi: https://doi.org/10.1016/0013-4686(71)85146-0.
59. O. Kubaschewski, *Oxidation of metals and alloys*, 2d ed. London: Butterworths, 1962.
60. “2.7: Ions as Acids and Bases - Chemistry LibreTexts,” *Chemistry LibreTexts*, May 08, 2021. https://chem.libretexts.org/Courses/Mount_Royal_University/Chem_1202/Unit_2%3A_Acids_and_Bases/15.07%3A_Ions_as_Acids_and_Bases (accessed Dec. 04, 2021).
61. E. Koubek, C. McWherter, and G. L. Gilbert, “Acid-Base Chemistry of the Aluminum Ion in Aqueous Solution,” *Journal of Chemical Education*, vol. 75, no. 1, pp. 60–60, 1998, doi: https://doi.org/10.1021/ED075P60.1.
62. J. A. Tossell, “Theoretical studies on aluminate and sodium aluminate species in models for aqueous solution; $Al(OH)_3$, $Al(OH)^-{}_4$, and $NaAl(OH)_4$,” *American Mineralogist*, vol. 84, no. 10, pp. 1641–1649, Oct. 1999, doi: https://doi.org/10.2138/AM-1999-1019.
63. V. A. Pokrovskii and H. C. Helgeson, “Thermodynamic properties of aqueous species and the solubilities of minerals at high pressures and temperatures; the system Al_2O_3-H_2O-NaCl,” *American Journal of Science*, vol. 295, no. 10, pp. 1255–1342, Dec. 1995, doi: https://doi.org/10.2475/AJS.295.10.1255.
64. E. Cauët, S. A. Bogatko, E. J. Bylaska, and J. H. Weare, “Ion association in $AlCl_3$ aqueous solutions from constrained first-principles molecular dynamics,” *Inorganic Chemistry*, vol. 51, no. 20, pp. 10856–10869, Oct. 2012, doi: https://doi.org/10.1021/IC301346K/SUPPL_FILE/IC301346K_SI_001.PDF.
65. R. G. Buchheit, R. P. Grant, P. F. Hlava, B. Mckenzie, and G. L. Zender, “Local Dissolution Phenomena Associated with S Phase (Al_2CuMg) Particles in Aluminum Alloy 2024-T3,” *Journal of The Electrochemical Society*, vol. 144, no. 8, pp. 2621–2628, Aug. 1997, doi: https://doi.org/10.1149/1.1837874/XML.
66. E. Bardal, R. Johnsen, and P. O. Gartland, “Prediction of Galvanic Corrosion Rates and Distribution by Means of Calculation and Experimental Models,” *Corrosion*, vol. 40, no. 12, pp. 628–633, Dec. 1984, doi: https://doi.org/10.5006/1.3593898.
67. M. van Soestbergen, A. Mavinkurve, J. J. M. Zaal, G. M. Ohalloran, R. T. H. Rongen, and M. L. Farrugia, “Crevice Corrosion of Ball Bond Intermetallics of Cu and Ag Wire,” in *Proceedings - Electronic Components and Technology Conference*, Aug. 2016, vol. 2016-August, pp. 774–781. doi: https://doi.org/10.1109/ECTC.2016.192.
68. N. K. Othman, J. Zhang, and D. J. Young, “Water Vapour Effects on Fe–Cr Alloy Oxidation,” *Oxidation of Metals 2010 73:1*, vol. 73, no. 1, pp. 337–352, Oct. 2010, doi: https://doi.org/10.1007/S11085-009-9183-9.

69. C. D. Breach and T. K. Lee, "Conjecture on the chemical stability and corrosion resistance of Cu-Al and Au-Al intermetallics in ball bonds," in *2011 International Conference on Electronic Packaging Technology and High Density Packaging*, 2011, pp. 275–283. doi: https://doi.org/10.1109/ICEPT.2011.6066835.
70. N. Ross *et al.*, "Mechanistic study of copper wire-bonding failures on packaging devices in acidic chloride environments," *Microelectronics Reliability*, vol. 113, p. 113917, Oct. 2020, doi: https://doi.org/10.1016/J.MICROREL.2020.113917.
71. UHLIG HH, COOK EW, and JR, "MECHANISM OF INHIBITING STRESS CORROSION CRACKING OF 18–8 STAINLESS STEEL IN MGCL2 BY ACETATES AND NITRATES," *Electrochem Soc-J*, vol. 116, no. 2, pp. 173–177, Feb. 1969, doi: https://doi.org/10.1149/1.2411789/XML.
72. R. W. Revie and H. H. Uhlig, *Corrosion and Corrosion Control: An Introduction to Corrosion Science and Engineering: Fourth Edition*. John Wiley and Sons, 2008. doi: https://doi.org/10.1002/9780470277270.
73. C. B. Bargeron and R. C. Benson, "Analysis of the Gases Evolved during the Pitting Corrosion of Aluminum in Various Electrolytes," *Journal of The Electrochemical Society*, vol. 127, no. 11, pp. 2528–2530, Nov. 1980, doi: https://doi.org/10.1149/1.2129511/XML.
74. S. Kaimori, T. Nonaka, and A. Mizoguchi, "The development of Cu bonding wire with oxidation-resistant metal coating," *IEEE Transactions on Advanced Packaging*, vol. 29, no. 2, pp. 227–231, May 2006, doi: https://doi.org/10.1109/TADVP.2006.872999.
75. H. Abe *et al.*, "Cu wire and Pd-Cu wire package reliability and molding compounds," in *Proceedings - Electronic Components and Technology Conference*, 2012, pp. 1117–1123. doi: https://doi.org/10.1109/ECTC.2012.6248975.
76. W. Qin, T. Anderson, and G. Chang, "Mechanism to improve the reliability of copper wire bonding with palladium-coating of the wire," *Microelectronics Reliability*, vol. 99, pp. 239–244, Aug. 2019, doi: https://doi.org/10.1016/J.MICROREL.2019.04.010.
77. D. Horwat *et al.*, "Chemistry, phase formation, and catalytic activity of thin palladium-containing oxide films synthesized by plasma-assisted physical vapor deposition," *Surface and Coatings Technology*, vol. 205, no. SUPPL. 2, pp. S171–S177, Jul. 2011, doi: https://doi.org/10.1016/J.SURFCOAT.2010.12.021.
78. I. Platzman, R. Brener, H. Haick, and R. Tannenbaum, "Oxidation of polycrystalline copper thin films at ambient conditions," *Journal of Physical Chemistry C*, vol. 112, no. 4, pp. 1101–1108, Jan. 2008, doi: https://doi.org/10.1021/JP076981K/SUPPL_FILE/JP076981K-FILE001.PDF.
79. H. Clauberg, B. Chylak, N. Wong, J. Yeung, and E. Milke, "Wire bonding with Pd-coated copper wire," 2010. doi: https://doi.org/10.1109/CPMTSYMPJ.2010.5679678.
80. A. Rezvani, M. Mayer, A. Shah, N. Zhou, S. J. Hong, and J. T. Moon, "Free-air ball formation and deformability with Pd coated Cu wire," in *Proceedings - Electronic Components and Technology Conference*, 2011, pp. 1516–1522. doi: https://doi.org/10.1109/ECTC.2011.5898711.
81. F. Gobal and R. Arab, "A preliminary study of the electro-catalytic reduction of oxygen on Cu–Pd alloys in alkaline solution," *Journal of Electroanalytical Chemistry*, vol. 647, no. 1, pp. 66–73, Aug. 2010, doi: https://doi.org/10.1016/J.JELECHEM.2010.05.009.
82. M. Shao, "Palladium-based electrocatalysts for hydrogen oxidation and oxygen reduction reactions," *Journal of Power Sources*, vol. 196, no. 5, pp. 2433–2444, Mar. 2011, doi: https://doi.org/10.1016/J.JPOWSOUR.2010.10.093.
83. S. Trasatti, "Work function, electronegativity, and electrochemical behaviour of metals: III. Electrolytic hydrogen evolution in acid solutions," *Journal of Electroanalytical Chemistry and Interfacial Electrochemistry*, vol. 39, no. 1, pp. 163–184, Sep. 1972, doi: https://doi.org/10.1016/S0022-0728(72)80485-6.
84. Y. Wu, K. N. Subramanian, S. C. Barton, and A. Lee, "Electrochemical studies of Pd-doped Cu and Pd-doped Cu-Al intermetallics for understanding corrosion behavior in wire-bonding packages," *Microelectronics Reliability*, vol. 78, pp. 355–361, Nov. 2017, doi: https://doi.org/10.1016/J.MICROREL.2017.09.024.

85. T. Müller, L. Schräpler, F. Altmann, H. Knoll, and M. Petzold, “Influence of intermetallic phases on reliability in thermosonic Au-Al wire bonding,” in *ESTC 2006 - 1st Electronics System integration Technology Conference*, 2006, vol. 2, pp. 1266–1273. doi: https://doi.org/10.1109/ESTC.2006.280174.

86. S. Murali, N. Srikanth, and C. J. Vath, “An analysis of intermetallics formation of gold and copper ball bonding on thermal aging,” *Materials Research Bulletin*, vol. 38, no. 4, pp. 637–646, Mar. 2003, doi: https://doi.org/10.1016/S0025-5408(03)00004-7.