Rasit O. Topaloglu *Editor*

More than Moore Technologies for Next Generation Computer Design

Springer

More than Moore Technologies for Next Generation Computer Design

Rasit O. Topaloglu
Editor

More than Moore Technologies for Next Generation Computer Design

Springer

Editor
Rasit O. Topaloglu
IBM
Hopewell Junction, NY, USA

ISBN 978-1-4939-2162-1 ISBN 978-1-4939-2163-8 (eBook)
DOI 10.1007/978-1-4939-2163-8

Library of Congress Control Number: 2015931070

Springer New York Heidelberg Dordrecht London

Printed on acid-free paper

Springer Science+Business Media LLC New York is part of Springer Science+Business Media
(www.springer.com)

Foreword

This book edited by Rasit Topaloglu is a valuable addition to the library of the student and practicing engineer interested in learning about three emerging aspects of modern system design: three-dimensional integration, emerging memories and their architecture, and photonic communication.

These topics are assuming increased importance due to a confluence of multiple factors: the slowing down of conventional silicon scaling that enabled the enormously successful System on a Chip era (mainly for economic reasons), the need for ever-increasing memory at every level of the hierarchy, and finally the need to increase system bandwidth beyond what can be done by conventional interconnects. This book examines the design aspects of these three broad areas along with relevant physics and circuits background for these design constraints. The first two chapters deal with three-dimensional integration and the interconnect challenges of placing Through Silicon Vias (TSVs) in a design environment, especially the impact on area and power. It is important to note that while the TSVs may under certain circumstances increase chip power and area slightly (a few percent), the reduction in these parameters at the board or system level is enormous. For example, in the Hybrid Memory Cube embodiment,[1] main memory footprint may be reduced by 90 % while simultaneously reducing active power by about 30 % compared to DDR protocols. Even larger reductions are possible with further optimization.

While DRAM is the main memory workhorse of the memory subsystem, emerging memories such as magnetic memory are promising avenues to augment (*not* replace) DRAM. While the scaling potential of Magnetic Tunnel Junctions (MTJs) remains to be seen, they do offer the potential for persistence with very high endurance as magnetic memory does not suffer the fatigue experienced by phase change and charge trapping memories, radiation hardness, and perhaps offers some power reduction. This is the subject of three chapters in the memory section, while a fourth addresses improvements in architecture to reduce power and increase bandwidth in DRAM technology.

[1] http://www.hybridmemorycube.org/.

The last two chapters of this book address the application of photonics to improve communication at the chip and subsystem level. Li et al. address the important aspect of photonics for multiprocessor environments including the challenge of high-bandwidth communication with ever-increasing memory, while Condrat et al. address ideas for design automation—an important requirement if photonics is to be widely adopted.

We are entering an exciting era of "orthogonal scaling"[2] where classical scaling of silicon technology will continue to offer advantages albeit at lower levels of cost performance, and more advantage and cost reduction will occur due to smarter ways of system integration.[3] This book addresses some very important aspects of this transition.

Hopewell Junction, NY, USA
24 November 2014

Subramanian S. Iyer

[2] S.S. Iyer, "The evolution of embedded memory in high performance systems," *Proc. IEDM*, 33.1, 2012.

[3] S.S. Iyer, "Three-Dimensional Integration: An Industry Perspective," March 2015 issue of the MRS bulletin.

Preface

Traditional scaling in semiconductor integrated circuits where device sizes are made progressively smaller has significant limitations. More than Moore technologies offer many advantages by combining multiple types of devices on one chip. In this book, we focus on three More than Moore technologies in particular based on the impact they can bring and readiness of technology: 3D integration, novel memories, and nanophotonics.

Two of the big obstacles in improving performance of computing systems are (1) increased interconnection delays and (2) memory wall. The former indicates limitations on the speed of moving data from one portion of a chip to another, while the latter refers to limitations on the bandwidth of data transfer between memory and the processing units. In this book, we focus on More than Moore technologies that can alleviate these issues.

3D integration enables stacking multiple dies within a small footprint. Close proximity of devices improves performance by reducing interconnection delays. Additional input/output connections between processing units and memory can also help with the memory wall issue. The first two chapters cover design, modeling, architecture, and performance aspects of 3D integration.

Traditional memory scaling has shown to exacerbate the memory wall issue over time. Hence a drastic change is needed. Unless a different architecture is the solution, perhaps novel memory design may help with this issue. Chapters 3–5 target modeling, circuit design, and architecture aspects of spin-transfer torque memory. Chapter 6 discusses architecture implications of utilizing novel memories in modern computer architecture.

Although significant research was conducted and progress was made in utilizing optical electronics in computing systems, constraints such as area, manufacturing, and thermal issues have prevented practical implementations in computing. It is likely that optical interconnects will be introduced first to replace global interconnection. The last two chapters discuss network-on-chip architecture and design automation enablement of nanophotonic systems.

More than Moore technologies presented herein offer solutions to the interconnection and memory wall issues in modern computer design. Whether or when such solutions are accepted depends on keeping other design constraints such as cost, variability, and power within limits.

Hopewell Junction, NY, USA Rasit O. Topaloglu

Contents

Chapter 1
Impact of TSV and Device Scaling on the Quality of 3D ICs

Dae Hyun Kim and Sung Kyu Lim

Abstract TSVs have negative effects such as area, delay, and power overhead because of non-negligible TSV area and capacitance. Therefore, obtaining benefits such as wirelength reduction and performance improvement from 3D integration is highly dependent on the TSV size and capacitance. To reduce the negative effects, TSVs have been downscaled and sub-micron TSVs are expected to be commercially available in the near future. Meanwhile, devices have also been downscaled beyond 32 and 22 nm, so future 3D ICs will very likely be built with sub-micron TSVs and advanced device technologies. In this chapter, the impact of TSVs on the quality of today and future 3D ICs is investigated based on GDSII-level layouts.

1.1 Introduction

Three-dimensional integration provides many benefits such as higher bandwidth, smaller form factor, shorter wirelength, lower power, and better performance than traditional two-dimensional integrated circuits (2D ICs) [1–4]. These benefits are obtained by die stacking and use of through-silicon vias (TSVs) for inter-die connections. However, TSVs occupy silicon area and have non-negligible capacitance, which have negative effects on the quality of 3D ICs. The reason is as follows. If large TSVs are inserted into a 3D IC, the footprint area of the design becomes larger, so the amount of wirelength reduction decreases [5]. In addition, non-negligible TSV capacitance increases the total capacitance of 3D signal paths and degrades timing and dynamic power consumption of the 3D paths.

To reduce the negative effects, TSVs have been downscaled as devices have been downscaled [6–8]. However, since process technology is also advancing, future 3D ICs will very likely be fabricated with smaller TSVs and state-of-the-art process

D.H. Kim (✉)
Washington State University Pullman, WA, USA
e-mail: daehyun@eecs.wsu.edu

S.K. Lim
Georgia Institute of Technology Atlanta, GA, USA
e-mail: limsk@ece.gatech.edu

R.O. Topaloglu (ed.), *More than Moore Technologies for Next Generation Computer Design*, DOI 10.1007/978-1-4939-2163-8_1

technology. In this case, the negative effects of TSVs might remain the same or even increase depending on the relative size and electrical properties among transistors, local interconnects, and TSVs.

In this chapter, the impact of TSVs on the area, wirelength, critical path delay, and power of today and future 3D ICs is investigated based on GDSII-level layouts. To simulate future process technologies, a 22 and a 16 nm process and standard cell libraries are developed. Today and future 3D IC layouts are generated using these future process technologies as well as an existing 45 nm library and the impact of TSVs are studied thoroughly.

1.2 Preliminaries

1.2.1 Negative Effects of TSVs

Use of TSVs in 3D ICs causes area, delay, and power overhead because of non-negligible area TSVs occupy and non-negligible TSV capacitance as follows.

1.2.1.1 Area Overhead

TSVs are fabricated in the silicon bulk. If TSVs are inserted into existing whitespace, the die area does not increase and TSV insertion does not cause area overhead. However, TSVs have similar properties as standard cells,[1] so TSV insertion requires new whitespace to achieve fixed area utilization. Area overhead caused by inserting a TSV is represented by $A_{\mathrm{TSV}} + A_{\mathrm{KOZ}}$ where A_{TSV} is the area of a TSV and A_{KOZ} is the area of the whitespace called keep-out zone (KOZ) around the TSV. Assuming each TSV insertion needs new whitespace, the total area overhead caused by TSV insertion is computed by $N_{\mathrm{TSV}} \cdot (A_{\mathrm{TSV}} + A_{\mathrm{KOZ}})$ where N_{TSV} is the total number of TSVs. As the formula shows, the total area overhead is proportional to the total TSV count and the TSV size. Figure 1.1 shows the ratio between the total TSV area and the total silicon area. As the figure shows, when the area required for each TSV insertion is comparable to the average gate area ($A_{\mathrm{TSV+KOZ}} = A_{\mathrm{gate}}$), the area overhead caused by TSV insertion is less than 5 % until the TSV count reaches 5 % of the gate count. However, if the area required for each TSV insertion is greater than $1.5^{*}A_{\mathrm{gate}}$, the area overhead goes over 10 %. This area overhead increases as the TSV becomes larger, the keep-out zone size becomes larger, more TSVs are inserted, the gate count decreases, or the average gate area goes down (smaller gates are used or newer technology is used).

[1] Both standard cells and TSVs occupy silicon area and need routing from/to them.

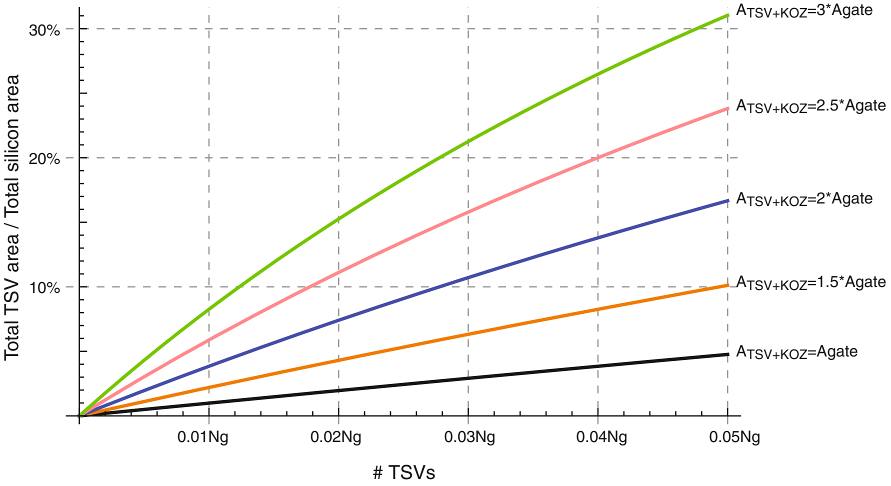


Fig. 1.1 Ratio between the total TSV area and the total silicon area. N_g is the total gate count, $A_{\mathrm{TSV+KOZ}}$ is the sum of the TSV and KOZ area, and A_{gate} is the average gate area

1.2.1.2 Delay Overhead

The total wirelength of 3D designs is expected to be shorter than their 2D counterparts by 20–35 % although TSV insertion causes area overhead [9]. However, delay of a 3D net, which connects gates placed in different dies, is dependent on both the length of the net and the capacitance of the TSVs inserted into the net, which means that shorter wirelength might not be converted into lower delay if TSV capacitance is not sufficiently small or many TSVs exist on the net. For instance, assume that the length of a net is $X(\mu\mathrm{m})$, the unit resistance and capacitance of the net are $r(\Omega/\mu\mathrm{m})$ and $c(\mathrm{fF}/\mu\mathrm{m})$, respectively, the load capacitance is C_{L}, the output resistance of the driver of the net is R_{o}, the capacitance of a TSV is C_{TSV}, and the number of TSVs in the net is N_{TSV}. Assuming the TSVs are evenly distributed along the net, the Elmore delay of the net is represented as follows:

$$\begin{aligned} &R_{\mathrm{o}} \cdot (cX + C_{\mathrm{L}}) + 0.5rcX^2 + rX \cdot (0.5c + C_{\mathrm{L}}) + R_{\mathrm{o}} \cdot N_{\mathrm{TSV}} \cdot C_{\mathrm{TSV}} \\ &\quad + \cdot 0.5rX \cdot N_{\mathrm{TSV}} \cdot C_{\mathrm{TSV}} \end{aligned} \tag{1.1}$$

where the first term is the delay by the driver output resistance driving the net and the load capacitance, the second and the third terms are wire delays, the fourth term is the delay by the driver output resistance driving the TSVs, and the fifth term is the delay by the wire resistance driving the TSVs. In the above equation, TSV resistance is not considered because in general TSV resistance is very small (less than 1 Ω) compared to the wire resistance (a few Ω/μm), so the impact of TSV resistance on the net delay is negligible. Figure 1.2 shows the Elmore delay of 2D

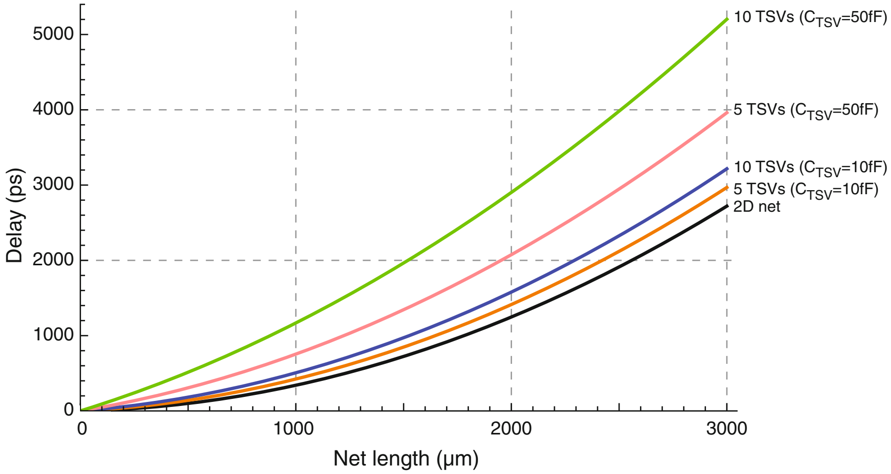


Fig. 1.2 Net delay. $R_{\mathrm{o}} = 305\,\Omega$. $r = 3.31\,\Omega/\mu\mathrm{m}$. $c = 0.171$ fF/μm. $C_{\mathrm{L}} = 1.55$ fF. It is assumed that TSVs are evenly distributed along the net

and 3D nets for 45 nm technology parameters. As the figure shows, inserting five and ten TSVs having 10 fF capacitance into a 3,000 μm net causes 9 and 18 % delay overhead, respectively, compared to the delay of the 2D net of the same length. If TSV capacitance increases to 50 fF, inserting five and ten TSVs into the same net causes 45 and 91 % delay overhead, respectively.

1.2.1.3 Power Overhead

Non-negligible TSV capacitance causes dynamic power consumption overhead. Using the well-known formula for dynamic power consumption, $P = \alpha f C V^2$, power overhead by TSV insertion is represented by $P = \alpha f C_{\mathrm{TSV}} V^2$ where C_{TSV} is the TSV capacitance. Since 3D ICs have in general shorter wirelength than 2D ICs, however, total dynamic power consumption could still be lower than 2D ICs depending on the amount of wirelength reduction and the power overhead by TSVs. For instance, assume that the switching factor is 0.2, operation frequency is 1 GHz, supply voltage is 1.0 V, and unit wire capacitance is 0.171 fF/μm. If the total wirelength is 100 m, its dynamic power consumption is 3.42 W. When the 2D design is converted into 3D and the total wirelength becomes 20 % shorter than that of the 2D design, its dynamic power consumption becomes the sum of 2.74 W and P_{TSV}, which is the power consumed by TSVs. If C_{TSV} is 10 fF and 10,000 TSVs are inserted, P_{TSV} becomes 0.02 W and the total power is 2.76 W, which is still almost 20 % lower than the power of the 2D design. However, if a million TSVs are inserted, P_{TSV} becomes 2 W and the total power becomes 4.74 W, which is 39 % higher than that of the 2D design.

1.2.2 Motivation

Process scaling reached 22 nm node [10,11] in 2012, and 16 and 11 nm technologies are currently under development as of 2013. As devices are downscaled, TSVs are also downscaled as TSV manufacturing technology advances. Recently, it was demonstrated that 0.7 µm-diameter TSVs could also be fabricated reliably [8]. In addition, according to the ITRS prediction, TSV diameter will continue to decrease while TSV aspect ratio will increase. Therefore, sub-micron TSVs are expected to be developed and ready for use within the next few years. Thus, it is worthwhile studying the impact of both super- and sub-micron TSVs on the quality of 3D IC designs. In addition, the impact of TSVs on the quality of 3D ICs varies depending on which process technology is used to build the 3D ICs. For instance, inserting 5 µm-diameter TSVs into a one-million-gate 3D IC built by 32 nm technology will have greater impact on the area overhead than inserting the same-size TSVs into the same 3D IC built by 90 nm technology because the average gate size of the former is smaller than that of the latter. Therefore, the goal in this chapter is to investigate the impact of super- and sub-micron TSVs on the area, wirelength, critical path delay, and power of 3D ICs built by different process technology.

1.3 Library Development

Creating 3D IC layouts requires process and standard cell libraries. Therefore, this section briefly describes the library development flow used to create a 22 and a 16 nm process and standard cell libraries.

1.3.1 Library Development Flow

Creation of process and standard cell libraries requires the followings:

- **Transistor models**: Transistor models are used to create timing and power libraries.
- **Interconnect layer definitions**: Interconnect layer definitions include the minimum width, thickness, minimum spacing, and routing direction of each metal layer, the width and the spacing of each via layer, and permittivity and height of each inter-layer dielectric. Although the libraries used for fabrication require much more elaborate and complex design rules, the above definitions are sufficient for the simulation in this chapter.
- **Standard cell layouts**: Standard cell libraries should have various simple and complex cells.

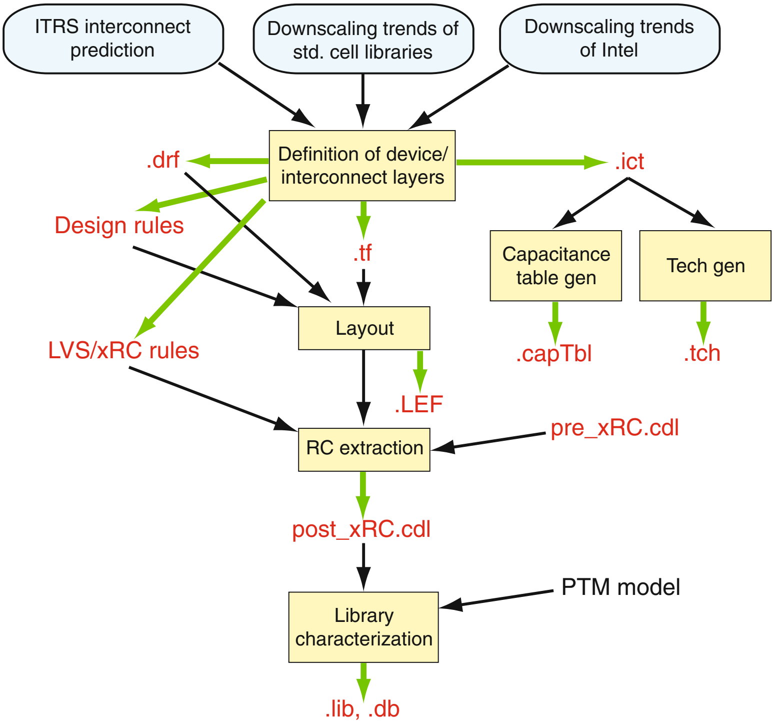


Fig. 1.3 Process and standard cell library development flow

For 22 and 16 nm transistor models, the predictive technology model (22 and 16 nm PTM HP model V2.1) is used [12]. To develop 22 and 16 nm process and standard cell libraries, a typical library development flow illustrated in Fig. 1.3 is used. First, device and interconnect layers are defined and then tech files (.tf), display resource files (.drf), interconnect technology files (.ict), design rule files, layout-versus-schematic (LVS) rule files, and RC parasitic extraction rule files are created from the device and interconnect layer definitions. With the tech and display resource files, DRC-clean standard cell layouts are created. After layout generation, abstraction is performed to create library exchange format files (.LEF). SPICE netlists (post_xRC.cdl) containing parasitic resistance and capacitance are also created after RC extraction. With these SPICE netlists and the PTM transistor models, library characterization is performed to create timing and power libraries (.lib and .db). Various technology-dependent files such as capacitance tables are also generated for accurate RC extraction and timing analysis in digital IC design tools.

1.3.2 Interconnect Layers and Standard Cell Libraries

Interconnect layers for a specific technology can be defined by extrapolation or prediction. In this chapter, interconnect layers for 22 and 16 nm technologies are defined by prediction based on ITRS interconnect prediction [16], downscaling trends of other standard cell libraries, and the downscaling trends of Intel process technology [13–15]. According to ITRS prediction on interconnect layers, for example, the pitch of the metal 1 wire at 22 nm is about 72 nm and that at 16 nm is about 48 nm, and the pitch of a semi-global wire at 22 nm is about 160 nm and that at 16 nm is about 130 nm. From these values as well as extrapolation of interconnect layers of Intel process technology and other standard cell libraries, interconnect layers at 22 and 16 nm are predicted as shown in Table 1.1. Table 1.2 shows the width and thickness of each metal layer and the thickness of each barrier and inter-layer dielectric (ILD) layer in the 22 and 16 nm process libraries predicted in this chapter. The aspect ratio of the 22 nm library is set to 1.8 and that of the 16 nm library is set to 1.9. Since it is assumed that low-k inter-layer insulator material is used, 1.9 is used for the dielectric constant of the inter-layer dielectric material and 3.8 for the dielectric constant of the barrier material for both the 22 nm and the 16 nm libraries.

With the interconnect layers defined above and a minimum set of design rules such as the minimum poly-to-contact spacing and the minimum metal-to-metal spacing, standard cell layouts are created. About 90 standard cells are created as shown in Table 1.3. Figure 1.4 shows the smallest (1×) two-input NAND gates of the 45, 22, and 16 nm standard cell libraries. After creating the standard cell layouts, DRC, LVS, and RC extraction are executed for each layout. With the netlists having parasitic RC and the transistor SPICE models, all the standard cells are characterized and timing and power libraries are created.

Table 1.1 Interconnect layers of 65 nm [13], 45 nm [14], 32 nm [15], 22 nm, and 16 nm process technology

	Pitch (nm)				
Layer	65 nm	45 nm	32 nm	22 nm	16 nm
Contacted gate	220	160	112.5	86	62
Metal 1	210	160	112.5	76	46
Metal 2	210	160	112.5	76	46
Metal 3	220	160	112.5	76	46
Metal 4	280	240	168.8	130	72
Metal 5	330	280	225.0	206	98
Metal 6	480	360	337.6	206	146
Metal 7	720	560	450.1	390	240
Metal 8	1080	810	566.5	390	240

The 22 and the 16 nm layers are based on prediction and extrapolation

Table 1.2 Width (w) and thickness (t) of the metal, barrier, and inter-layer dielectric (ILD) layers used in the 22 and 16 nm process libraries

	22 nm		16 nm	
Layer	w (nm)	t (nm)	w (nm)	t (nm)
Metal 1, 2, 3	36	64.8	22	41.8
Barrier 1, 2, 3		6		4
ILD 1–2, 2–3, 3–4		53		34
Metal 4	60	108	32	60.8
Barrier 4		12		8
ILD 4–5		84		45
Metal 5	96	172.8	44	83.6
Barrier 5		18		8
ILD 5–6		137		68
Metal 6	96	172.8	66	125.4
Barrier 6		18		8
ILD 6–7		137		109
Metal 7, 8	180	324	110	209
Barrier 7, 8		24		14
ILD 7–8, 8–9		276		181

The aspect ratio of the metal layers for the 22 nm library is 1.8 and that for the 16 nm library is 1.9

Table 1.3 Standard cells in the 22 and 16 nm standard cell libraries developed in this chapter

Type	Available sizes
AND2/3/4, AOI21/211/221	1×, 2×, 4×
BUF, INV	1×, 2×, 4×, 8×, 16×, 32×
LOGIC 0, LOGIC 1	1×
MUX2	1×, 2×
NAND2/3/4/, NOR2/3/4	1×, 2×, 4×
OAI21/22/211/221/222	1×, 2×, 4×
OAI33	1×
OR2/3/4	1×, 2×, 4×
XNOR2, XOR2	1×, 2×
DFF	1×, 2×
FA, HA	1×

1.4 Comparison of Process and Standard Cell Libraries

This section validates the 22 and 16 nm libraries created by the library development flow presented in the previous section. Since the libraries are not based on fabrication and measurement data, the validation relies on the trend of the process scaling and the libraries are accepted unless their characteristics do not deviate too much from the expectation.

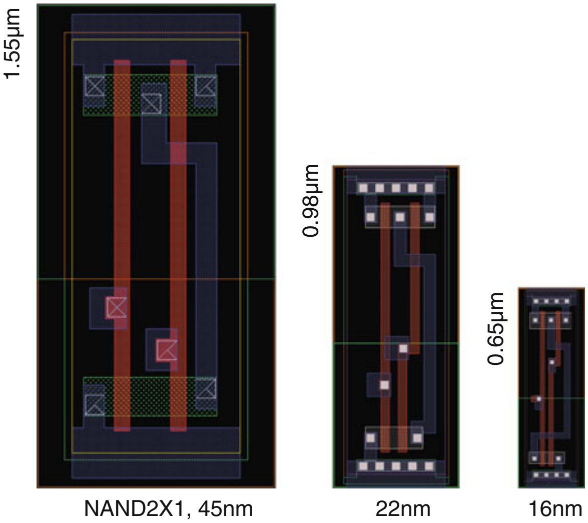


Fig. 1.4 The smallest (1×) two-input NAND gates of the 45 nm [17], and the 22 and 16 nm libraries (drawn to scale) created in this chapter

1.4.1 Gate Delay and Input Capacitance

The first simulation for the validation of the 22 and 16 nm libraries is to compare transistor characteristics. The simulation setting is as follows. The minimum-size inverter in each process library drives another minimum-size inverter, which drives an N× inverter of the same library. The delay of the second minimum-size inverter (driving the N× inverter) is obtained by SPICE simulation. Figure 1.5 shows the delay. As shown in the figure, the 16 nm inverter has the shortest delay and the 45 nm inverter has the longest delay. Quantitatively, when the process moves from 45 to 22 nm and from 22 to 16 nm, approximately 30 and 20 % delay improvement is obtained, respectively. Notice that two-generation gap exists between 45 and 22 nm while only one-generation gap exists between 22 and 16 nm, so moving from 45 to 22 nm has larger improvement than moving from 22 to 16 nm. Table 1.4 also shows the FO4 delay at each process technology.

Gate input capacitance, which is also an important factor determining delay and power, of some of the 45, 22, and 16 nm standard cells are listed in Table 1.5. As shown in the table, the average input capacitance of the 22 nm standard cells is approximately 48 % of that of the 45 nm standard cells. On the other hand, the average input capacitance of the 16 nm standard cells is approximately 83 % of that of the 22 nm standard cells. Since two-generation gap exists between 45 and 22 nm, the input capacitance difference between 45 and 22 nm is greater than that between 22 and 16 nm.

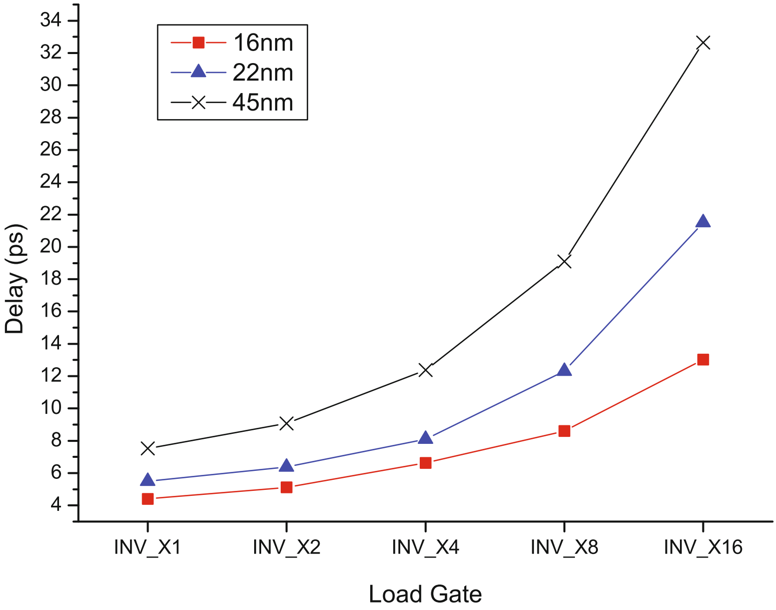


Fig. 1.5 Delay of a minimum-size inverter driving an N× inverter (N = 1, 2, 4, 8, 16), where both inverters are in the same process. RC parasitics are included

Table 1.4 FO4 delay, standard cell heights, wire sheet resistance, and unit wire capacitance (fF/μm)

	45 nm	22 nm	16 nm
FO4 delay (ps)	15.15	13.63	12.28ps
Std. cell. height (μm)	1.4	0.9	0.6
Wire sheet resistance (Metal 1)	0.38	0.26	0.40
(Metal 4)	0.21	0.16	0.28
(Metal 7)	0.08	0.05	0.08
Unit wire capacitance (Metal 1)	0.20	0.15	0.16
(Metal 4)	0.20	0.15	0.13
(Metal 7)	0.20	0.14	0.14

1.4.2 Interconnect Layers

Characteristics of interconnect layers also have a significant effect on the performance. Table 1.4 shows wire sheet resistance and unit wire capacitance of short, semi-global, and global metal layers in the 45, 22, and 16 nm libraries. The resistivity of the 45 nm technology is about 5.0×10^{-8}, so the sheet resistance of the 45 nm library is relatively high compared to the 22 nm library. On the other hand, the resistivity of the 22 and 16 nm technology is 1.7×10^{-8}, which is the resistivity of copper. This is why the sheet resistances of the 22 nm metal layers are lower than those of the 45 nm metal layers although the thickness of the 45 nm metal layers is larger than that of the 22 nm metal layers. As the technology moves from 22 to

Table 1.5 Input capacitance of selected standard cells in the 45, the 22, and the 16 nm libraries

	Cap (fF)		
Cell	45 nm	22 nm	16 nm
AND2 1×	0.54 (1.00)	0.25 (0.46)	0.22 (0.41)
AOI211 1×	0.64 (1.00)	0.30 (0.47)	0.25 (0.39)
AOI21 1×	0.55 (1.00)	0.23 (0.42)	0.20 (0.36)
BUF 4×	0.47 (1.00)	0.28 (0.60)	0.29 (0.62)
DFF 1×	0.90 (1.00)	0.41 (0.46)	0.26 (0.29)
FA 1×	2.46 (1.00)	1.31 (0.53)	1.36 (0.55)
INV 4×	1.45 (1.00)	0.69 (0.48)	0.56 (0.39)
MUX2 1×	0.95 (1.00)	0.42 (0.44)	0.34 (0.36)
NAND2 1×	0.50 (1.00)	0.24 (0.48)	0.22 (0.44)
OAI21 1×	0.53 (1.00)	0.25 (0.47)	0.20 (0.38)
OR2 1×	0.60 (1.00)	0.26 (0.43)	0.20 (0.33)
XOR2 1×	1.08 (1.00)	0.55 (0.51)	0.45 (0.42)
Average	(1.00)	(0.48)	(0.40)

Table 1.6 Benchmark circuits

			Total cell area		
Circuit	# Gates (K)	# Nets (K)	45 nm	22 nm	16 nm
BM1	352	372	0.632	0.218	0.098
BM2	518	680	1.288	0.437	0.198

16 nm, the sheet resistance goes up because both of them use the same resistivity, but the metal layer thickness of the 16 nm library is smaller than that of the 22 nm library.

The unit wire capacitance of the 45 nm library is also slightly higher than that of the 22 nm library. This is because the dielectric constant used for the 45 nm library is 2.5 while the 22 nm library uses 1.9 for its dielectric constant. If the same dielectric material ($\epsilon_r = 1.9$) is used for the 45 nm library, the unit wire capacitance becomes 0.15, which is close to the unit wire capacitance of the 22 nm library.

1.4.3 Full-Chip 2D Design

In this simulation, 2D circuit layouts are designed using the three standard cell libraries and the area, wirelength, critical path delay, and power of the designs are compared. The simulation flow is as follows. Two benchmark circuits shown in Table 1.6 are synthesized, designed, and optimized using each standard cell library and commercial tools. The same area utilization (60 %) is used for all designs for fair comparison and the fastest operation frequency is found for each library.

Table 1.7 shows the comparison results for the 2D designs. The chip area of the 45 nm designs is about three times larger than that of the 22 nm designs on average, and the chip area of the 22 nm designs is approximately two times larger

Table 1.7 Comparison of 2D layouts

	BM1			BM2		
	45 nm	22 nm	16 nm	45 nm	22 nm	16 nm
Area (mm^2)	1.00	0.36	0.17	2.56	0.81	0.42
Wirelength (m)	10.65	4.22	2.75	15.17	8.90	6.19
Delay (ns)	3.19	2.61	2.38	6.51	4.10	3.93
Power (W)	0.352	0.0684	0.068	0.521	0.154	0.133

than that of the 16 nm designs on average. In addition, the total wirelength of the 16 nm designs is approximately 1.48× shorter than that of the 22 nm designs, and 3.08× shorter than that of the 45 nm designs. Regarding the critical path delay, the 16 nm designs are 1.49× faster than the 45 nm designs on average and 1.07× faster than the 22 nm designs on average. Power consumption of the 16 nm designs is approximately 4.5× smaller than that of the 45 nm designs and 1.1× smaller than that of the 22 nm designs. Overall, the delay and power enhancement coming from 22 to16 nm transition is not as significant as the enhancement coming from 45 to 22 nm transition because 45 and 22 nm technologies are two generations apart while 22 and 16 nm technologies are only one generation apart, and the quality (sheet resistance and unit wire capacitance) of the interconnect layers of the 45 nm library is worse than that of the 22 nm library.

1.5 3D IC Design and Analysis Methodology

To generate 3D IC layouts, the 3D RTL-to-GDSII tool obtained from [18] are used. This tool works as follows: For a given 2D gate-level (flattened) netlist, this tool partitions gates in the x-, y-, and z- directions iteratively to globally place gates in grids in 3D. After global placement, it constructs a 3D Steiner tree for each 3D net and inserts TSVs into each placement grid based on the locations of vertical edges of the 3D Steiner tree. Then, it runs detailed placement in each placement grid using Cadence Encounter [19]. Routing for each die is also performed by Encounter. The output of the tool consists of a Verilog netlist, a design exchange format (DEF) file containing TSV locations, and a standard parasitic exchange format (SPEF) file for each die, and a top-level Verilog netlist containing die-to-die connections and a top-level SPEF file. One thing to notice is that the minimum number of TSVs to be inserted in the 3D design is dependent on the cut sequence, which is the order of the x-, y-, and z- direction partitioning applied to global placement. For example, if the z-direction partitioning is applied early, fewer inter-die connections will likely be obtained. On the other hand, if the z-direction partitioning is applied later, more inter-die connections will likely be obtained [18]. This variation of the number of TSVs enables producing different global placement solutions with different TSV counts.

After 3D IC layouts are generated, 3D timing optimization is performed as follows. First, an initial timing optimization is performed in each die. Then, all the layouts, timing analysis results, and the target clock frequency are fed into the 3D timing optimization tool obtained from [20]. This 3D timing optimization tool iterates the following steps: (a) it performs RC extraction and obtains an SPEF file for each die; (b) it performs 3D timing analysis using the SPEF files and the top-level SPEF file using Synopsys PrimeTime [21]; (c) based on the timing analysis result and the target clock frequency, the tool determines the target delay of each 3D path and creates a timing constraint file for each die; (d) since each die has its own netlist and timing constraint file, timing optimization is performed for each die separately using Encounter. This timing optimization process is repeated several times until the overall timing improvement saturates.

3D power analysis needs (a) a netlist for each die and a top-level netlist, (b) an SPEF file for each die and a top-level SPEF file, and (c) switching activities of cells and nets. To obtain switching activities of cells and nets, Verilog netlists generated by the 3D RTL-to-GDSII tool obtained from [18] are loaded into Encounter and power analysis is performed. The power analysis internally generates and stores switching activities of the cells and nets, so this information is dumped into an output file after the power analysis. Then, all the netlists, SPEF files, and the switching activity files are loaded into PrimeTime and power analysis is performed. This power analysis method produces true full-chip 3D power analysis results.

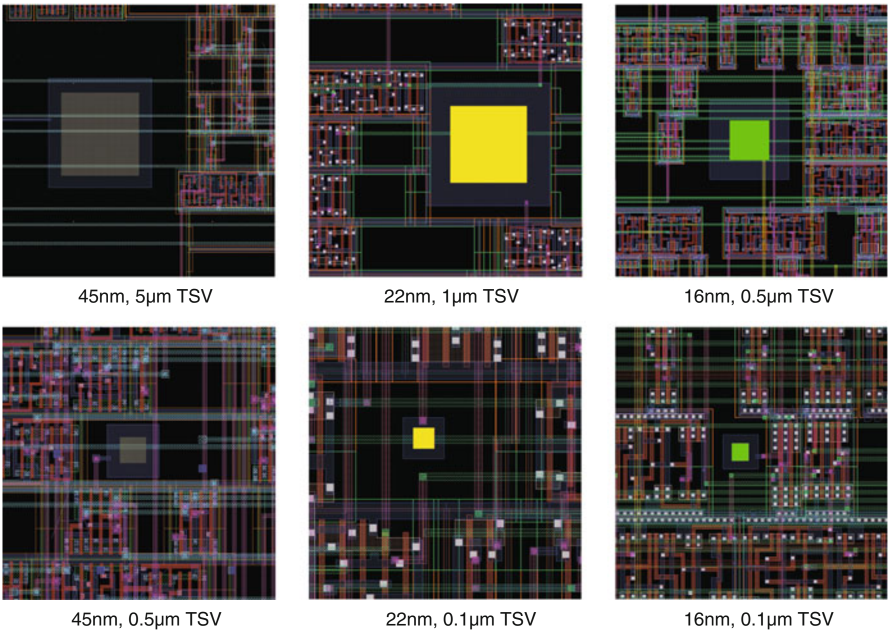
1.6 Simulation Results

1.6.1 Simulation Settings

Two benchmark circuits, BM1 and BM2 shown in Table 1.6, are used to evaluate the quality of 3D IC layouts. For the 45 nm process node, the Nangate 45 nm standard cell library is used [17]. The four sets of TSV-related dimensions listed in Table 1.8 are used to simulate today and future TSVs. Especially, the 5 and 0.5 µm TSVs are used with the 45 nm technology, the 1 and 0.1 µm TSVs are used with the 22 nm technology, and the 0.5 and 0.1 µm TSVs are used with the 16 nm technology. Since the standard cell height of the 45 nm library is 1.4 µm, a 5 µm TSV including keep-out zone occupies five standard cell rows while a 0.5 µm TSV including keep-out zone occupies one standard cell row. Similarly, a 1 µm TSV and a 0.1 µm TSV occupy three standard cell rows and 0.26 standard row, respectively, when they are used with the 22 nm standard cell library. If 0.5 and 0.1 µm TSVs are used with the 16 nm standard cell library, a 0.5 µm TSV occupies 1.33 standard cell rows and a 0.1 µm TSV occupies 0.5 standard cell row. Figure 1.6 shows GDSII images of TSVs and standard cells at 45, 22, and 16 nm technology. Die thickness for each TSV dimension set is the same as the TSV height, which ranges from 5 to 25 µm. Although 5 µm thickness is extremely thin, it is practical [1, 22].

Table 1.8 TSV-related dimensions, design rules, and TSV capacitance

Dimensions	TSV-5	TSV-1	TSV-0.5	TSV-0.1
Width (μm)	5	1	0.5	0.1
Height (μm)	25	5	8	5
Aspect ratio	5	5	16	50
Liner thickness (nm)	100	30	20	10
Barrier thickness (nm)	50	15	10	5
Landing pad width (μm)	6	1.6	1	0.18
TSV-to-TSV spacing (μm)	2	0.8	0.6	0.1
TSV-to-device spacing (μm)	1	0.4	0.3	0.1
TSV capacitance (fF)	20	2.67	3.2	0.8
Used with	45 nm		45 nm	
		22 nm		22 nm
			16 nm	16 nm


45nm, 5μm TSV

22nm, 1μm TSV

16nm, 0.5μm TSV

45nm, 0.5μm TSV

22nm, 0.1μm TSV

16nm, 0.1μm TSV

Fig. 1.6 GDSII images (zoom-in shots) of the six types of designs studied in this chapter. Each TSV has keep-out zone around it

1.6.2 *Impact on Silicon Area*

Figure 1.7 shows the footprint area of 2D and two-die 3D BM1 and BM2 designs at each technology node. If the TSV size is zero, the footprint area of a two-die 3D design is approximately half of its 2D counterpart. Since the TSV size is not

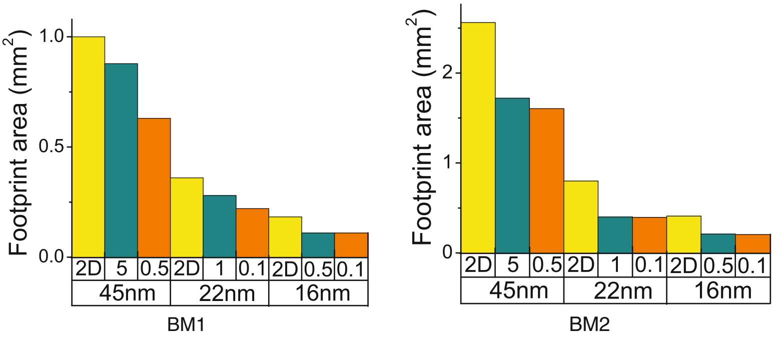


Fig. 1.7 Footprint area of the optimized 2D and two-die 3D designs (*left*: BM1, *right*: BM2). The x-axis shows the technology combination (the *first row* shows TSV diameter in μm)

zero in reality, however, the footprint area of a two-die 3D design is usually greater than half of its 2D counterpart if the same utilization is applied to both 2D and 3D designs. For example, the area of the 45 nm 2D design is 1.0 mm^2, but the area of the 45 nm 3D design using 5 μm TSVs is about 0.85 mm^2, which is 85 % of the 2D design. Similarly, the area of the 45 nm 3D design using 0.5 μm TSVs is about 0.63 mm^2, which is 63 % of the 2D design. The same trend is found in the 22 and 16 nm designs. However, if the TSV size is 0.1 μm, the footprint area of two-die 3D designs becomes almost half of the area of their 2D counterparts. Similar trends are found in the BM2 designs.

All these trends depend on the TSV size and the number of TSVs used in the designs. Using smaller TSVs helps achieve smaller footprint area, which can reduce the chip cost. However, smaller TSVs could be more expensive due to manufacturing difficulties, so the use of smaller TSVs might not necessarily lead to lower chip cost. Using fewer TSVs also helps achieve smaller footprint area. However, several studies show that using more TSVs than the minimum number of TSVs helps reduce wirelength and improve performance [18, 23, 24]. Thus, tradeoffs exist among the TSV size, the number of TSVs used in the design, chip cost, the footprint area, and the chip performance.

1.6.3 Impact on Wirelength

The left figure in Fig. 1.8 shows the wirelength of the BM1 benchmark circuit. When 5 μm TSVs are used with the 45 nm technology, the 3D design has longer wirelength than the 2D design. However, when 0.5 μm TSVs are used with the 45 nm technology, the wirelength of the 3D design is about 10 % shorter than that of the 2D design. When 1 and 0.1 μm TSVs are used with the 22 nm technology, however, the amount of wirelength reduction is less than 4 %. On the other hand, when 0.5 and 0.1 μm TSVs are used with the 16 nm technology, 15 % wirelength reduction is achieved.

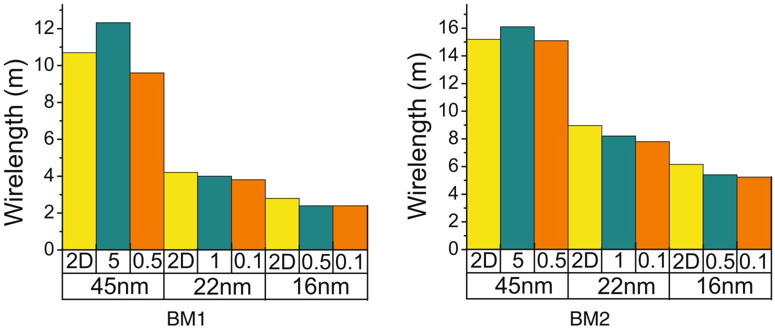


Fig. 1.8 Wirelength of the optimized 2D and two-die 3D designs (*left*: BM1, *right*: BM2). The x-axis shows the technology combination (the *first row* shows TSV diameter in μm)

Similar trends are found in the BM2 designs as shown in the right figure in Fig. 1.8. The 45 nm 3D design has longer wirelength than the 2D design. However, when 1 and 0.1 μm TSVs are used with the 22 nm technology, 9 and 13 % wirelength reduction is achieved, respectively. Similarly, 12 and 15 % wirelength reduction is achieved when 0.5 and 0.1 μm TSVs are used with the 16 nm technology, respectively.

One thing to note is that 3D designs at n-th generation process node could have longer wirelength than 2D designs at $(n + 1)$-th generation process node. For instance, the 22 nm 3D layouts designed with 0.1 μm TSVs have longer wirelength than the 16 nm 2D layouts in Fig. 1.8. Therefore, shrinking the TSV size is important to reduce the wirelength, but switching to advanced process nodes is also important for wirelength reduction. This observation also coincides with the prediction result presented in [25].

1.6.4 Impact on Performance

Figure 1.10 shows the critical path delay for the BM1 and BM2 benchmark circuits. In general, using smaller TSVs leads to shorter critical path delay because of smaller TSV capacitance and area overhead. In addition, as seen in Figs. 1.8 and 1.10, the critical path delay of a 3D design having longer wirelength than (or similar wirelength to) its 2D counterpart can be smaller than that of the 2D design. For example, the wirelength of the 3D design built with 5 μm TSVs and the 45 nm technology is 15 % longer than that of the 2D design, but the critical path delay of the 3D design is 12 % smaller than that of the 2D design. Similar trends are also found in the BM2 benchmark circuit (Fig. 1.9).

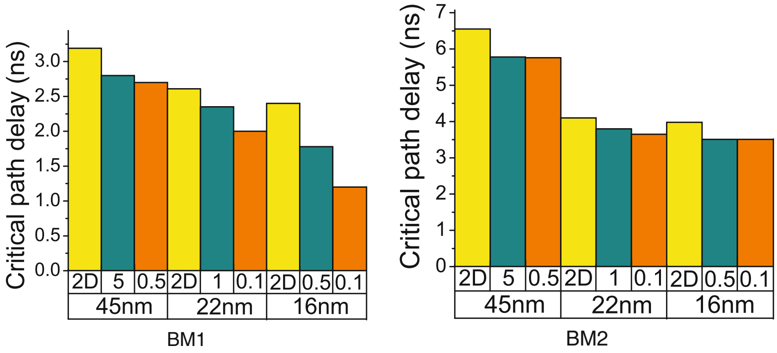


Fig. 1.9 Critical path delay of the optimized 2D and two-die 3D designs (*left*: BM1, *right*: BM2). The x-axis shows the technology combination (the *first row* shows TSV diameter in μm)

An important observation is that the critical path delay of 3D designs built with n-th generation process node could be smaller than that of 2D designs built with $(n + 1)$-th generation process node. For example, the BM1 3D design built with 0.1 μm TSVs and the 22 nm technology has approximately 20 % smaller delay than the 2D design built with the 16 nm technology. Similarly, the BM2 3D design built with 0.1 μm TSVs and the 22 nm technology has about 9 % smaller delay than the 2D design built with the 16 nm technology.

1.6.5 Impact on Power

Figure 1.10 shows power consumption for the BM1 and BM2 benchmark circuits. As seen in the figures, moving from 2D ICs to 3D ICs does not necessarily lead to power reduction even if 3D designs have shorter wirelength than 2D designs. The reason is as follows. Reduction in power consumption by building 3D ICs comes mainly from smaller dynamic power consumption due to shorter wirelength.[2] However, TSV capacitance can essentially be thought of as wire capacitance. Therefore, the total capacitance is the sum of the total TSV capacitance and the total wire capacitance. This means that the total TSV capacitance should be less than the reduced wire capacitance to achieve power reduction.[3] In other words, achievement of power reduction needs smaller TSV capacitance, use of fewer TSVs, and wirelength reduction. However, there again exists trade-offs among the number of TSVs, the amount of wirelength reduction, and power consumption.

[2] There exist many kinds of 3D integration and some of them (e.g., core-DRAM stacking) provide a huge amount of power saving by removing long chip-to-chip connections.

[3] Note that this is a simplified analysis. In reality, the total power should be computed in a more sophisticated fashion taking switching activities of nets and gates into account.

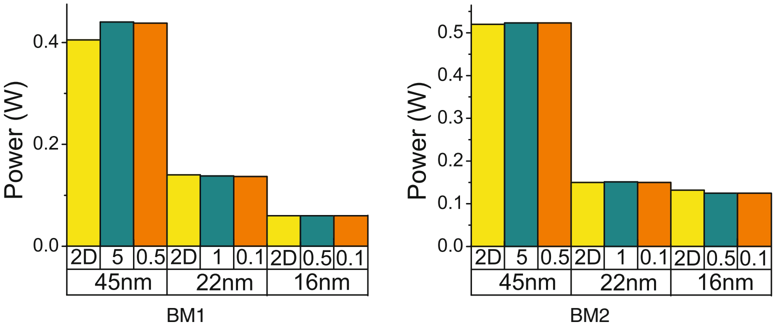


Fig. 1.10 Power consumption of the optimized 2D and two-die 3D designs (*left*: BM1, *right*: BM2). The x-axis shows the technology combination (the *first row* shows TSV diameter in μm)

Inserting fewer TSVs may not reduce the total wirelength as much as expected. Similarly, inserting fewer TSVs may not reduce the dynamic power consumption. Inserting more TSVs, however, may reduce the total wirelength more than 10–20 % [23], but then the total TSV capacitance also increases, so the total capacitance could be larger than the total capacitance of 2D designs.

Another reason that the total power does not decrease in 3D designs is related to the wirelength distribution. If wirelength reduction is achieved by shortening short wires in a net, the capacitance of the input pins connected to the net dominates the capacitance of the net, so the power consumption does not decrease. However, if long wires are shortened, the wire capacitance dominates the capacitance of the net, so the power will reduce. In the simulation, however, the wirelength reduction comes primarily from shortening short wires.

1.6.6 Area, Wirelength, Performance, and Power vs. # Dies

The number of dies stacked in a 3D IC also has a non-negligible impact on the area, wirelength, critical path delay, and power [23]. In this section, therefore, the impact of TSVs on the four quality metrics are studied when the die count varies. Figures 1.11 and 1.12 show the footprint area, wirelength, critical path delay, and power of the BM1 and BM2 benchmarks when the number of dies varies from two to five. To limit the simulation space size, 0.5, 1, and 0.5 μm TSVs are used for the 45, 22, and 16 nm technologies, respectively.

As the number of dies increases, the footprint area decreases as expected. Assuming that the TSV size is zero and the same utilization is used for all layouts, the footprint area of an n-die design of a circuit is approximately A_{2D}/n where A_{2D} is the area of the 2D design of the circuit. However, the TSV size is not zero and stacking more dies usually increases the number of TSVs inserted, so the footprint area of the circuit designed in n dies is larger than A_{2D}/n. A noticeable result found

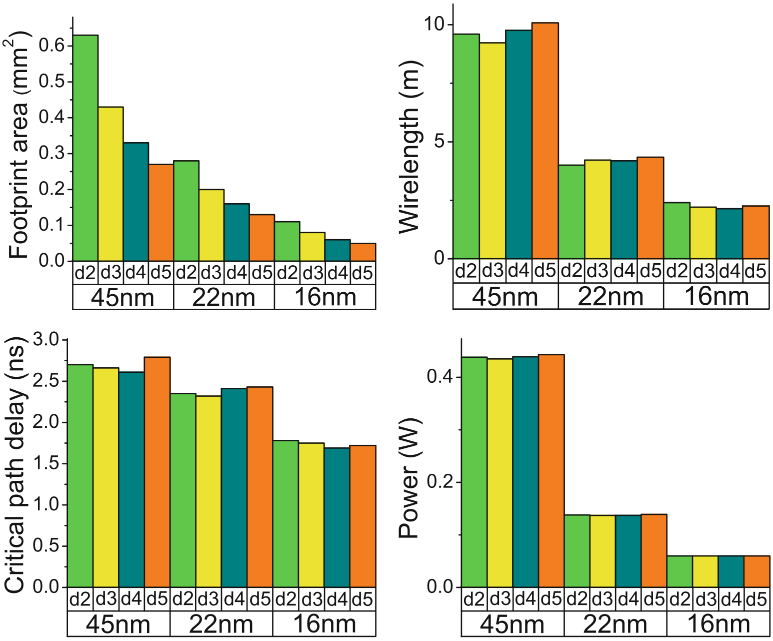


Fig. 1.11 Comparison of optimized 3D designs (BM1) implemented in multiple dies. "d*n*" denotes n-die implementation. 0.5, 1, and 0.5 μm TSVs are used for the 45, 22, and 16 nm technologies, respectively

in the figures is that 3D ICs stacked in more than four or five dies can have smaller footprint area than their 2D counterparts built by more advanced technology. For example, the five-die 3D IC built with 45 nm technology in Fig. 1.11 overcomes the two-generation gap and has smaller footprint area than the 2D IC built with 22 nm technology. Similarly, the five-die 3D IC built with 22 nm technology has smaller footprint area than the 2D IC built with 16 nm technology in Fig. 1.12.

On the other hand, stacking more dies does not necessarily result in shorter wirelength although stacking more than two dies helps reduce the wirelength. The largest wirelength reduction ratio between more-than-two-die designs and two-die designs is about 11 % in the simulation (the 16 nm two-die implementation vs. the 16 nm four-die implementation of BM1). In addition, stacking five dies does not produce shorter wirelength than stacking two to four dies. The main reason is because stacking more dies generally needs more TSVs, which causes wirelength overhead because of area overhead.

Regarding the critical path delay, stacking three or four dies reduces the critical path delay more effectively than stacking two dies. The largest critical path delay ratio between more-than-two-die designs and two-die designs is about 5 % in the figure (the 16 nm four-die implementation of BM1). However, stacking more than four dies does not reduce the critical path delay effectively. On the other hand, power

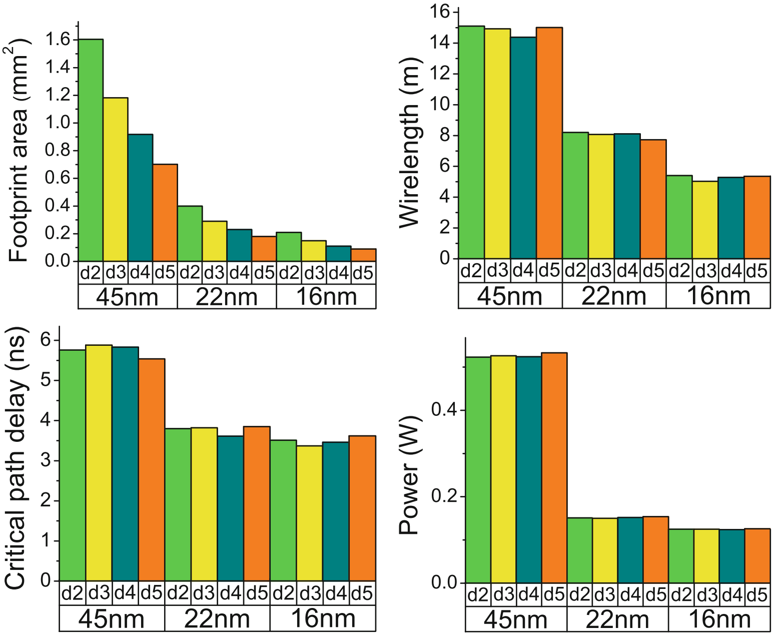


Fig. 1.12 Comparison of optimized 3D designs (BM2) implemented in multiple dies. "d*n*" denotes n-die implementation. 0.5, 1, and 0.5 μm TSVs are used for the 45, 22, and 16 nm technologies, respectively

consumption varies in a very small range. The reason is because gate switching power is much more dominant, so the combination of reducing wirelength (a positive effect) and inserting more TSVs (a negative effect due to TSV capacitance) leads to the very small change in total power consumption.

1.7 Conclusion

In this chapter, the impact of TSVs on the quality of today and future 3D ICs has been investigated using GDSII-level layouts. To generate the layouts of the future 3D ICs, a 22 and a 16 nm process and standard cell libraries have been created and validated. With these realistic libraries and a 45 nm library, several 3D IC layouts have been generated and various quality metrics such as the footprint area, wirelength, critical path delay, and power consumption have been compared. The simulation results show that (1) footprint area is strongly dependent on the TSV size, so the use of sub-micron TSVs is the most important factor for area reduction; (2) wirelength is also dependent on the TSV size, but if the TSV size is sufficiently small (e.g., 0.5 μm TSVs for 16 nm technology), shrinking the TSV size further does not help reduce the wirelength; (3) critical path delay is strongly dependent on the

TSV capacitance, but the footprint area also has a non-negligible effect on critical path delay; (4) transition from 2D ICs to 3D ICs does not necessarily lead to less power consumption even when the TSV capacitance is small.

References

1. Kim DH, Athikulwongse K, Healy MB, Hossain MM, Jung M, Khorosh I, Kumar G, Lee Y-J, Lewis DL, Lin T-W, Liu C, Panth S, Pathak M, Ren M, Shen G, Song T, Woo DH, Zhao X, Kim J, Choi H, Loh GH, Lee H-HS, Lim SK. 3D-MAPS: 3D massively parallel processor with stacked memory. In: Proc. IEEE int. solid-state circuits conference, 2012. pp. 188–90.
2. Kim DH, Athikulwongse K, Lim SK. Study of through-silicon-via impact on the 3D stacked IC layout. IEEE Trans Very Large Scale Integr VLSI Syst. 2013;21(5):862–74.
3. Panth S, Samadi K, Du Y, Lim SK. Design and CAD methodologies for low power gate-level monolithic 3D ICs. In: Proc. IEEE int. symposium on low power electronics and design, 2014.
4. Jung M, Song T, Wan Y, Peng Y, Lim SK. On enhancing power benefits in 3D ICs: block folding and bonding styles perspective. In: Proc. ACM/IEEE design automation conference, 2014.
5. Kim DH, Kim S, Lim SK. Impact of sub-micron through-silicon vias on the quality of today and future 3D IC designs. In: Proc. ACM/IEEE int. workshop on system level interconnect prediction, 2011.
6. Farhane RE, Assous M, Leduc P, Thuaire A, Bouchu D, et al. A successful implementation of Dual Damascene architecture to copper TSV for 3D high density. In: Proc. IEEE int. 3D systems integration conf., 2010.
7. Motoyoshi M. Through-silicon via (TSV). Proc IEEE. 2009;97(1):43–8.
8. Koyanagi M, Fukushima T, Tanaka T. High-density through silicon vias for 3-D LSIs. Proc IEEE. 2009;97(1):49–59.
9. Kim DH, Mukhopadhyay S, Lim SK. Through-silicon-via aware interconnect prediction and optimization for 3D stacked ICs. In: Proc. ACM/IEEE int. workshop on system level interconnect prediction, 2009. pp. 85–92.
10. Radosavljevic M, et al. Electrostatics improvement in 3-D tri-gate over ultra-thin body planar InGaAs quantum well field effect transistors with high-K gate dielectric and scaled gate-to-drain/gate-to-source separation. In: Proc. IEEE int. electron devices meeting, 2011.
11. Hsu S, et al. A 280 mV to 1.1 V 256b reconfigurable SIMD vector permutation engine with 2-dimensional shuffle in 22 nm CMOS. In: Proc. IEEE int. solid-state circuits conference, 2012.
12. PTM. Predictive Technology Model. http://ptm.asu.edu.
13. Bai P, et al. A 65 nm logic technology featuring 35 nm gate lengths, enhanced channel strain, 8 Cu interconnect layers, low-k ILD and 0.57 μm^2 SRAM cell. In: Proc. IEEE int. electron devices meeting, 2004.
14. Mistry K, et al. A 45 nm logic technology with high-k + metal gate transistors, strained silicon, 9 Cu interconnect layers, 193 nm dry patterning, and 100 % Pb-free packaging. In: Proc. IEEE int. electron devices meeting, 2007.
15. Packan P, et al. High performance 32 nm logic technology featuring 2nd generation high-k + metal gate transistors. In: Proc. IEEE int. electron devices meeting, 2009.
16. ITRS. International Technology Roadmap for Semiconductors 2007 Edition Interconnect. http://www.itrs.net.
17. Nangate. Nangate FreePDK45 Open Cell Library. http://www.nangate.com.
18. Pathak M, Lee Y-J, Moon T, Lim SK. Through-silicon-via management during 3D physical design: when to add and how many? In: Proc. IEEE int. conf. on computer-aided design, 2010. pp. 387–94.
19. Cadence Design Systems. Encounter digital implementation system. http://www.cadence.com.

20. Lee Y-J, Lim SK. Timing analysis and optimization for 3D stacked multi-core microprocessors. In: Proc. int. 3D system integration conference, 2010.
21. Synopsys. PrimeTime. http://www.synopsys.com.
22. Kim YS, Tsukune A, Maeda N, Kitada H, Kawai A, et al. Ultra thinning 300-mm wafer down to 7-μm for 3D wafer integration on 45-nm node CMOS using strained silicon and Cu/low-k interconnects. In: Proc. IEEE int. electron devices meeting, 2009. pp. 14.6.1–4.
23. Kim DH, Athikulwongse K, Lim SK. A study of through-silicon-via impact on the 3D stacked IC layout. In: Proc. IEEE int. conf. on computer-aided design, 2009. pp. 674–80.
24. Kim DH, Mukhopadhyay S, Lim SK. TSV-aware interconnect length and power prediction for 3D stacked ICs. In: Proc. IEEE int. interconnect technology conference, 2009. pp. 26–8.
25. Kim DH, Lim SK. Impact of through-silicon-via scaling on the wirelength distribution of current and future 3D ICs. In: Proc. IEEE int. interconnect technology conference, 2011.

Chapter 2
3D Integration Technology

Yuan Xie and Qiaosha Zou

Abstract The emerging three-dimensional (3D) chip architectures, with their intrinsic capability of reducing the wire length, is one of the promising solutions to mitigate the interconnect problem in modern microprocessor designs. To leverage the benefits of fast latency, high bandwidth, and heterogeneous integration capability that are offered by 3D technology, new design methodologies should be developed targeting the unique feature of 3D integration. In this chapter, various approaches to model 3D electrical behavior, handle 3D thermal reliability problems, and design future 3D microprocessors are surveyed.

2.1 Introduction

With continued technology scaling, interconnect has emerged as the dominant source of circuit delay and power consumption. The reduction of interconnect delay and power consumption are of paramount importance for deep-sub-micron designs. Three-dimensional integrated circuits (3D ICs) [11] are attractive options for overcoming the barriers in interconnect scaling, thereby offering an opportunity to continue performance improvements using CMOS technology.

3D integration technologies offer many benefits for future microprocessor designs. Such benefits include: (1) *The reduction in interconnect wire length*, which results in improved performance and reduced power consumption; (2) *Improved*

This chapter includes portions reprinted with permission from the following publications: Qiaosha Zou, Tao Zhang, Eren Kursun, and Yuan Xie. Thermomechanical stress-aware management for 3D IC designs. Proceedings of Design, Automation Test in Europe Conference Exhibition (DATE) (2013). Copyright 2013 IEEE.

Y. Xie (✉)
University of California, Santa Barbara, CA, USA
e-mail: yuanxie@ece.ucsb.edu

Q. Zou
The Pennsylvania State University, State College, PA, USA
e-mail: qszou@cse.psu.edu

R.O. Topaloglu (ed.), *More than Moore Technologies for Next Generation Computer Design*, DOI 10.1007/978-1-4939-2163-8_2

memory bandwidth, by stacking memory on microprocessor cores with TSV connections between the memory layer and the core layer; (3) *The support for realization of heterogeneous integration*, which could result in novel architecture designs. (4) *Smaller form factor*, which results in higher packing density and smaller footprint due to the addition of a third dimension to the conventional two dimensional layout, and potentially results in a lower cost design.

To design the 3D microprocessor that can fully leverage the benefits of fast latency, higher bandwidth, and heterogeneous integration capability that are offered by 3D technology, understanding of 3D electrical behavior is necessary and mature techniques should be developed to handle the unique thermal reliability challenge in 3D technology.

This chapter first presents the background on 3D integration technology, and then reviews the models to capture the 3D electrical behaviors. The challenges of 3D thermal reliability are then presented. Various approaches to design future 3D microprocessors, leveraging the benefits from 3D technology, are surveyed. The challenges for future 3D architecture design are also discussed in the last section.

2.2 3D Integration Technology

The 3D integration technologies [56, 57] can be classified into one of the two following categories. (1) *Monolithic approach*. This approach involves sequential device process. The frontend processing (to build the device layer) is repeated on a single wafer to build multiple active device layers before the backend processing builds interconnects among devices. (2) *Stacking approach*, which could be further categorized as wafer-to-wafer, die-to-wafer, or die-to-die stacking methods. This approach processes each layer separately, using conventional fabrication techniques. These multiple layers are then assembled to build up 3D IC, using bonding technology. Since the stacking approach does not require the change of conventional fabrication process, it is easier to adopt compared to the monolithic approach, and has become the focus of recent 3D integration research.

Several 3D stacking technologies have been explored recently, including wire bonded, microbump, contactless (capacitive or inductive), and *through-silicon vias (TSV)* vertical interconnects [11]. Among all these integration approaches, TSV-based 3D integration has the potential to offer the greatest vertical interconnect density, and therefore is the most promising one among all the vertical interconnect technologies. Figure 2.1 shows a conceptual 2-layer 3D integrated circuit with TSV and microbump.

3D stacking can be carried out using two main techniques [16]: (1) *Face-to-Face (F2F)* bonding: two wafers/dies are stacked so that the very top metal layers are connected. Note that the die-to-die interconnects in face-to-face wafer bonding does not go through a thick buried Silicon layer and can be fabricated as *microbump*. The connections to C4 I/O pads are formed as TSVs; (2) *Face-to-Back (F2B)* bonding: multiple device layers are stacked together with the top metal layer of

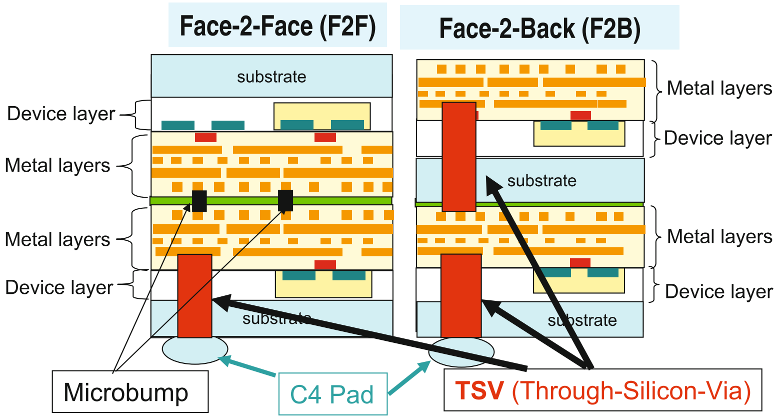


Fig. 2.1 Illustration of F2F and F2B 3D bonding

one die bonding together with the substrate of the other die, and direct vertical interconnects (which are called *through-silicon vias (TSV)*) tunneling through the substrate. In such F2B bonding, TSVs are used for both between-layer-connections and I/O connections. Figure 2.1 shows two conceptual 2-layer 3D ICs with F2F and F2B bonding, with both TSV connections and microbump connections between layers.

All TSV-based 3D stacking approaches share the following three common process steps [16]: (a) *TSV formation*; (b) *Wafer thinning* and (c) *Wafer alignment or die bonding*, which could be wafer-to-wafer(W2W) bonding or die-to-wafer(D2W) bonding. Wafer thinning is used to reduce the area impact of TSVs. The thinner the wafer, the smaller (and shorter) the TSV is (with the same aspect ratio constraint) [16]. The wafer thickness could be in the range of 10–100 μm and the TSV size is in the range of 0.2–10 μm [11].

In TSV-based 3D stacking bonding, the dimension of the TSVs is not expected to scale at the same rate as feature size because alignment tolerance during bonding poses limitation on the scaling of the vias. The vertical connection size, length, and the pitch density, as well as the bonding method (face-to-face or face-to-back bonding, SOI-based 3D or bulk CMOS-based 3D), can have a significant impact on the 3D microprocessor design. For example, relatively large size of TSVs can hinder partitioning a design at fine granularity across multiple device layers, and make the true 3D component design less possible. On the other hand, the monolithic 3D integration provides more flexibility in vertical 3D connection because the vertical 3D via can potentially scale down with feature size due to the use of local wires for connection. Availability of such technologies makes it possible to partition a design at a very fine granularity. Furthermore, face-to-face bonding or SOI-based 3D integration may have a smaller via pitch size and higher via density than face-to-back bonding or bulk-CMOS-based integration. Such influence of the 3D technology parameters on the microprocessor design must be thoroughly studied before an appropriate partition strategy is adopted.

2.3 TSV-Based 3D Electrical Model

For circuit performance (delay, power consumption, and heat dissipation) estimation, RLC model is the most straightforward modeling method by treating the device as composed of resistance, capacitance and inductance. In recently explored 3D stacking approach, TSV is working as the key enabling component, making the electrical modeling of TSV nontrivial. Therefore, this section primarily reviews the previous work that modeling TSV as RLC model under different conditions (frequency, temperature, etc.), followed by the brief introduction of electrical modeling of microbumps and back-end-of-line (BEOL).

2.3.1 TSV RC Model in Low Frequency Region

The TSV geometry description influences the final modeling accuracy. Most papers assume that TSVs are equivalent cylindrical structure [26, 44, 52, 58] and this assumption is examined by paper [44]. The researchers compared the electrical parameters extracted from Ansoft electromagnetic simulation tool with two geometry descriptions. The first one is a cylinder structure containing both top and bottom copper landing pads and another is the proposed simple structure without landing. The results show that only less than a 7 % difference is found in the RLC value, indicating that using simple cylindrical structure is sufficient for TSV modeling. However, this examination is performed with frequency as high as 1 GHz under stationary temperature. The conclusion may not be applicable for higher frequency beyond this point.

Most of previous work are on developing simple analytical model for TSV delay estimation. In [26], a physical dimension dependent analytical model for the propagation delay of TSVs is proposed. A lumped element model using dimensional analysis method is proposed with the observation that TSVs have a MOS-like capacitor structure [43]. Similarly, a lumped TSV model and the corresponding TSV propagation delay analysis are demonstrated in [23]. Besides the TSV's structure, the process method also influences the TSV electrical characteristic. An electrical and reliability study based on a fabricated via last TSV is presented in [35]. As it is illustrated in the work, the structural and material parameters both have impact on 3D TSV electrical characteristic. A 3D full wave and SPICE circuit simulation is performed and eye-diagrams at different frequencies are used to study the impact [38].

The analytical models from above mentioned work can provide electrical behavior analysis, however, closed-form equations are needed for real value calculation. The RLC model for single TSV and coupled TSVs with closed-form expressions are given [8, 23, 44, 52]. In [8], in addition to the analysis, a guard ring structure is proposed to suppress the noise coupling in TSVs. Closed-form expressions derived in [44] consider various effects, such as skin effect, therefore, the expression is relatively complicated with several parameters that need to be determined based on the given operation frequency. The expression is consistent with simulation results up to 2 GHz frequency in the paper. The expressions given in this paper are accurate, however, they are not suitable for fast circuit simulation. Empirical parameters are used in [52], which is more practical for full chip circuit simulation. Due to the relatively large size of TSVs, coupling effect is usually prominent which should be taken into consideration for full chip analysis. In the following section, the low frequency RLC equations for isolated TSV are introduced followed by the expressions for the coupling capacitance and mutual inductance in TSV grid.

2.3.1.1 RLC Model for an Isolated TSV

For an isolated TSV, the RLC model is shown in Fig. 2.2a. Capacitances exist between TSV and the adjacent substrate while resistance and inductance are in series along the TSV.

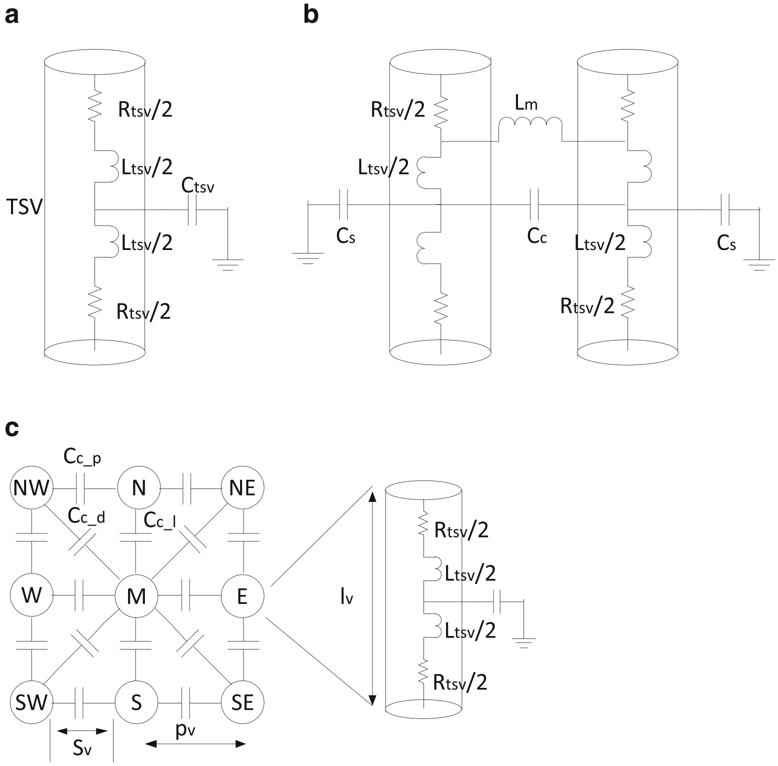


Fig. 2.2 The resistance, inductance, and capacitance components for (**a**) an isolated TSV; (**b**) two coupled TSVs; (**c**) a TSV bundle [52]

The resistance calculation can be described as the function of TSV conductivity (σ), TSV length (l), and radius (r):

$$R_{tsv} = \frac{l_v}{\sigma \pi r_v^2} \tag{2.1}$$

TSV has a MOS-like capacitor structure, therefore, the effective capacitance in TSV is the depletion capacitance and the oxide capacitance acting in series. During TSV formation, a dielectric layer is deposited between TSV metal and surrounding silicon, which makes TSV has the similar MOS structure. Based on the MOS effect modeling, a depletion region appears with introduced depletion capacitance. The depletion region width is determined by the voltage on TSV, threshold voltage (derived from flatband energy), interface charge density, and material properties. The depletion width changes with correspondence to the bias voltage when other conditions are fixed [59]. The final effective capacitance is a function of its geometry and the effective permittivity (ϵ_0) of surrounding dielectric liner. The following expression is based on empirical formula which assumes the thickness of dielectric layer is smaller than 1 μm:

$$C_{tsv} = \frac{63.36\varepsilon_0 l_v}{ln\left(1 + 5.26\frac{l_v}{r_v}\right)} \tag{2.2}$$

When the TSV is treated as a lossy transmission line in the model, the inductance has great impact on signal propagation delay. The propagation delay study in [26] shows that without the presence of inductance in TSVs, the average error is 55.2 % higher than the value of the distributed RLC model. The inductance of an isolated TSV is depended on the geometry parameters. It can be expressed as follows:

$$L_{tsv} = \frac{\mu l_v}{2\pi} ln\left(1 + \frac{2.84}{\pi}\frac{l_v}{r_v}\right) \tag{2.3}$$

All the above empirical closed-form equations are verified by a 3D/2D quasi-static electromagnetic-field solver tool and results show that the maximum error is within 6 %. By using these closed-form expressions, the resistance, capacitance, and inductance values in a single TSV can be easily calculated for fast circuit simulation.

2.3.1.2 RLC Model for Coupled TSVs

The RLC models for two coupled TSVs and a TSV bundle are shown in Fig. 2.2b, c, respectively. For coupled TSVs, the resistance expression is the same as that in an isolated TSV since coupling effect has negligible impact on the resistance.

But for inductance and capacitance, the inter-via coupling effects are prominent. In the following analysis, capacitance and inductance are divided into two parts: self parameter and mutual parameter.

The capacitance of the whole coupled bundle TSVs can be expressed as follows:

$$C_{bundle} = \begin{bmatrix} C_{1,1} & -C_{1,2} & \dots & -C_{1,n} \\ -C_{2,1} & C_{2,2} & \dots & -C_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ -C_{n,1} & -C_{n,2} & \dots & C_{n,n} \end{bmatrix} \tag{2.4}$$

The diagonal element means the sum of self and inter-via coupling capacitances. This capacitance matrix is sparse because only the diagonal elements and elements that represent nearest neighbors contain meaningful values. From the researchers' experiments [52], for a 7 × 7 TSV bundle, coupling terms for nearest neighbors are more significant than those that are non-adjacent.

The self capacitance formula is different from the isolated TSV, which is given as:

$$C_s = C_{tsv} - k_1 C_{tsv} e^{(k_2 \frac{pv}{r_v} + k_3 \frac{pv}{l_v})} \left[k_4 \left(\frac{L_v}{r_v} \right)^{k_5} + k_6 \left(\frac{p_v}{r_v} \right)^{k_7} + k_8 \right] \tag{2.5}$$

where C_{tsv} is the capacitance of an isolated TSV, the parameters from k_1 to k_8 are empirical constants. However, these constants are based on the simulation results and varied with different TSV configurations, making it hard to be directly used in circuit simulation. When k_2 and k_3 are negative and p_v approaches infinity, C_s equals to C_{tsv}.

The formula for the coupling capacitance where $i \neq j$ in the matrix is given as follows:

$$C_{coupled} = \frac{k_1 \epsilon_0 l_v}{ln(k_2 \frac{pv}{rv})} \left[1 + k_5 \left(\frac{L_v}{r_v} \right)^{k_6} + k_3 \left(\frac{p_v}{r_v} \right)^{k_4} + k_7 \left(\frac{p_v}{l_v} \right)^{k_8} \right] \tag{2.6}$$

The coupling inductance terms is defined similarly to the coupling capacitance. Different from capacitance, inductive coupling effect has long range, therefore, the matrix is not sparse. The mutual inductance between any two TSVs can be captured with the following formula:

$$L_m = 0.199 \mu l_v ln \left(1 + 0.438 \frac{d_v}{l_v} \right) \tag{2.7}$$

where d_v is the center-to-center distance between two TSVs.

Based on the simulation results, the maximum error for coupling capacitance and inductance are within 6 and 8 %, respectively.

2.3.2 TSV RLCG Model in High Frequency Region

Previous content introduces previous work that use simple RLC model in low frequency region. In the following section, the relatively complex RLCG model is introduced for high frequency operation which is up to 100 GHz [41, 59]. In this model, the substrate is not assumed as ideal conductor, therefore, the impact of substrate resistance is taken into consideration. The RLCG model contains two components: admittance per unit TSV height which consists of conductance and susceptance; impedance per unit TSV height which is composed of resistance and reactance.

Skin effect and eddy currents in silicon should be also taken into consideration for TSV modeling at high frequencies [59]. Skin effect means the current density drops by a certain factor below the surface of a conductor. It has great impact on the high frequency resistance.

The RLCG model developed in [59] is introduced in detail. The equivalent distributed circuit model is shown in Fig. 2.3a, the simplified model is given in Fig. 2.3b. The impedance which is represented by Z is inside TSV, similar to the resistance and inductance in series in RLC model. Capacitance $C1$ resides between TSV and substrate representing the final effective capacitance (oxide capacitance and depletion capacitance in series). Admittance, represented by Y, exists in the silicon substrate between two adjacent TSVs. In this figure, Y_{open} is the input admittance between ports 1 and 2 if ports 3 and 4 are open while Z_{short} represents the impedance between ports 1 and 2 when ports 3 and 4 are short circuited.

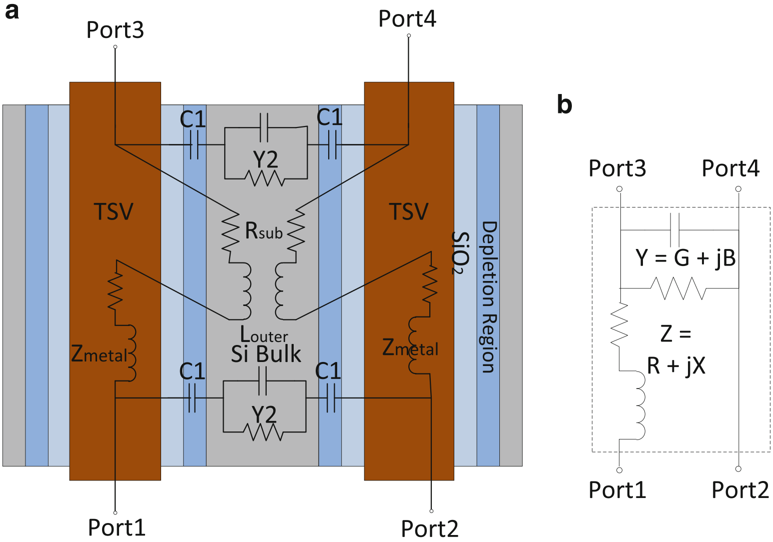


Fig. 2.3 RLCG model for TSV, (**a**) the equivalent distributed RLCG model of two coupled TSVs; (**b**) a simplified distributed transmission line model [59]

2.3.2.1 Admittance of TSV in RLCG Model

The admittance (CG) per unit TSV height can be treated as two components working in series, one is the effective capacitance (C_1) and the other is the coupling admittance (Y_2) due to the bulk silicon. The admittance expression is shown in the following equation:

$$Y = [2(j\omega C_1)^{-1} + Y_2^{-1}]^{-1} \tag{2.8}$$

where ω is the radial frequency. Since there are two TSVs contributing to C_1 in series with Y_2, the equation contains a factor of 2 before C_1. The detailed equations to calculate C_1 and Y_2 can be found in [59].

The CG model is verified with a 2-D quasi-electrostatic simulation tool. The results suggest that at low frequencies, if the depletion region is not considered, the error is not negligible, however this difference is not so significant at high frequencies.

2.3.2.2 Impedance of TSV in RLCG Model

The serial impedance (RL) per unit height is not so straightforward. The final expression results are shown here without detailed deduction steps. For simplicity, the serial impedance can be treated as the sum of three components: the inner impedance of TSV (Z_{metal}), the outer inductance (L_{outer}), and the resistance due to eddy currents in silicon substrate (R_{sub}), where the equations for these three components can be found in [59]:

$$Z = 2Z_{metal} + j\omega L_{outer} + R_{sub} \tag{2.9}$$

The model is compared with the simulation tool and the results indicate skin effect in TSV is of great importance for high frequency analysis when the higher frequency resistance is dominant over DC resistance.

2.3.2.3 TSV Electrical Performance with RLCG Model

As technology scales, the diameter and pitch of TSVs shrink, however, the substrate thickness almost remains the same as predicted by the ITRS. When the radius of TSVs reduce, C, G and L do not change much due to the proportional scaling of geometrical parameters. Nevertheless, resistance increases significantly when the frequency reaches the region of tens of GHz due to the decreasing TSV cross sectional area.

In terms of circuit performance sensitivity, capacitance has the most important impact on circuit behavior while resistance is of the least importance. The interconnect exhibits the short-transmission line behavior on signal propagation, which indicates that simple RLC model is enough for delay and signal rise/fall calculation. However, the L and G are crucial factors for the estimation of voltage variations in V_{DD} and *GND*. More accurate whole circuit performance evaluation can only be done with all RLCG models.

2.3.3 TSV RC Model with Temperature Consideration

Although 3D stacking provides a number of benefits over traditional 2D circuits, 3D exacerbates the thermal dissipation problems due to higher power density in smaller footprint. The temperature gradients on chip result in TSV electrical characteristic variation. Several work have explored the temperature dependent TSV modeling [22, 24, 55].

Due to the complexity of temperature-dependent TSV modeling, only semi-analytical capacitance model and empirical RC model are brought out in previous work. First, the semi-analytical capacitance model is briefly introduced. Then the empirical RC model formulations are given for straightforward RC value computation.

2.3.3.1 Semi-Analytical Temperature-Dependent Capacitance Model

This model is called semi-analytical because the close-form expression is not given, instead, a four-step algorithm is given to calculate the capacitance until the convergence conditions are satisfied. Normally, the behavior of TSV is similar to a MOS capacitor, and the analytical expression for TSV capacitance is derived by solving a 1D Poisson equation in the radial direction in a cylindrical coordinate system. Considering the depletion region, a semi-analytical algorithm for depletion capacitance calculation is proposed [24].

The algorithm first identifies the initial maximum depletion radius. The assumption neglects the hole and electron charges. The initial maximum depletion radius can be obtained from the following equation:

$$\frac{qN_aR_{OX}^2}{4\varepsilon_{Si}} - \frac{qN_aR_{max}^2}{2\varepsilon_{Si}}Ln(R_{OX}) + \frac{qN_aR_{max}^2}{4\varepsilon_{Si}}(2ln(R_{max}) - 1 = \psi_s \quad (2.10)$$

with the assumption that the surface potential ψ_s equals to $2(K_BT/q)ln(N_a/n_i)$. In the equation, q is the electron charge, N_a is the density of ionized acceptors or the doping concentration, ε is the silicon permittivity, and $\psi(r)$ represents the electrostatic potential with respect to the radius.

The second step is trying to identify electron-hole densities in the substrate from the potential with the initial depletion radius calculated from previous step. The potential at every point is calculated as follows:

$$\psi(r) = \frac{qN_a r^2}{4\varepsilon_{Si}} - \frac{qN_a R_{max}^2}{2\varepsilon_{Si}} ln(r) + \frac{qN_a R_{max}}{4\varepsilon_{Si}}(2ln(R_{max} - 1) \tag{2.11}$$

The value of the potential at distance r is used to compute the hole and electron charge densities in the substrate using $p(r) = p_{Po}exp(-\beta\psi(r))$ and $n(r) = n_{Po}exp(\beta\psi(r))$.

Step three calculates the new maximum depletion radius with consideration of the hole and electron charge densities derived from the previous step. The new maximum depletion radius is calculated from the following equation:

$$\frac{q(N_a + p - n)R_{OX}^2}{4\varepsilon_{Si}} - \frac{q(N_a + p - n)R_{max}^2}{2\varepsilon} ln(R_{OX}) + \frac{q(N_a + p - nR_{max}^2)}{4\varepsilon_{Si}}(2ln(R_{max}) - 1) = \psi_s \tag{2.12}$$

The last step calculates the depletion capacitance by using equation $C_{dep} = 2\pi\varepsilon_{Si}L_{TSV}/ln(2R_{max}/\phi_{TSV})$. The final depletion capacitance is obtained by continuing these four steps until the new depletion radius approaches the initial maximum depletion radius. The total TSV capacitance can be viewed as the oxide capacitance and depletion capacitance in series.

Comparison between the semi-analytical results and the measurement results shows that the error is within 3 %. When the temperature rises, the TSV capacitance increases due to the reduction of maximum depletion radius.

2.3.3.2 Empirical Temperature-Dependent RC Model

Besides the TSV capacitance, resistance also changes with temperature variation. Lumped RC model should be enhanced by considering TSV capacitance and resistance change due to temperature variation [22]. However, due to lack of close-form expression for temperature-dependent resistance, [22] builds a 2D/3D ring oscillator to measure the model parameters at different temperatures. Thus, an empirical RC model of TSV is discussed. This RC model is similar to the simple signal-transmission line model at DC and low frequencies without inductances.

The expressions of resistance and capacitance from empirical data are given in the following:

$$R_{TSV}(T) = R_0(1 + \alpha(T - T_0)) \tag{2.13}$$

$$C_{TSV}(T) = 0.0007T^2 - 0.0333T + 44.4 \tag{2.14}$$

The measurement results suggest that with temperature rise, substantial increment in TSV capacitance and resistance can be seen.

The temperature-dependent RLC model is still far beyond maturity in current research. Furthermore, from these work, we can see that the resistance and capacitance have great dependency on temperature. Moreover, these dependencies can be translated into further influence on the on-chip temperature by producing Joule heating [55]. Accurate modeling and electrical-thermal co-analysis framework are required for precise circuit performance and on-chip temperature estimation.

2.3.4 RC Model for Microbumps, RDL, and BEOL

In addition to TSV which is the key enabling component in 3D ICs, other components (microbumps, redistribution layers, C4, etc.) are necessary for electrical modeling to perform full chip and package analysis. Several work modeled the redistribution layer (RDL), back-end-of-line (BEOL) and microbumps [1,40,54].

The BEOL and RDL can be treated as traditional metal layers for signal transmission with resistance and capacitance. The theoretical values of resistance for RDL can be calculated from [40]:

$$R_{Th} = R_{RDL} + \frac{R_{Ground}}{2} = \alpha \frac{1}{w_{RDL} * t_{RLD}} \frac{3}{2} \tag{2.15}$$

where w_{RDL} and t_{RDL} are the width and the oxide thickness of the metal layers. The resistance value of BEOL can be obtained in a similar way. The capacitance calculation is absent in previous work.

Microbump can not be simply treated as metal layers due to the structure difference. In [54], the microbump has the RLC model similar to the TSV and shares the same expressions. The microbump model contains resistance and inductance in series from input to output port. Two capacitances reside between microbump and connected tiers. One of the capacitance represents the capacitance between microbump and substrate of the first tier while the other captures the capacitance between microbump and the substrate of the second tier.

2.4 Thermal Stress-Aware Design for 3D ICs

Stacked chips on 3D architecture increase the packaging density and thermal resistances, which results in higher on-chip temperatures. Plenty of studies have focused on the 3D thermal modeling, analysis [6, 7, 19, 51], and thermal-aware design methodology [9, 10, 17] to manage the on-chip thermal issues of 3D ICs. These work, however, failed to consider the TSV lateral thermal blockage effect and thermomechanical stress. Moreover, they used TSVs as thermal vias to build vertical heat dissipation path, which in turn results in increased thermal load on TSVs as well as thermomechanical stresses, and thus weakens the reliability. On the other hand, prior work on analyzing the mechanical stresses in 3D ICs [2, 21] only consider the static stress management by adjusting TSV keep-out zone size, TSV placement, or TSV structure. In [61], the work not only accounts for the static (design-time) management of TSV thermal stress and thermal load but also takes into account the run-time TSV stress analysis and management.

For better thermal management, profound understandings of 3D heat transfer and cycling effects are necessary. Accurate 3D ICs thermal modelings have been conducted. An analytical and numerical model for temperature distribution in a 3D stack considering multiple heat sources is developed to help 3D thermal analysis [19]. An analytical thermal model for the top layer in 3D architecture with TSVs vertical thermal conductivity model is proposed to determine TSVs density during design time [51]. TSV is one of the most important component in 3D ICs, precise thermal modeling of TSVs can significantly improve the thermal analysis of 3D architectures. The equivalent thermal conductivity model [7] and lateral thermal blockage model [6] of TSVs are demonstrated. The thermal modeling for both silicon devices and TSVs in vertical and horizontal directions should be used for precise temperature analysis in 3D architectures.

Based on the thermal modeling and analysis of 3D architectures, several work have performed thermal-aware design. Thermal-aware 3D design placement techniques with TSVs for thermal vias are introduced to alleviate the on-chip temperature [9, 10, 17]. These work, however, fail to take the TSV lateral thermal blockage effects into consideration. As thermal vias, TSVs are likely to place near hotspot for vertical thermal dissipation, but the lateral blockage effects may worsen the thermal problem in horizontal direction.

The above mentioned work on thermal-aware design only makes effort on reducing the on-chip temperature without considering the thermomechanical stresses related reliability issues in 3D architectures. Analysis of reliability problems induced by thermomechanical stresses and strains is performed but it includes one single TSV [3]. Full-chip thermomechanical stresses and reliability analysis tool is generated to alleviate the reliability problems in 3D ICs [21]. The 3D FEA (finite element analysis) simulations are performed to examine the effect of TSV structure and liner material/thickness on TSV radial stress. Superposition method which is proved to be effective is applied for full chip analysis with TSV bundles. Besides the thermal stress analysis, stress-aware reliability schemes are also developed. Both design-time and run-time thermal stress management strategies are developed with the consideration of TSV horizontal thermal blockage effects in [61]. During design time, the management scheme tries to reduce the thermal load on TSV to reduce the TSV thermal stress, preventing the early time TSV interfacial delamination and wafer cracking. Moreover, thermal cycling effect is considered during run-time and thermal control mechanism is used to provide mechanical equilibrium for whole chip reliability.

The detailed TSV thermal stress model and 3D thermal cycling effect are introduced in the following subsection.

2.4.1 Analysis of TSV Thermal Stress

In 3D IC fabrication, copper (Cu) is usually used as TSV filling material. Copper has more than five times larger coefficient of thermal expansion (CTE) than silicon. The CTE mismatch between TSV and silicon substrate in turn introduces mechanical stresses that can lead to high probability of die cracking and interfacial delamination [33, 34, 45]. The coefficients of thermal expansion of TSV materials and silicon are listed in Table 2.1. The CTE of four possible TSV materials are all larger than the CTE of silicon substrate. As an example, Fig. 2.4 illustrates the potential cracking and delamination damage. Once heating is applied to the die, TSVs tend to expand much faster than silicon; this finally results in TSVs stretching out of silicon substrate. As a consequence, damage is generated in back-end-of-line (BEOL) and wire layers [33]. On the other hand, the contracted TSVs pull the surface of surrounding silicon during the cooling process, causing surface delamination and tensile stress in the surrounding region. Since silicon substrate is thinned drastically to expose TSVs, it is more vulnerable to mechanical stresses than 2D circuits.

Table 2.1 Coefficient of thermal expansion of TSV materials and silicon [42]

Material	CTE (ppm/K)
Silicon	2.3
Copper	17
Aluminium	20
Tungsten	4.4
Nickel	13

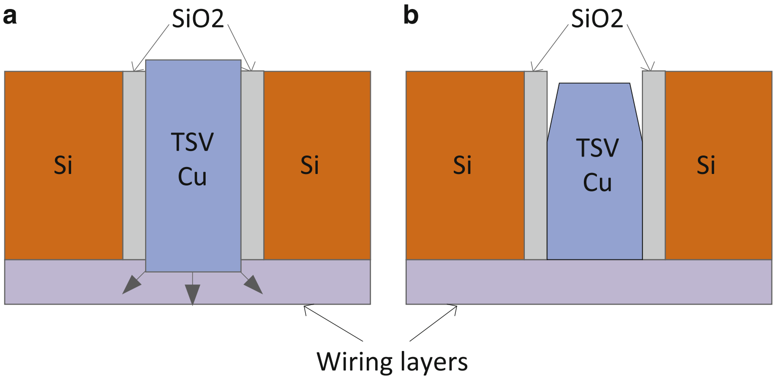


Fig. 2.4 TSV thermal expansion and delamination due to CTE mismatch. (**a**) TSV expansion during heating; (**b**) the delamination between TSV and silicon during cooling [33]

To minimize thermomechanical stresses, TSV farms should be placed smartly during design time. Therefore, the corresponding analysis on the thermal stresses around TSVs is critical to the solution. Several work [32, 42] have targeted thermal stress analysis showing that the stress field in TSVs is uniform and can be represented by radial, circumferential, and axial stresses. The stresses can be expressed as following:

$$\sigma_r = \sigma_\theta = \frac{-E(\alpha_{tsv} - \alpha_{si})T_{tsv}}{2 - 2\upsilon}, \sigma_z = 2\sigma_\theta \tag{2.16}$$

where σ_r, σ_θ, and σ_z are radial, circumferential, and axial stresses, respectively. α_{tsv} is the CTE of TSVs and α_{si} represents the CTE of silicon. T_{tsv} is the thermal load on TSV, E is the Young's modulus and υ is the Poisson's ratio.[1]

TSVs thermal load estimation during design-time is usually based on accurate thermal modeling of TSVs and TSV temperature is used to represent the corresponding thermal load assuming the stress-free temperature is at room temperature. Both vertical high thermal conductivity and lateral thermal blockage effect [6] should be considered in the TSV model for more accurate temperature modeling. The lateral thermal blockage effect is due to the relatively low thermal conductivity of dielectric layer surrounding TSV. For example, the normal dielectric layer material is SiO_2 with thermal conductivity of $1.4\,\mathrm{W\,m^{-1}\,K}$ compared to the silicon thermal conductivity of $149\,\mathrm{W\,m^{-1}\,K}$. Therefore, the thermal resistances exist between lateral TSV walls and all neighboring blocks, resulting in high thermal resistances on the lateral thermal dissipation path.

In general, the thermal resistance of TSV farms can be captured by:

$$R_{TSV} = \frac{h}{k \cdot A} \tag{2.17}$$

where h is the material thickness, k is the thermal conductivity of the material per volume, and A is the cross sectional area where heat flow passes through. Note that this equation can be used to calculate both vertical and lateral thermal resistance of TSVs. For vertical thermal resistance calculation, the TSV metal thermal conductivity is used, otherwise, lateral TSV farm thermal conductivity (including low thermal conductivity insulator) is adopted. To this end, the lateral heat blockage effect has been taken into account during design-time floorplan and the TSV thermal stress is proportional to the thermal load on TSVs.

[1] In this formula, the difference of elastic between materials is omitted for simplicity.

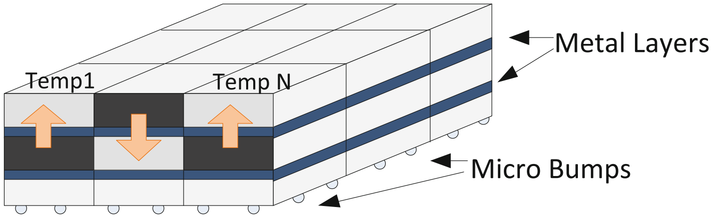


Fig. 2.5 Stack level thermal cycling effect in 3D structure. Thermal stresses are pointing from hot blocks (*dark color*) to cool blocks (*light color*). Alternating direction of stresses (the *arrows*) easily cause cracking on thinned substrate

2.4.2 3D Thermal Cycling Effects

Thermal cycling effect is another factor that can cause reliability issues for 3D ICs [36, 61]. Particularly, the thermal cycling effects in 3D ICs become prominent because the dynamic thermal gradients and stresses in x, y, z directions during run-time can no longer be ignored. Moreover, the thermal cycling effects are more complicated since each functional block now has two more proximity blocks in the vertical direction. As shown in Fig. 2.5 [61], the generated thermal expansion forces are highlighted by arrows, which are from the hotter blocks (dark color) to the cooler blocks (light color). When the force direction varies in the stacked chips, it makes the thinned silicon substrate more vulnerable to damage. A run-time thermal cycling management scheme should be proposed to eliminate the damaging thermal cycling pattern.

Most of the traditional 2D and 3D techniques can not mitigate the problem because they only strive to minimize the peak temperature on chip but disregard the thermomechanical stresses. Sometimes the traditional thermal management techniques can even worsen the problem by wrongly forcing the thermal patterns to exert maximum stress on the device layers in checkerboard configurations, where cold and hot structures are overlaid. In [61], the analysis of vertical thermal cycling pattern and temperature gradients between neighbors in x, y, z directions is employed as part of the thermomechanical stresses management scheme. As a result, the management scheme can alleviate the temperature gradients on cell granularity and achieve mechanical equilibrium.

2.5 Designing 3D Processor Architecture

The following subsections discuss various architecture design approaches that leverage different benefits that 3D integration technology can offer, namely, wirelength reduction, high memory bandwidth, heterogeneous integration, and cost reduction. They also briefly review 3D network-on-chip architecture designs.

2.5.1 Wirelength Reduction

Designers have resorted to technology scaling to improve microprocessor performance. Although the size and switching speed of transistors benefit as technology feature sizes continue to shrink, global interconnect wire delay does not scale accordingly with technologies. The increasing wire delays have become one major impediment to performance improvement.

Three-dimensional integrated circuits (3D ICs) are attractive options for overcoming the barriers to interconnect scaling, thereby offering an opportunity to continue performance improvements using CMOS technology. Compared to a traditional two dimensional chip design, one of the important benefits of a 3D chip over a traditional two-dimensional (2D) design is the reduction on global interconnects. It has been shown that three-dimensional architectures reduce wiring length by a factor of the square root of the number of layers used [20]. The reduction of wire length due to 3D integration can result in two obvious benefits: *latency improvement* and *power reduction*.

2.5.1.1 Latency Improvement

Latency improvement can be achieved due to the reduction of average interconnect length and the critical path length.

Early work on fine-granularity 3D partitioning of processor components shows that the latency of 3D components could be reduced. For example, since interconnects dominate the delay of cache accesses which determines the critical path of a microprocessor, and the regular structure and long wires in a cache make it one of the best candidates for 3D designs, 3D cache design is one of the early design example for fine-granularity 3D partition [56]. Wordline partitioning and bitline partitioning approaches divide a cache bank into multiple layers and reduce the global interconnects, resulting in fast cache access time. Depending on the design constraints, the 3DCacti tool [47] automatically explores the design space for a cache design, and finds out the optimal partitioning strategy, and the latency reduction can be as much as 25 % for a two-layer 3D cache. 3D arithmetic-component designs also show latency benefits. For example, various designs [15, 37, 39, 48] have shown that the 3D arithmetic unit design can achieve around 6–30 % delay reduction due to the wire length reduction. Such fine-granularity 3D partitioning was also demonstrated by Intel [4], showing that by targeting the heavily pipelined wires, the pipeline modifications resulted in approximately 15 % improved performance, when the Intel Pentium-4 processor was folded onto 2-layer 3D implementation.

Note that such fine-granularity design of 3D processor components increases the design complexity, and the latency improvement varies depending on the partitioning strategies and the underlying 3D process technologies. For example, for the same Kogge–Stone adder design, a partitioning based on logic level [48] demonstrates that the delay improvement diminishes as the number of 3D layers increases;

while a bit-slicing partitioning [37] strategy would have better scalability as the bit-width or the number of layers increases. Furthermore, the delay improvement for such bit-slicing 3D arithmetic units is about 6 % when using a bulk-CMOS-based 180 nm 3D process [15], while the improvement could be as much as 20 % when using a SOI-based 180 nm 3D process technology [37], because the SOI-based process has much smaller and shorter TSVs (and therefore much smaller TSV delay) compared to the bulk-CMOS-based process.

2.5.1.2 Power Reduction

Interconnect power consumption becomes a large portion of the total power consumption as technology scales. The reduction of the wire length translates into power savings in 3D IC design. For example, 7–46 % of power reduction for 3D arithmetic units were demonstrated in [37]. In the 3D Intel Pentium-4 implementation [4], because of the reduction in long global interconnects, the number of repeaters and repeating latches in the implementation is reduced by 50 %, and the 3D clock network has 50 % less metal RC than the 2D design, resulting in a better skew, jitter and lower power. Such 3D stacked redesign of Intel Pentium 4 processor improves performance by 15 % and reduces power by 15 % with a temperature increase of 14°. After using voltage scaling to lower the peak temperature to be the same as the baseline 2D design, their 3D Pentium 4 processor still showed a performance improvement of 8 %.

2.5.2 Memory Bandwidth Improvement

It has been shown that circuit limitations and limited instruction level parallelism will diminish the benefits of modern superscalar microprocessors by increased architectural complexity, which leads to the advent of Chip Multiprocessors (CMP) as a viable alternative to the complex superscalar architecture. The integration of multi-core or many-core microarchitecture on a single die is expected to accentuate the already daunting memory-bandwidth problem. Supplying enough data to a chip with a massive number of on-die cores will become a major challenge for performance scalability. Traditional off-chip memory will not suffice due to the I/O pin limitations. Three-dimensional integration has been envisioned as a solution for future micro-architecture design (especially for multi-core and many-core architectures), to mitigate the interconnect crisis and the "memory wall" problem [18,30,31]. It is anticipated that memory stacking on top of logic would be one of the early commercial uses of 3D technology for future chip-multiprocessor design, by providing improved memory bandwidth for such multi-core/many-core microprocessors. In addition, such approaches of memory stacking on top of core layers do not have the design complexity problem as demonstrated by the fine-granularity design approaches, which require re-designing all processor components for wire length reduction.

Intel [4] explored the memory bandwidth benefits using a base-line Intel Core2 Duo processor, which contains two cores. By having memory stacking, the on-die cache capacity is increased, and the performance is improved by capturing larger working sets, reducing off-chip memory bandwidth requirements. For example, one option is to stack an additional 8 MB L2 cache on top of the base-line 2D processor (which contains 4 MB L2 cache), and the other option is to replace the SRAM L2 cache with a denser DRAM L2 cache stacking. Their study demonstrated that a 32 MB 3D stacked DRAM cache can reduce the cycles per memory access by 13 % on average and as much as 55 % with negligible temperature increases.

PicoServer project [25] follows a similar approach to stack DRAM on top of multi-core processors. Instead of using stacked memory as a larger L2 cache (as shown by Intel's work [4]), the fast on-chip 3D stacked DRAM main memory enables wide low-latency buses to the processor cores and eliminates the need for an L2 cache, whose silicon area is allocated to accommodate more cores. Increasing the number of cores by removing the L2 cache can help improve the computation throughput, while each core can run at a much lower frequency, and therefore result in an energy-efficient many core design. For example, it can achieve a 14 % performance improvement and 55 % power reduction over a baseline multi-core architecture.

As the number of the cores on a single die increases, such memory stacking becomes more important to provide enough memory bandwidth for processor cores. Recently, Intel [49] demonstrated an 80-tile terascale chip with network-on-chip. Each core has a local 256 KB SRAM memory (for data and instruction storage) stacked on top of it. TSVs provide a bandwidth of 12 GB/s for each core, with a total about 1 TB/s bandwidth for Tera Flop computation. In this chip, the thin memory die is put on top of the CPU die, and the power and I/O signals go through memory to CPU.

Since DRAM is stacked on top of the processor cores, the memory organization should also be optimized to fully take advantages of the benefits that TSVs offer [29, 31]. For example, the numbers of ranks and memory controllers are increased, in order to leverage the memory bandwidth benefits. A multiple-entry row buffer cache is implemented to further improve the performance of the 3D main memory. Comprehensive evaluation shows that a 1.75× speedup over commodity DRAM organization is achieved [31]. In addition, the design of MSHR was explored to provided a scalable L2 miss handling before accessing the 3D stacked main memory. A data structure called the Vector Bloom Filter with dynamic MSHR capacity tuning is proposed. Such structure provides an additional 17.8 % performance improvement. If stacked DRAM is used as the last-level caches (LLC) in chip multiple processors (CMPs), the DRAM cache sets are organized into multiple queues [29]. A replacement policy is proposed for the queue-based cache to provide performance isolation between cores and reduce the lifetimes of dead cache lines. Approaches are also proposed to dynamically adapt the queue size and the policy of advancing data between queues.

The latency improvement due to 3D technology can also be demonstrated by such memory stacking design. For example, Li et al. [28] proposed a 3D chip multiprocessor design using network-in-memory topology. In this design, instead of partitioning each processor core or memory bank into multiple layers (as shown in [47, 56]), each core or cache bank remains to be a 2D design. Communication among cores or cache banks are via the network-on-chip (NoC) topology. The core layer and the L2 cache layer are connected with TSV-based bus. Because the short distance between layers, TSVs provide a fast access from one layer to another layer, and effectively reduce the cache access time because of the faster access to cache banks through TSVs.

2.5.3 Heterogenous Integration

3D integration also provides new opportunities for future architecture design, with a new dimension of design space exploration. In particular, the heterogenous integration capability enabled by 3D integration gives designers new perspective when designing future CMPs.

3D integration technologies provide feasible and cost-effective approaches for integrating architectures composed of heterogeneous technologies to realize future microprocessors targeted at the “More than Moore” technology projected by ITRS. 3D integration supports heterogeneous stacking because different types of components can be fabricated separately, and layers can be implemented with different technologies. It is also possible to stack optical device layers or non-volatile memories [such as magnetic RAM (MRAM) or phase-change memory (PCRAM)] on top of microprocessors to enable cost-effective heterogeneous integration. The addition of new stacking layers composed of new device technology will provide greater flexibility in meeting the often conflicting design constraints (such as performance, cost, power, and reliability), and enable innovative designs in future microprocessors.

2.5.3.1 Non-volatile Memory Stacking

Stacking layers of non-volatile memory technologies such as Magnetic Random Access Memory (MRAM) [13] and Phase Change Random Access Memory (PRAM) [53] on top of processors can enable a new generation of processor architectures with unique features. There are several characteristics of MRAM and PRAM architectures that make them promising candidates for on-chip memory. In addition to their non-volatility, they have zero standby power, low access power and are immune to radiation-induced soft errors. However, integrating these non-volatile memories along with a logic core involves additional fabrication challenges that need to be overcome (for example, MRAM process requires growing a magnetic stack between metal layers). Consequently, it may incur extra cost and additional fabrication complexity to integrate MRAM with conventional CMOS logic into a single 2D chip. The ability to integrate two different wafers developed with different technologies using 3D stacking offers an ideal solution to overcome this fabrication challenge and exploit the benefits of PRAM and MRAM technologies. For example, Sun et al. [46] demonstrated that the optimized MRAM L2 cache on top of multi-core processor can improve performance by 4.91 % and reduce power by 73.5 % compared to the conventional SRAM L2 cache with similar area.

2.5.3.2 Optical Device Layer Stacking

Even though 3D memory stacking can help mitigate the memory bandwidth problem, when it comes to off-chip communication, the pin limitations, the energy cost of electrical signaling, and the non-scalability of chip-length global wires are still significant bandwidth impediments. Recent developments in silicon nanophotonic technology have the potential to meet the off-chip communication bandwidth requirements at acceptable power levels. With the heterogeneous integration capability that 3D technology offers, one can integrate optical die together with CMOS processor dies. For example, HP Labs proposed a Corona architecture [50], which is a 3D many-core architecture that uses nanophotonic communication for both inter-core communication and off-stack communication to memory or I/O devices. A photonic crossbar fully interconnects its 256 low-power multithreaded cores at 20 TB/s bandwidth, with much lower power consumption.

Figure 2.6 illustrates such a 3D heterogenous processor architecture, which integrates non-volatile memories and optical die together through 3D integration technology.

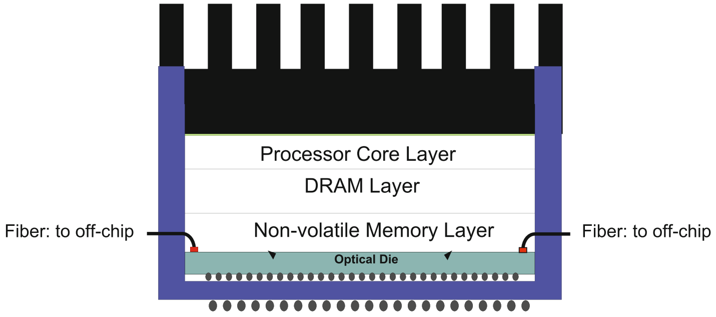


Fig. 2.6 An illustration of 3D heterogeneous architecture with non-volatile memory stacking and optical die stacking

2.5.4 Cost-Effective Architecture

Increasing integration density has resulted in large die size for microprocessors. With a constant defect density, a larger die typically has a lower yield. Consequently, partitioning a large 2D microprocessor to be multiple smaller dies and stacking them together may result in a much higher yield for the chip, even though 3D stacking incurs extra manufacturing cost due to extra steps for 3D integration and may cause a yield loss during stacking. Depending on the original 2D microprocessor die size, it may be cost-effective to implement the chip using 3D stacking [12], especially for large microprocessors. The heterogenous integration capability that 3D provides can also help reduce the cost.

In addition, as technology feature size scales to reach the physical limits, it has been predicted that moving to the next technology node is not only difficult but also prohibitively expensive. 3D stacking can potentially provide a cost-effective integration solution, compared to traditional technology scaling.

2.5.5 3D NoC Architecture

Network-on-chip (NoC) is a general purpose on-chip interconnection network architecture that is proposed to replace the traditional design-specific global on-chip wiring, by using switching fabrics or routers to connect processor cores or processing elements (PEs). Typically, the PEs communicate with each other using a packet-switched protocol. Even though both 3D integrated circuits and NoCs are proposed as alternatives for the interconnect scaling demands, the challenges of combining both approaches to design three-dimensional NOCs have not been addressed until recently [14, 27, 28]. Researchers have studied various NoC router design with 3D integration technology. For example, various design options for the NoC router for 3D NoC has been investigated: (1) symmetric NoC router design with a simple extension to the 2D NoC router; (2) NoC-bus hybrid router design which leverages the inherent asymmetry in the delays in a 3D architecture between the fast vertical interconnects and the horizontal interconnects that connect neighboring cores; (3) True 3D router design with major modification as dimensionally-decomposed router [27]; (4) Multi-layer 3D NoC router design which partitions a single router to multiple layers to boost the performance and reduce the power consumption [14]. 3D NoC topology design was also investigated [60]. More details can be found in [5].

2.6 Challenges for 3D Architecture Design

Even though 3D integrated circuits show great benefits, there are several challenges for the adoption of 3D technology for future architecture design: (1) *Thermal management.* The move from 2D to 3D design could accentuate the thermal concerns due to the increased power density. To mitigate the thermal impact, thermal-aware design techniques must be adopted for 3D architecture design [56]; (2) *Design Tools and methodologies.* 3D integration technology will not be commercially viable without the support of EDA tools and methodologies that allow architects and circuit designers to develop new architectures or circuits using this technology. To efficiently exploit the benefits of 3D technologies, design tools and methodologies to support 3D designs are imperative [57]; (3) *Testing.* One of the barriers to 3D technology adoption is insufficient understanding of 3D testing issues and the lack of design-for-testability (DFT) techniques for 3D ICs, which have remained largely unexplored in the research community.

References

1. Alam S, Jones R, Rauf S, Chatterjee R. Inter-strata connection characteristics and signal transmission in three-dimensional (3D) integration technology. In: International symposium on quality electronic design, 2007.
2. Athikulwongse K, Chakraborty A, Yang JS, Pan D, Lim SK. Stress-driven 3D-IC placement with TSV keep-out zone and regularity study. In: International conference on computer-aided design, 2010.
3. Barnat S, Fremont H, Gracia A, Cadalen E, Bunel C, Neuilly F, Tenailleau J. Design for reliability: Thermo-mechanical analyses of stress in through silicon via. In: International conference on thermal, mechanical multi-physics simulation, and experiments in microelectronics and microsystems, 2010.
4. Black B, et al. Die stacking 3D microarchitecture. In: MICRO, 2006. pp. 469–79.
5. Carloni L, Pande P, Xie Y. Networks-on-chip in emerging intercoonect paradigms: advantages and challenges. In: Intl. symp. on networks-on-chips, 2009.
6. Chen Y, Kursun E, Motschman D, Johnson C, Xie Y. Analysis and mitigation of lateral thermal blockage effect of through-silicon-via in 3D IC designs. In: International symposium on low power electronics and design, 2011.
7. Chien HC, Lau JH, Chao YL, Tain RM, Dai MJ, Lo WC, Kao MJ. Estimation for equivalent thermal conductivity of silicon-through vias TSV used for 3D IC integration. In: International microsystems, packaging, assembly and circuit technology conference, 2011.
8. Cho J, Song E, Yoon K, Pak JS, Kim J, Lee W, Song T, Kim K, Lee J, Lee H, Park K, Yang S, Suh M, Byun K, Kim J. Modeling and analysis of through-silicon via (TSV) noise coupling and suppression using a guard ring. IEEE Trans Compon Packag Manuf Technol. 2011;1:220–33.
9. Cong J, Luo G, Wei J, Zhang Y. Thermal-aware 3D IC placement via transformation. In: Asia and South Pacific design automation conference, 2007.
10. Cong J, Luo G, Shi Y. Thermal-aware cell and through-silicon-via co-placement for 3D ICs. In: Design automation conference, 2011.
11. Davis WR, Wilson J, Mick S, Xu J, Hua H, Mineo C, Sule AM, Steer M, Franzon PD. Demystifying 3D ICs: the pros and cons of going vertical. IEEE Des Test Comput. 2005;22(6):498–510.

12. Dong X, Xie Y. Cost analysis and system-level design exploration for 3D ICs. In: Asia and South Pacific design automation conference, 2009.
13. Dong X, Wu X, Sun G, Xie Y, Li H, Chen Y. Circuit and microarchitecture evaluation of 3D stacking Magnetic RAM (MRAM) as a universal memory replacement. In: Design automation conference, 2009. pp. 554–9.
14. Dongkook P, Eachempati S, Das R, Mishra AK, Xie Y, Vijaykrishnan N, Das CR. MIRA: a multi-layered on-chip interconnect router architecture. In: International symposium on computer architecture, 2008. pp. 251–61
15. Egawa R, Tada J, Kobayashi H, Goto G. Evaluation of fine grain 3D integrated arithmetic units. In: IEEE international 3D system integration conference, 2009.
16. Garrou P. Handbook of 3D integration: technology and applications using 3D integrated circuits. Wiley-CVH, chap Introduction to 3D integration, 2008.
17. Goplen B, Sapatnekar S. Thermal via placement in 3D ICs. In: International symposium on physical design, 2005.
18. Jacob P, et al. Mitigating memory wall effects in high clock rate and multi-core CMOS 3D ICs: processor memory stacks. Proc IEEE 2008;96(10):5.
19. Jain A, Jones R, Chatterjee R, Pozder S. Analytical and numerical modeling of the thermal performance of three-dimensional integrated circuits. IEEE Trans Compon Packag Technol. 2010;33(1):56–63.
20. Joyner J, Zarkesh-Ha P, Meindl J. A stochastic global net-length distribution for a three-dimensional system-on-a-chip (3D-SoC). In: International ASIC/SOC conference, 2001.
21. Jung M, Mitra J, Pan D, Lim SK. TSV stress-aware full-chip mechanical reliability analysis and optimization for 3D IC. In: Design automation conference, 2011
22. Katti G, Mercha A, Stucchi M, Tokei Z, Velenis D, Van Olmen J, Huyghebaert C, Jourdain A, Rakowski M, Debusschere I, Soussan P, Oprins H, Dehaene W, De Meyer K, Travaly Y, Beyne E, Biesemans S, Swinnen B. Temperature dependent electrical characteristics of through-si-via (TSV) interconnections. In: International interconnect technology conference, 2010.
23. Katti G, Stucchi M, De Meyer K, Dehaene W. Electrical modeling and characterization of through silicon via for three-dimensional ICs. IEEE Trans Electron Devices. 2010;57:256–62.
24. Katti G, Stucchi M, Velenis D, Soree B, De Meyer K, Dehaene W. Temperature-dependent modeling and characterization of through-silicon via capacitance. IEEE Electron Device Lett. 2011;32:563–5.
25. Kgil T, D'Souza S, Saidi A, Binkert N, Dreslinski R, Mudge T, Reinhardt S, Flautner K. PicoServer: using 3D stacking technology to enable a compact energy efficient chip multiprocessor. In: ASPLOS, 2006. pp. 117–28.
26. Khalil D, Ismail Y, Khellah M, Karnik T, De V. Analytical model for the propagation delay of through silicon vias. In: International symposium on quality electronic design, 2008.
27. Kim J, Nicopoulos C, Park D, Das R, Xie Y, Vijaykrishnan N, Das C. A novel dimensionally-decomposed router for on-chip communication in 3D architectures. In: International symposium on computer architecture, 2007.
28. Li F, Nicopoulos C, Richardson T, Xie Y, Vijaykrishnan N, Kandemir M. Design and management of 3D chip multiprocessors using network-in-memory. In: International symposium on computer architecture, 2006.
29. Loh G. Extending the effectiveness of 3D-stacked DRAM caches with an adaptive multi-queue policy. In: International symposium on microarchitecture, 2009.
30. Loh G, Xie Y, Black B. Processor design in three-dimensional die-stacking technologies. IEEE Micro. 2007;27(3):31–48.
31. Loh GH. 3D-stacked memory architectures for multi-core processors. In: International symposium on computer architecture, 2008.
32. Lu KH, Zhang X, Ryu SK, Im J, Huang R, Ho P. Thermo-mechanical reliability of 3-D ICs containing through silicon vias. In: Electronic components and technology conference, 2009.
33. Lu KH, Ryu SK, Zhao Q, Zhang X, Im J, Huang R, Ho PS. Thermal stress induced delamination of through silicon vias in 3D interconnects. In: Electronic components and technology conference, 2010.

34. Lu KH, Ryu SK, Im J, Huang R, Ho P. Thermomechanical reliability of through-silicon vias in 3D interconnects. In: International reliability physics symposium, 2011.
35. Majeed B, Sabuncuoglu Tezcan D, Vandevelde B, Duval F, Soussan P, Beyne E. Electrical characterization, modeling and reliability analysis of a via last TSV. In: Electronics packaging technology conference, 2010.
36. Noritake C, Limaye P, Gonzalez M, Vandevelde B. Thermal cycle reliability of 3D chip stacked package using PB-free solder bumps: Parameter study by FEM analysis. In: International conference on thermal, mechanical and multi-physics simulation and experiments in microelectronics and microsystems, 2006.
37. Ouyang J, Sun G, Chen Y, Duan L, Zhang T, Xie Y, Irwin M. Arithmetic unit design using 180 nm TSV-based 3D stacking technology. In: international 3D system integration conference, 2009.
38. Pak JS, Ryu C, Kim J. Electrical characterization of through silicon via (TSV) depending on structural and material parameters based on 3D full wave simulation. In: International conference on electronic materials and packaging, 2007.
39. Puttaswamy K, Loh GH. Scalability of 3D-integrated arithmetic units in high-performance microprocessors. In: Design automation conference, 2007.
40. Roullard J, Capraro S, Farcy A, Lacrevaz T, Bermond C, Leduc P, Charbonnier J, Ferrandon C, Fuchs C, Flechet B. Electrical characterization and impact on signal integrity of new basic interconnection elements inside 3D integrated circuits. In: Electronic components and technology conference, 2011.
41. Ryu C, Chung D, Lee J, Lee K, Oh T, Kim J. High frequency electrical circuit model of chip-to-chip vertical via interconnection for 3-D chip stacking package. In: Topical meeting on electrical performance of electronic packaging, 2005.
42. Ryu SK, Lu KH, Zhang X, Im JH, Ho P, Huang R. Impact of near-surface thermal stresses on interfacial reliability of through-silicon vias for 3-D interconnects. IEEE Trans Device Mater Reliab. 2011;11:35–43.
43. Salah K, El Rouby A, Ragai H, Amin K, Ismail Y. Compact lumped element model for TSV in 3D-ICs. In: International symposium on circuits and systems, 2011.
44. Savidis I, Friedman E. Closed-form expressions of 3-D via resistance, inductance, and capacitance. IEEE Trans Electron Devices. 2009;56:1873–81.
45. Selvanayagam C, Lau J, Zhang X, Seah S, Vaidyanathan K, Chai T. Nonlinear thermal stress/strain analyses of copper filled TSV (through silicon via) and their flip-chip microbumps. IEEE Trans Adv Packag. 2009;32(4):720–8.
46. Sun G, Dong X, Xie Y, Li J, Chen Y. A novel 3D stacked MRAM cache architecture for CMPs. In: International symposium on high performance computer architecture, 2009.
47. Tsai YF, Wang F, Xie Y, Vijaykrishnan N, Irwin MJ. Design space exploration for three-dimensional cache. IEEE Trans Very Large Scale Integr VLSI Syst. 2008;16(4):444–55.
48. Vaidyanathan B, Hung WL, Wang F, Xie Y, Narayanan V, Irwin MJ. Architecting microprocessor components in 3D design space. In: Intl. conf. on VLSI design, 2007.
49. Vangal S, et al. An 80-tile Sub-100-W TeraFLOPS processor in 65-nm CMOS. IEEE J Solid State Circuits. 2008;43(1):29–41.
50. Vantrease D, Schreiber R, Monchiero M, McLaren M, Jouppi NP, Fiorentino M, Davis A, Binkert N, Beausoleil RG, Ahn JH. Corona: system implications of emerging nanophotonic technology. In: international symposium on computer architecture, 2008.
51. Wang F, Zhu Z, Yang Y, Wang N. A thermal model for the top layer of 3D integrated circuits considering through silicon vias. In: International conference on ASIC, 2011.
52. Weerasekera R, Grange M, Pamunuwa D, Tenhunen H, Zheng LR. Compact modelling of through-silicon vias (TSVs) in three-dimensional (3-D) integrated circuits. In: International conference on 3D system integration, 2009.
53. Wu X, Li J, Zhang L, Speight E, Xie Y. Hybrid cache architecture. In: International symposium on computer architecture, 2009.

54. Wu X, Zhao W, Nakamoto M, Nimmagadda C, Lisk D, Gu S, Radojcic R, Nowak M, Xie Y. Electrical characterization for intertier connections and timing analysis for 3-D ICs. IEEE Trans Very Large Scale Integr VLSI Syst. 2012;20:186–91.
55. Xie J, Chung D, Swaminathan M, Mcallister M, Deutsch A, Jiang L, Rubin B. Electrical-thermal co-analysis for power delivery networks in 3D system integration. In: International conference on 3D system integration, 2009.
56. Xie Y, Loh G, Black B, Bernstein K. Design space exploration for 3D architectures. ACM J Emerg Technol Comput Syst. 2006;2:65–103.
57. Xie Y, Cong J, Sapatnekar S. Three-dimensional integrated circuit design: EDA, design and microarchitectures. Springer: New York, 2009.
58. Xu C, Li H, Suaya R, Banerjee K. Compact ac modeling and analysis of Cu, W, and CNT based through-silicon vias (TSVs) in 3-D ICs. In: International electron devices meeting, 2009.
59. Xu C, Li H, Suaya R, Banerjee K. Compact ac modeling and performance analysis of through-silicon vias in 3-D ICs. IEEE Trans Electron Devices. 2010;57:3405–17.
60. Xu Y, et al. A low-radix and low-diameter 3D interconnection network design. In: Intl. symp. on high performance computer architecture, 2009.
61. Zou Q, Zhang T, Kursun E, Xie Y. Thermomechanical stress-aware management for 3D IC designs. In: Design, automation test in Europe conference exhibition, 2013.

Chapter 3
Design and Optimization of Spin-Transfer Torque MRAMs

Xuanyao Fong, Sri Harsha Choday, and Kaushik Roy

Abstract In this chapter, reviews the basics and modeling of spin-transfer torque magnetic RAM (STT-MRAM) for circuit-level failure analysis. A methodology for analyzing failures in STT-MRAM bit-cells is also presented. The optimization of STT-MRAM bit-cells using the presented framework is then discussed, along with several circuit and array architecture-level failure mitigation techniques. We will show that despite the relatively high write energy in STT-MRAM, large capacity last level caches based on STT-MRAM can be more energy efficient than their SRAM counterparts due to the unique characteristics of STT-MRAM.

The cache capacity of high-performance microprocessors is increasing as transistor technology is scaled down. Since the leakage power also increases exponentially with the scaling down of transistor technology, the power dissipation of on-chip caches is an increasingly dominant component of power dissipation in high-performance microprocessors. Non-volatile memories have been proposed as a solution for mitigating the increasing power dissipation in high-performance on-chip caches. Among the currently available non-volatile memory technologies, only spin-transfer torque magnetic random access memory (STT-MRAM) has the desired characteristics for high-performance on-chip cache applications [1]. In this chapter, we discuss the design optimization and modeling of STT-MRAMs, and its potential application in high-performance on-chip caches.

3.1 MRAM Storage Device: The Magnetic Tunnel Junction

The storage device in MRAM is the magnetic tunnel junction or MTJ. An MTJ, as shown in Fig. 3.1, consists of a soft ferromagnetic layer which stores the information (also called the "free" layer), a tunneling layer (usually AlO_x or more commonly,

X. Fong • S.H. Choday • K. Roy (✉)
School of Electrical and Computer Engineering, Purdue University,
West Lafayette, IN 47907, USA
e-mail: xfong@purdue.edu; schoday@purdue.edu; kaushik@purdue.edu

R.O. Topaloglu (ed.), *More than Moore Technologies for Next Generation Computer Design*, DOI 10.1007/978-1-4939-2163-8_3

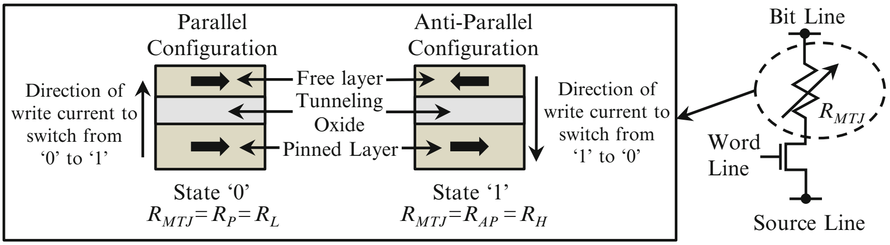


Fig. 3.1 The storage device in the MRAM memory cell is the magnetic tunnel junction (illustrated inset). The memory cell consists of an access transistor and the storage device connected as shown. The current direction for programming the cell using spin-transfer torque is also shown

MgO), and a reference ferromagnetic layer (also called the "fixed" or "pinned" layer). The MTJ can be switched between two stable states. When both the free and the pinned layers are magnetically aligned, the configuration is called the "parallel" state (P), and when the free and the pinned layers are anti-aligned magnetically, the configuration is called the "anti-parallel" state (AP). A metric for MTJ as shown in [2] is its resistance–area (*RA*) product. The *RA* product of the MTJ depends exponentially on the tunnel oxide thickness (t_{MgO}) since the mechanism for electron transport is tunneling. At the same t_{MgO}, the MTJ resistance, R_{MTJ}, depends linearly on the cross-sectional area of the MTJ (A_{MTJ}), similar to an Ohmic conductor. R_{MTJ} also depends on the relative magnetic polarization of the free layer with respect to the pinned layer. The dependence of R_{MTJ} on magnetic polarization is due to the difference in density of states around the Fermi energy, E_F, in the ferromagnetic layers [3]. When the MTJ is in the P state, the density of states of like-spins around E_F is very high in the ferromagnetic layers. Conversely, the density of states of like-spins around E_F in the ferromagnetic layers is very low when the MTJ is in AP state. Thus, R_{MTJ} is low in the P state ($R_{MTJ} = R_P = R_L$) and high in the AP state ($R_{MTJ} = R_{AP} = R_L$). This difference in R_{MTJ}, termed the "tunneling magneto-resistance ratio" (or *TMR*), is given by

$$TMR = \frac{R_{AP} - R_P}{R_P} \times 100\,\% \tag{3.1}$$

and is an important metric for the performance of MTJs as memory elements. Since binary data are represented by and stored as the resistance state of the MTJ, a larger *TMR* also means that the MTJ states can be distinguished more easily. A constant voltage or constant current scheme can be used to sense R_{MTJ} and hence, the MTJ state [4, 5]. In the constant voltage scheme, a fixed voltage is applied across the MTJ and the resulting current through the MTJ is compared to a reference current. The current flowing through the MTJ can be either higher or lower than the reference current, depending on the resistance state of the MTJ. The advantage of the constant voltage scheme is that the current flowing through the MTJ during read operations may be amplified in the sense amplifier to improve sensing speed. However, the disadvantage is that the result of the sensing needs to be converted into an output voltage. In case of constant current scheme, a fixed current is passed through the MTJ and the voltage developed across the bit-line and the source-line is compared with a reference voltage. The constant current scheme has the advantage that the result of the sensing is already in the voltage domain and hence, no conversion is required. However, the current required to generate sufficient voltage signal for sensing may be large enough to cause *disturb failures*, which will be discussed in detail later.

The magnetic layers are stabilized against thermal effects by engineering them with anisotropies during fabrication. The most common form of anisotropy engineered into the magnetic layers of an MTJ is the *uniaxial anisotropy*. This causes the magnetization of the magnetic layers to have a preferential alignment axis—the magnetization will align along this axis when no external stimulus is present. When the volume of the magnet is reduced, the *uniaxial anisotropy energy* must be proportionally increased to maintain the same stability. We will discuss this in more detail in the later sections.

Nano-scale MTJs may be switched using the spin-transfer torque phenomenon which was theoretically predicted by Slonczewski and Berger independently in 1996 [6, 7]. Since then many experiments have observed spin-transfer torque (STT) switching [8–10]. STT exists because magnetism in ferromagnetic metals arises due to the spin property of electrons. The magnetization of the ferromagnet points in a particular direction when the majority of electron spins in it are aligned in that direction. Hence, when current flows through the MTJ, the ferromagnetic layers act as spin filters that polarizes the flowing electrons. Electrons in a spin polarized current flowing into a ferromagnetic layer are able to transfer their spin momentum to it. The spin momentum transferred exerts a torque on the magnetization of the ferromagnetic layer. The magnetization of the ferromagnetic layer is switched if the torque is large enough to overcome all other energies in the ferromagnetic layer. The rate of spin momentum transfer and the torque exerted are proportional to the rate of electron flow or the current, and determine the switching time. The current or current density needed to achieve a specific switching time is the *critical current*, I_C, or *critical current density*, J_C.

In an MTJ, the pinned layer is magnetically pinned whereas the free layer is not. Hence, it is easier for spin-transfer torque to switch the free layer than to switch the pinned layer. Let us consider what happens when electrons are flowing from the pinned layer to the free layer in an MTJ. The pinned layer polarizes the incoming electrons which then flow into the free layer. These electrons are polarized in the spin direction of the pinned layer and transfer their spin momentum to the free layer. Hence, a spin-transfer torque is exerted on the free layer to align its magnetization parallel with the pinned layer. Consider instead when electrons flow from the free layer to the pinned layer. Electrons entering the free layer from the metallic interconnect are not polarized and can have any spin direction. Electrons with same spin direction as the pinned layer are able to tunnel across the oxide easily. However, electrons with the opposite spin-polarization may not tunnel across the oxide easily and accumulate in the free layer. These electrons transfer their spin angular momentum to the free layer and exert a torque that aligns the free layer magnetization anti-parallel with the pinned layer. When the electrons transfer their spin angular momentum to the free layer, their spin directions become aligned with the spin polarization of the pinned layer. They may then tunnel across the oxide easily. From this discussion, we can see that the process of parallelizing the free and pinned layers is more efficient than the anti-parallelizing process, resulting in asymmetry in I_C and J_C [3, 11]. It has been reported that J_C when anti-parallelizing the MTJ can be 10–200 % larger than for parallelizing the MTJ [11, 12].

3.2 Modeling Magnetic Tunnel Junctions

The transient behavior of an MTJ can be modeled only if the essential physics in it are captured. The *I–V* characteristic of the MTJ depends on physical parameters of the MTJ, such as the thickness of the tunneling oxide and the cross-sectional area of the MTJ, and on the magnetization directions of the free and the pinned layers. Since the magnetization of the free layer does not change instantaneously during switching, R_{MTJ} also transition smoothly during MTJ switching. Accurate modeling of the transient behavior of MTJs must model the transient behavior of the free layer and relate it to the *I–V* characteristic of the MTJ. The transient behavior of the free layer magnetization may be modeled using the *Landau–Lifshitz–Gilbert* (LLG) [13] equation, and the *I–V* characteristic of the MTJ may be modeled using the *Non-Equilibrium Green's Function* (NEGF) [3] approach.

3.2.1 The Non-Equilibrium Green's Function (NEGF) Approach

The Non-Equilibrium Green's Function (NEGF) approach may be used to simulate electronic transport through an MTJ [3]. The approach requires the effective mass Hamiltonian representing the MTJ and the MTJ biasing conditions, to be written first. The *I–V* characteristics may be calculated by solving Non-Equilibrium Green's Function (NEGF) equations. Details of the approach are published in [3, 14] and are beyond the scope of this chapter.

The NEGF approach to modeling the *I–V* characteristic of the MTJ has the advantage that model parameters correspond to material parameters and may be obtained from experimental measurements. The model may then be used to predict MTJ characteristics and then validated experimentally. Figure 3.2 shows the successful calibration of the NEGF model to experimentally measured data published in [15, 16].

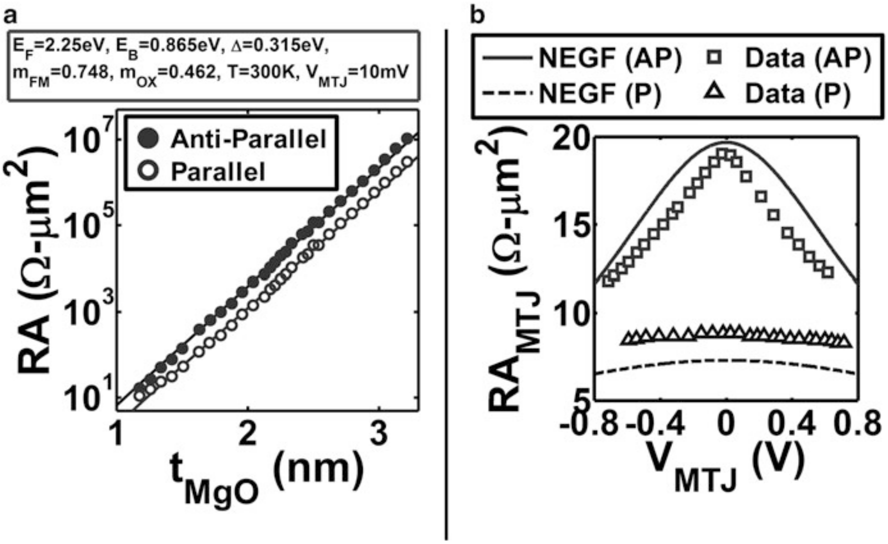


Fig. 3.2 The successful calibration of the NEGF model to experimentally measured data reported in the literature. (**a**) and (**b**) show results for calibration to data from [15] and from [16], respectively

3.2.2 The Landau–Lifshitz–Gilbert (LLG) Equation

The typical approach to simulating the magnetization dynamics in an MTJ is the micromagnetic approach. In this approach, the free magnetic layer in the MTJ is discretized into a 3-D grid of ferromagnetic mono-domains. Since micromagnetic simulations solve the Landau–Lifshitz–Gilbert (LLG) equation numerically, they need to be repeated such that the solutions converge when parameters that should not affect them are varied. For example, the magnetization dynamics are independent of the discretization resolution and the discretization resolution is increased until the numerical solutions of micromagnetic simulations converge.

The LLG equation describes magnetization dynamics of each ferromagnetic mono-domain, and is given by [13]

$$\frac{\partial \widehat{m}}{\partial t} = -|\gamma|\,\widehat{m} \times \overrightarrow{H}_{EFF} + \alpha \widehat{m} \times \frac{\partial \widehat{m}}{\partial t} \tag{3.2}$$

where $\widehat{m}$ is the unit vector describing the magnetization direction of the mono-domain, γ is the electron gyromagnetic ratio (17.5 MHz/Oe or 2.21×10^5 m/A s), and α is the Gilbert damping factor [13]. An effective magnetic field, $\overrightarrow{H}_{EFF}$, models the forces acting on the mono-domain. In an MTJ, $\overrightarrow{H}_{EFF}$ may be written as

$$\overrightarrow{H}_{EFF} = \overrightarrow{H}_{Ani} + \overrightarrow{H}_{Dip} + \overrightarrow{H}_{Demag} + \overrightarrow{H}_{Ex} + \overrightarrow{H}_{Ext} + \overrightarrow{H}_{TH} + \overrightarrow{H}_{STT} \tag{3.3}$$

$\overrightarrow{H}_{Ani}$, $\overrightarrow{H}_{Dip}$, $\overrightarrow{H}_{Demag}$, $\overrightarrow{H}_{Ex}$, $\overrightarrow{H}_{Ext}$, $\overrightarrow{H}_{TH}$, and $\overrightarrow{H}_{STT}$ describe the effective magnetic fields due to magnetic anisotropies (including uniaxial anisotropy), dipolar coupling of the mono-domain to other magnetic dipoles, the demagnetization field due to the arrangement of the magnetic ensemble, the exchange coupling between mono-domains, any externally applied magnetic field, effects due to temperature, and spin-transfer torque, respectively. The first term in the right-hand side of Eq. (3.2) describes the precession of the magnetization around the axis of the effective magnetic field. On the other hand, the remaining term in the right-hand side of Eq. (3.2) describes the dampening of the precession which forces the magnetization to align with the effective magnetic field.

The free layer in the MTJ is stabilized against thermal effects using shape anisotropy, crystalline anisotropy, etc. Uniaxial anisotropy result in the free layer magnetization to preferentially align itself along a single axis, $\hat{u}$, and the effective anisotropy field may be calculated using

$$\overrightarrow{H}_{Ani} = 2K_{u2} \left(\widehat{m} \cdot \widehat{u}\right) \ \widehat{u} \tag{3.4}$$

where K_{u2} is the second order uniaxial anisotropy constant.

When an ensemble of mono-domains is considered, the demagnetization field due to the geometry of the ensemble needs to be considered. Since $\overrightarrow{\nabla} \times \overrightarrow{H}_{Demag} = 0$ and $\overrightarrow{\nabla} \cdot \overrightarrow{B}_{Demag} = 0$ in a uniformly magnetized mono-domain, the demagnetization field can be written as the gradient of a scalar potential

$$\overrightarrow{H}_{Demag} = -\overrightarrow{\nabla} \Phi M \tag{3.5}$$

where

$$\Phi_M \left(\mathbf{r}\right) = \frac{1}{4\pi} \int M \left(\mathbf{r}'\right) \cdot \overrightarrow{\nabla} \left(\frac{1}{\left|\mathbf{r} - \mathbf{r}'\right|}\right) \, d^3\mathbf{r}' \tag{3.6}$$

and $M(\mathbf{r}')$ is the magnetization of the whole ensemble relative to the origin. Details of the calculation of the demagnetization field in numerical solvers are beyond the scope of this chapter and may be found in [17, 18].

Mono-domains that are far apart may appear to be magnetic dipoles to each other. The magnetic field on a mono-domain due to a magnetic dipole is given by

$$\overrightarrow{H}_{DIP} = \frac{3\left(\overrightarrow{M} \cdot \overrightarrow{r}\right) \overrightarrow{r} - \left|\overrightarrow{r}\right|^2 \overrightarrow{M}}{4\pi \left|\overrightarrow{r}\right|^5} \tag{3.7}$$

where $\overrightarrow{M}$ is the magnetic moment of the dipole (or $\overrightarrow{M} = M_S \widehat{m}$ if a mono-domain with magnetization direction $\widehat{m}$ is approximated as a point dipole) and $\overrightarrow{r}$ is the vector pointing from the magnetic dipole to the mono-domain.

Thermal energy may also perturb the spin interaction between electrons in a mono-domain and needs to be modeled as well. The formulation of the effect thermal energy has on a mono-domain was presented by Brown in [19]. This effect is captured in Eq. (3.3) using the effective thermal field $\overrightarrow{H}_{TH}$. The thermal field is related to the mono-domain properties by

$$\overrightarrow{H}_{TH} = \overrightarrow{\xi} \sqrt{\frac{2k_B T}{|\gamma| \mu_0 M_S V_{Domain} \Delta t}} \tag{3.8}$$

where $\overrightarrow{\xi}$ is a vector with components that are independent standard Gaussian random variables, k_B is the Boltzmann constant, T is the temperature of the magnetic ensemble, μ_0 is the permeability of free space, Δt is the constant time step used in the numerical simulation, M_S and V_{Domain} are the saturation magnetization and the volume of the mono-domain, respectively. The statistics of $\overrightarrow{H}_{TH}$ are such that

$$\langle H_{TH,u} \rangle = 0 \ \text{ where } \ u = x, y, z \tag{3.9}$$

$$\langle H_{TH,u}(t) H_{TH,v} (t + \tau) \rangle = \frac{2k_B T}{|\gamma| \mu_0 M_S V_{Domain}} \delta(\tau) \delta_{uv} \tag{3.10}$$

where u and v denote the component of $\overrightarrow{H}_{TH}$.

Slonczewski and Berger independently showed that when a spin-polarized electron current (spins of every electron in the current are aligned in one direction) flows into a ferromagnetic layer, the electrons transfer their spin momentum to the ferromagnetic layer, exerting a torque on the magnetization of the ferromagnetic layer [6, 7]. The spin-transfer torque effect can be written as

$$-|\gamma| \widehat{m} \times \overrightarrow{H}_{STT} = \beta (\widehat{m} \times (\widehat{m} \times \widehat{m}_P)) + \beta' \widehat{m} \times \widehat{m}_P \tag{3.11}$$

where β and β' depend on the current, and $\widehat{m}_P$ is the unit vector describing the spin direction of the electrons entering the ferromagnetic layer. In the case of spin valves and of MTJs, $\widehat{m}_P$ corresponds to the magnetization direction of the pinned ferromagnetic layer. It may be convenient to write the spin-transfer torque in Eq. (3.11) as an effective field instead, which is given by

$$\overrightarrow{H}_{STT} = \frac{\beta}{|\gamma|} (\widehat{m}_P \times \widehat{m}) - \frac{\beta'}{|\gamma|} \widehat{m}_P \tag{3.12}$$

In Eqs. (3.11) and (3.12),

$$\beta = a_J \frac{|\gamma|}{\mu_0 M_S V_{Domain}} \frac{\hbar}{2} \frac{I_{Curr}}{e} \tag{3.13}$$

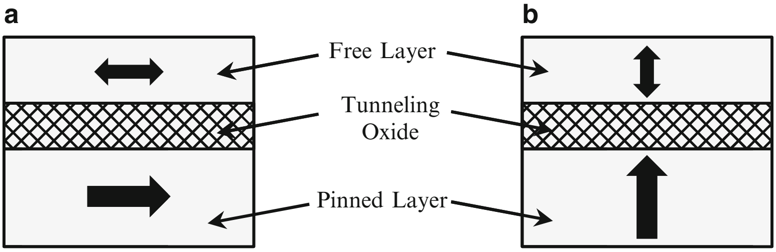


Fig. 3.3 The (**a**) in-plane magnetic anisotropy (IMA) MTJ has magnetizations which are in the plane of the thin film ferromagnetic layers whereas the (**b**) perpendicular-magnetic anisotropy (PMA) MTJ has magnetizations that are perpendicular to the plane of the thin film ferromagnetic layers

where e is the electronic charge, $\hbar$ is the reduced Planck constant, I_{Curr} is the electronic current flowing from the mono-domain into the polarizing ferromagnetic layer, and a_J is dimensionless. β' has the same form as β except a_J is replaced by $a_J^{'}$. The vector direction of the effective magnetic flux density is the spin direction, $\widehat{m}_P$, of the spin-carrying particles. a_J and $a_J^{'}$ are fitting functions that describe the in-plane and perpendicular-to-plane torques, respectively, relative to the plane containing $\widehat{m}$ and $\widehat{m}_P$. They may be interpreted as the effectiveness of spin-transfer (i.e. the proportion of total available spin-angular momentum that is transferred to the mono-domain).

MTJs with *perpendicular magnetic anisotropy* (PMA) are currently the technology of choice for STT-MRAM application. The magnetic layers in MTJs with PMA have magnetizations that are perpendicular to the plane of the magnetic layers. Previously, MTJs have *in-plane anisotropy* (IMA) in which the magnetic layers have magnetizations that are in-plane to the magnetic layers. The difference between MTJs with IMA and with PMA is illustrated in Fig. 3.3. In IMA, the STT has to overcome both $\overrightarrow{H}_{Ani}$ and $\overrightarrow{H}_{Demag}$. The strength of the effective field that STT needs to overcome is approximately $4\pi M_S$. Furthermore, it is difficult to increase the retention time as the MTJ with IMA is scaled down. These two issues are absent in MTJs with PMA. Since $\overrightarrow{H}_{Ani}$ and $\overrightarrow{H}_{Demag}$ are collinear in MTJs with PMA, the MTJ free layer can be modeled with only uniaxial anisotropy. The relationship between switching the energy barrier, E_A, and the critical switching field is then given by

$$\overrightarrow{H}_C = \frac{2E_A}{\mu_0 M_S V_{FL}} \tag{3.14}$$

where V_{FL} is the volume of the free layer. Also, $E_A = K_{u2} V_{FL}$.

In conventional MRAM, the MTJ free layer magnetization is switched using magnetic fields generated by current carrying wires as shown in Fig. 3.4. The required current for switching the MTJ is

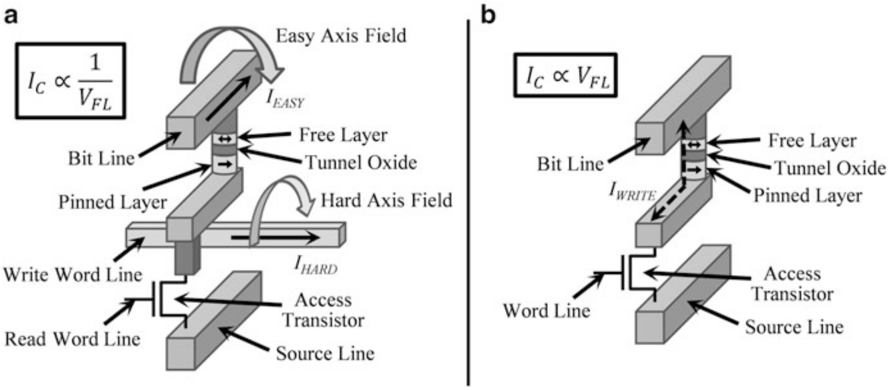


Fig. 3.4 Structures of (**a**) the field-switched MRAM, and (**b**) the spin-transfer torque MRAM

$$I_C = \frac{4\pi r E_A}{\mu_0 M_S V_{FL}} \tag{3.15}$$

where r is the spacing between the wire and the center of the free layer. When the MTJ is scaled down, I_C increases and hence MRAM is not scalable. On the other hand, if the MTJ free layer is approximated as a mono-domain, the effective switching field due to spin-transfer torque, which may be written as

$$\vec{H}_{STT} = \frac{\hbar I_{Curr}}{2e\mu_0 M_S V_{FL}} \left(a_J\left(\widehat{m}_P \times \widehat{m}\right) - a'_J \widehat{m}_P\right) \tag{3.16}$$

scales up at the same rate as $\vec{H}_C$ when the MTJ is scaled down. Hence, spin-transfer torque MRAM overcomes the scalability issue in MRAM.

3.2.3 SPICE Compatible Model of Magnetic Tunnel Junctions

The interaction between device dynamics within the MTJ and the external circuit needs to be considered in the design of STT-MRAM memory cells. Hence, a SPICE compatible model for the MTJ needs to be developed to include MTJ physics during circuit simulations in SPICE. Figure 3.5 shows how an SPICE compatible model for an MTJ with a mono-domain free layer may be implemented. This model captures the magnetization dynamics of the MTJ free layer, and the dependence of the I–V characteristics of the MTJ on the MTJ biasing conditions.

The LLG equation for the free layer may be solved by rewriting Eq. (3.2) in spherical coordinates and noting that the radial component of $\widehat{m}$ is constant. A circuit block consisting of current sources driving a capacitor may then be used to implement a differential equation solver in SPICE by noting that

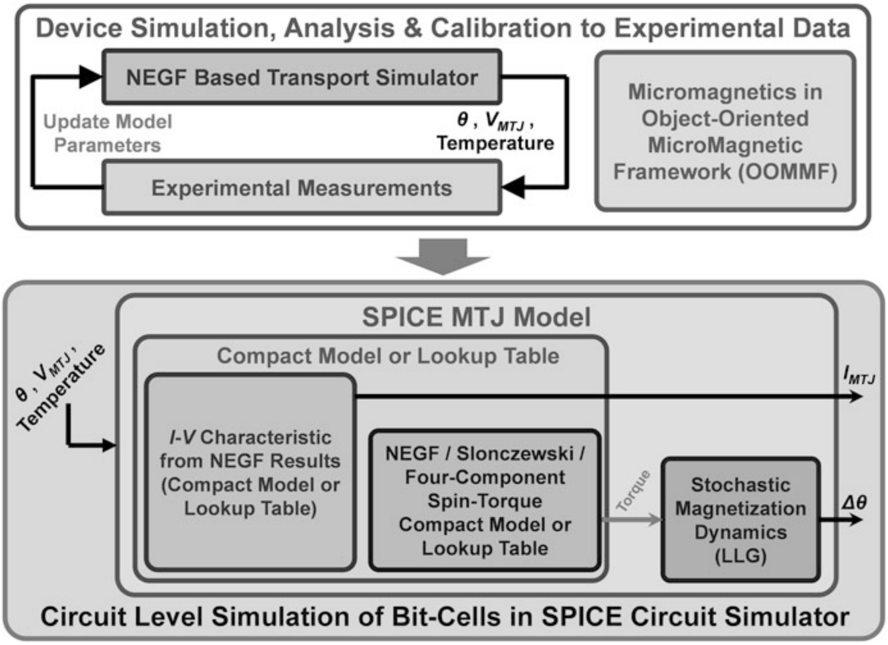

Device Simulation, Analysis & Calibration to Experimental Data
NEGF Based Transport Simulator
Update Model Parameters
θ , V_MTJ , Temperature
Experimental Measurements
Micromagnetics in Object-Oriented MicroMagnetic Framework (OOMMF)
SPICE MTJ Model
Compact Model or Lookup Table
θ , V_MTJ , Temperature
I_MTJ
I-V Characteristic from NEGF Results (Compact Model or Lookup Table)
NEGF / Slonczewski / Four-Component Spin-Torque Compact Model or Lookup Table
Torque
Stochastic Magnetization Dynamics (LLG)
Δθ
Circuit Level Simulation of Bit-Cells in SPICE Circuit Simulator

Fig. 3.5 Device/circuit simulation framework used to evaluate STT-MRAM. Device level simulation results are validated using experimental data before parameters are imported into the SPICE model for circuit level simulation of STT-MRAM bit-cells

$$\frac{dv_C}{dt} = \frac{i_C}{C} \tag{3.17}$$

where v_C and i_C are the voltage across and current through the capacitor with capacitance C, respectively. A pair of such circuit blocks can be used to solve the angular components of Eq. (3.2) by representing the right-hand side of Eq. (3.2) as a sum of currents.

The I–V characteristics of the MTJ may be stored in a lookup table by noting that the current flowing through the MTJ depends on both the voltage across the MTJ, as well as the magnetizations $\widehat{m}$ and $\widehat{m}_P$ of the free and pinned layers, respectively. Such a lookup table may consume a lot of memory and is impractical to implement. An alternate method is to note that the dependence of MTJ current on MTJ voltage and on the magnetizations may be decoupled by

$$I_{MTJ}\left(V_{MTJ}\right) = I_{AP}\left(V_{MTJ}\right)\ sin^2\left(\frac{\theta}{2}\right) + I_P\left(V_{MTJ}\right)\ cos^2\left(\frac{\theta}{2}\right) \tag{3.18}$$

where $\widehat{m} \cdot \widehat{m}_P = cos\ \theta$, and $I_{AP}(V_{MTJ})$ and $I_P(V_{MTJ})$ are the MTJ currents in the antiparallel and parallel configurations, respectively, when the voltage applied across the MTJ is V_{MTJ}. Hence, the I–V characteristics of the MTJ may be implemented using lookup tables or equations for I_{AP} and for I_P. The lookup tables or equations need to capture the dependence of I_{AP} and I_P on V_{MTJ}, MTJ cross-sectional area, and MTJ tunneling oxide thickness also [20].

3.3 Design of STT-MRAM Memory Cells

The STT-MRAM memory cell may be thought of as a programmable resistor connected with an access transistor as shown in Fig. 3.1. In an on-chip cache array, the gates of the access transistors in each row of memory cells are connected together so that they may be accessed in parallel. The bit and source lines are shared along the column of the array so that individual memory cells along the row being accessed may be written to or read from in parallel. When a memory cell is being accessed, the word line connected to the cell is charged to the supply voltage, V_{DD}, to enable the access transistor. Write operations are performed by charging the bit line and source line to the required voltages so that current will flow through the MTJ to program it. The directionality of the current determines the data being stored in the memory cell. Read operations may be performed either by passing a fixed current through the cell and sensing the voltage developed across the bit and source lines (also called *voltage sensing scheme*), or by clamping the voltages of the bit and source lines and sensing the current flowing through the memory cell (also called *current sensing scheme*). Figure 3.6 shows the biasing conditions of the STT-MRAM memory cell for different operations.

Under process variations, failures may occur during the operation of STT-MRAM memory cells. Variations in MTJ tunnel oxide thickness, t_{MgO}, and MTJ cross-sectional area affect R_{MTJ}, which in turn affect the ability to write into the memory cell, the ability to correctly sense R_{MTJ} of the memory cell, and the ability of the MTJ to retain its configuration when the bit-cell is being read. *Write failures* occur when the MTJ cannot be switched between anti-parallel and parallel

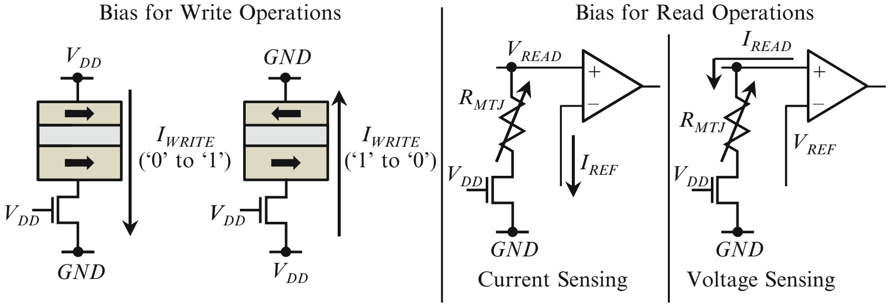


Fig. 3.6 Biasing conditions for read and for write operations of STT-MRAM. In current sensing read operation, the bit-line is clamped at V_{READ}, and the bit-cell current is compared to the reference current, I_{REF}. In voltage sensing, a read current (I_{READ}) is passed through the bit-cell and the voltage on the bit-line is compared to the reference voltage, V_{REF}

configurations. This occurs when the current through the MTJ falls below I_C during write. Read failures occur when R_{MTJ} is incorrectly determined (*decision* failure) or when the MTJ configuration is accidentally switched during read (*disturb* failure). The failure probability of each type of STT-MRAM failure may be calculated using D.C. load line analyses, discussed in the following sections.

3.3.1 Modeling STT-MRAM Failures

The common approach to calculating STT-MRAM failure probabilities assumes distributions for R_{MTJ} and the *TMR* of the MTJ [21], which may be physically incorrect. We now show how STT-MRAM failure probabilities may be calculated without the need to assume distributions for R_{MTJ} and *TMR* of the MTJ.

Write failure occurs when data cannot be written into a STT-MRAM bit-cell within the write cycle. This occurs when R_{MTJ} is too large for the access transistor to provide the required I_C. Write failure may occur when t_{MgO} is too thick, when the access transistor has a threshold voltage (V_T) that is too high, when access transistor width is too small, or when other factors or a combination of factors that results in a write current smaller than I_C flowing through the MTJ occur. The write failure probability ($P_{WR,i}$) for a particular bit-cell may be calculated using D.C. load line analysis as shown in Fig. 3.7a. Consider a bit-cell having an MTJ with cross-sectional area $A_{MTJ,j}$, and I_{MTJ} is exactly I_C for parallel-to-anti-parallel (P-to-AP) switching corresponding to $A_{MTJ,j}$. Further, consider that the MTJ is in parallel (P) configuration with $t_{MgO} = t_{WR,MAX}$ and resistance R_P. I_{MTJ} falls below I_C if $t_{MgO} > t_{WR,MAX}$, and hence data cannot be written into this bit-cell within one write cycle. The same argument holds for an MTJ in AP configuration. Since $t_{WR,MAX}$ depends on $A_{MTJ,j}$, $P_{WR,i}$ for this particular bit-cell can be written as

$$P_{WR,i} = \lim_{\delta \to 0} \sum_{all \quad j} P\left(X - \delta \le X \le X + \delta\right) \cdot P\left(t_{MgO} \ge t_{WR-MAX,j}\right) \qquad (3.19)$$

where $X = A_{MTJ,j}$. Since $t_{WR,MAX,j}$ depends on A_{MTJ} and A_{MTJ} is allowed to vary, A_{MTJ} is divided into bins (indexed as j) for numerical calculation of $P_{WR,i}$. The write failure probability of the array (P_{WR}) may be calculated by first using Monte Carlo simulation to generate N access transistor I–V characteristic and calculating $P_{WR,i}$ for each I–V characteristic. P_{WR} may then be calculated as

$$P_{WR} = \sum_{i=1}^{N} P_{WR,i} \qquad (3.20)$$

The *disturb failure* probability for a STT-MRAM cell ($P_{RD,i}$) may also be calculated in a similar way by noting that disturb failure occurs when data is accidentally written into the cell during read operations. The D.C. load line used

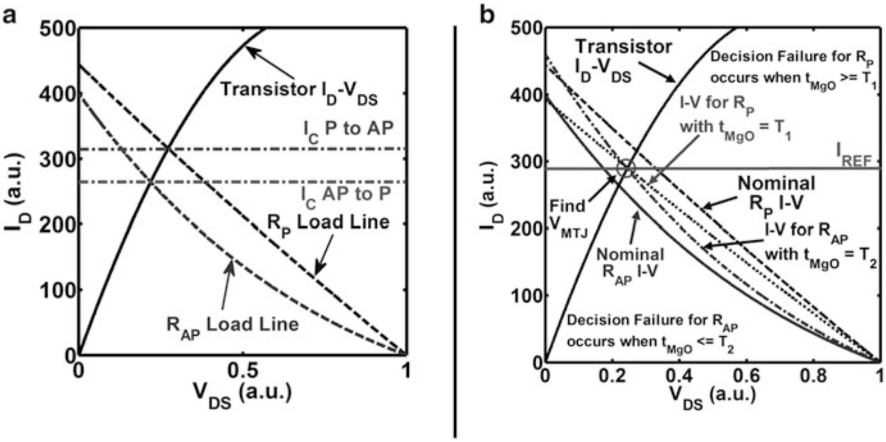


Fig. 3.7 Load lines used for analyzing (**a**) write and disturb failures, and (**b**) decision failures

for calculating $P_{RD,i}$ is the same as that in Fig. 3.7a except that the I–V curve of the MTJ intersects the horizontal axis at $V = V_{READ}$. Consider a memory cell having an MTJ with cross-sectional area $A_{MTJ,j}$, and I_{MTJ} is exactly I_C for P-to-AP switching corresponding to $A_{MTJ,j}$. Further, consider that the MTJ in the bit-cell is in P configuration with $t_{MgO} = t_{RD-MIN}$ and resistance R_P. If $t_{MgO} < t_{RD-MIN}$, I_{MTJ} will rise above I_C, and hence the data gets written into this memory cell within one read cycle, causing a disturb failure. The same argument holds for an MTJ in the AP configuration. However, the read operation involves only one direction of current flow, and for a specific direction of read current flow, either P-to-AP disturbs or AP-to-P disturbs will occur but not both. t_{RD-MIN} depends on $A_{MTJ,j}$, and $P_{RD,i}$ for this particular bit-cell is

$$P_{RD,i} = \lim_{\delta \to 0} \sum_{all\ \ j} P\left(X - \delta \le X \le X + \delta\right) \cdot P\left(t_{MgO} \le t_{RD-MIN,j}\right) \tag{3.21}$$

where $X = A_{MTJ,j}$. Since $t_{RD,MIN,j}$ depends on A_{MTJ} and A_{MTJ} is allowed to vary, A_{MTJ} is divided into bins (indexed as j) for numerical calculation of $P_{RD,i}$. The disturb failure probability of the array (P_{RD}) may be calculated by first using Monte Carlo simulation to generate N access transistor I–V characteristic and calculating $P_{RD,i}$ for each I–V characteristic. P_{RD} may then be calculated as

$$P_{RD} = \sum_{i=1}^{N} P_{RD,i} \tag{3.22}$$

The calculation for the decision failure probability of a STT-MRAM memory cell (P_{DEC}) depends on the sensing scheme and sense amplifier used. Consider the current sensing scheme where during STT-MRAM read operation, the voltage across the bit line and the source line is clamped at V_{READ} and a current sense amplifier (SA) compares the current flowing through the memory cell (I_{Cell}) with a reference current, I_{REF}. If $I_{Cell} < I_{REF}$, the MTJ in the memory cell is in the anti-parallel configuration (AP) or $R_{MTJ} = R_{AP}$ and the SA outputs logic '1'. If $I_{Cell} > I_{REF}$, the MTJ in the memory cell is in the parallel configuration (P) or $R_{MTJ} = R_P$ and the amplifier outputs logic '0'. However, due to process variations, I_{Cell} may be higher than I_{REF} when the MTJ is in AP, or lower than I_{REF} when the MTJ is in P. When this occurs, the SA outputs logic '0' when $R_{MTJ} = R_{AP}$ or logic '1' when $R_{MTJ} = R_P$. Such a failure is called a *decision* failure. I_{REF} needs to be carefully chosen to minimize decision failures.

Figure 3.7b illustrates the D.C. load lines used to calculate the decision probability for a particular memory cell ($P_{DEC,i}$) with a particular I_{REF}. For an MTJ in AP at the nominal t_{MgO} and cross-sectional area $A_{MTJ,j}$, its resistance is R_{AP} and the load line is the solid red line. $I_{MTJ} = I_{REF}$ when $t_{MgO} = T_2$. If $t_{MgO} < T_2$, I_{MTJ} will be more than I_{REF} and the SA incorrectly outputs logic '0'. Similarly, $I_{MTJ} = I_{REF}$ if the MTJ is in P and has cross-sectional area $A_{MTJ,j}$, and $t_{MgO} = T_1$. If $t_{MgO} > T_1$, I_{MTJ} will be less than I_{REF} and the SA incorrectly outputs logic '1'. Thus, for this particular STT-MRAM memory cell

$$P_{DEC,i} = \lim_{\delta \to 0} \sum_{all\ \ j} P\left(X - \delta \le X \le X + \delta\right) \cdot P\left(T_1 \le t_{MgO} \le T_2\right) \tag{3.23}$$

where $X = A_{MTJ,j}$. Since T_1 and T_2 depend on A_{MTJ} and A_{MTJ} is allowed to vary, A_{MTJ} is divided into bins (indexed as j) for numerical calculation of $P_{DEC,i}$. The decision failure probability of the array (P_{DEC}) may be calculated by first using Monte Carlo simulation to generate N access transistor I–V characteristic and calculating $P_{DEC,i}$ for each I–V characteristic. P_{DEC} may then be calculated as

$$P_{DEC} = \sum_{i=1}^{N} P_{DEC,i} \tag{3.24}$$

Because P_{DEC} depends on I_{REF}, I_{REF} may be used as a design parameter to minimize P_{DEC}. To determine the optimum I_{REF} ($I_{REF-OPT}$) that minimizes P_{DEC}, the nominal read currents through the bit-cell when the MTJ is in AP (I_{R-AP}) and when the MTJ is in P (I_{R-P}) are determined first. $I_{REF-OPT}$ is determined by minimizing P_{DEC} in the interval $[I_{R-AP}, I_{R-P}]$. A similar approach may be used to determine the decision failure probability with a voltage sensing scheme.

Finally, the total failure probability of the each memory cell ($P_{FAIL,i}$), may be calculated using

$$P_{FAIL,i} = \lim_{\delta \to 0} \sum_{all\ \ j} P\left(X - \delta \le X \le X + \delta\right) \cdot \left[1 - P\left(T_3 \le t_{MgO} \le T_4\right)\right] \tag{3.25}$$

$$T_3 = max\ \left(T_1, t_{RD-MIN,j}\right) \tag{3.26}$$

$$T_4 = min\ \left(T_2, t_{WR-MAX,j}\right) \tag{3.27}$$

where $X = A_{MTJ,j}$. Since T_3 and T_4 depend on A_{MTJ} and A_{MTJ} is allowed to vary, A_{MTJ} is divided into bins (indexed as j) for numerical calculation of $P_{FAIL,i}$. The total failure probability of the array (P_{FAIL}) may be calculated by first using Monte Carlo simulation to generate N access transistor I–V characteristic and calculating $P_{FAIL,i}$ for each I–V characteristic. P_{FAIL} may then be calculated as

$$P_{FAIL} = \sum_{i=1}^{N} P_{FAIL,i} \tag{3.28}$$

3.3.2 Optimization of STT-MRAM Memory Cells

Several STT-MRAM bit-cell designs have been published in the literature [16, 22]. STT-MRAM bit-cells can have two configurations as shown in Fig. 3.8: the "standard" connection (SC, Fig. 3.8a) and the "reversed" connection (RC, Fig. 3.8b). Furthermore, there are two possible configurations for sensing the data stored in the cell. Figure 3.6 shows one configuration where sensing is done by connecting the bit-line to the input of the sense amplifier. Note that sensing may also be done by connecting the source-line to the input of the sense amplifier instead.

Figures 3.9 and 3.10 shows the results of the failure analysis (using the methodology presented in the earlier sections) performed on SC and RC STT-MRAM bit-cells. It is clearly shown that the configurations for read and for write operations need to be carefully chosen to optimize the failure probabilities of the cell. Read failures for sensing through bit-line or through source-line may be significantly different, as Fig. 3.9a shows. For the SC bit-cell, sensing from the bit-line only has disturb failures that flip '1' to '0' (SC, P), whereas sensing from the source-line only has disturb failures that flip '0' to '1' (SC, AP). For the RC bit-

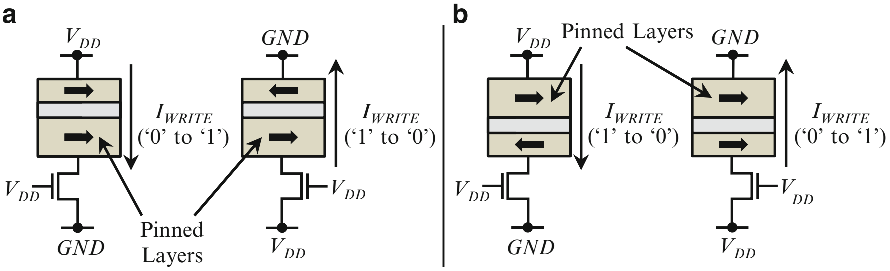


Fig. 3.8 The (**a**) standard, and (**b**) reversed connection 1T-1MTJ STT-MRAM bit-cell structures

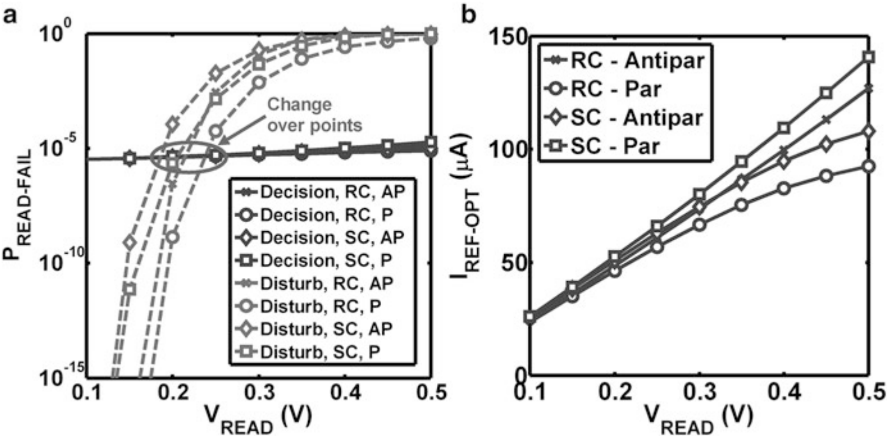


Fig. 3.9 (**a**) Decision and disturb failure probabilities were plotted with varying V_{READ} at constant ATx width. (**b**) The $I_{REF-OPT}$ corresponding to the decision failure in (**a**). V_{READ} was fixed at 0.1 V so that disturb failures are negligible

cell, sensing from the bit-line only has disturb failures that flip ‘0’ to ‘1’ (RC, AP), whereas sensing from the source-line only has disturb failures that flip ‘1’ to ‘0’ (RC, P). Interestingly, decision failures do not change significantly when V_{READ} is sufficiently small. However, the decision failure probability becomes increasingly sensitive to $I_{REF-OPT}$ (shown in Fig. 3.9b) as V_{READ} is reduced. The three failure probabilities are then plotted in the same graph, as shown in Fig. 3.10, to determine the optimum ATx width of the bit-cell. The optimum ATx width depends on whether read failures are decision dominated or disturb dominated.

3.3.3 The 2T-1MTJ STT-MRAM Bit-cell

Note that when read failures are decision dominated, the decision failure probability is minimized when ATx width is 908 nm (Fig. 3.10). However, the ATx width needs to be increased to reduce write failures. Alternatively, the design constraint can be relaxed by noting that multi-finger transistors are typically used to implement very wide transistors. Multi-finger transistors are just multiple transistors connected in such a way that their gate, source, and drain terminals are shared. When multi-finger transistors are used in the bit-cell design, the effective access transistor width may be varied using two word-lines instead of one (Fig. 3.11), and is called the *2T-1MTJ design* [21]. Word line 1 is used during read operations to switch M1 ON and OFF, while word line 2 keeps M2 OFF. During write operations, both word lines are turned ON and OFF simultaneously.

Let us analyze the 2T-1MTJ design using the failure characteristics in Fig. 3.10 as an example. The write operation of the 2T-1MTJ bit-cell requires both M1 and

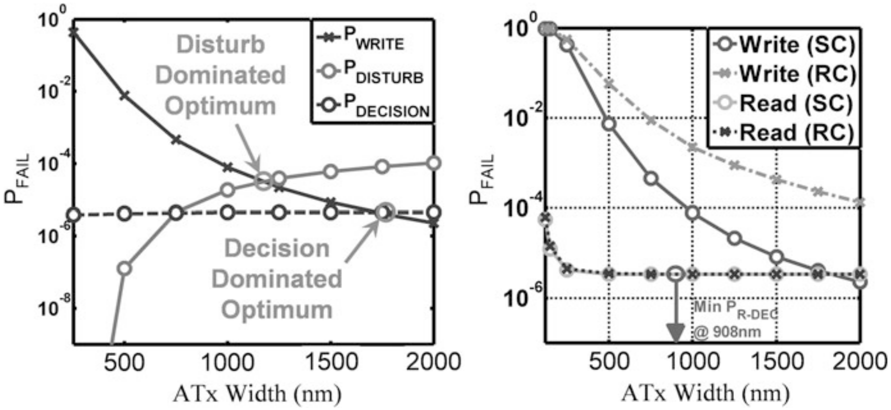


Fig. 3.10 Generally, all three failure graphs are plotted together to determine the optimum ATx width to use as shown on the *left*. The optimum point depends on whether the bit-cell failures are disturb dominated or decision dominated. However, if decision failures are the dominant read failure, then we only have to look at decision failures and write failures to determine the optimum ATx width. As shown on the *right*, decision failures are minimized at a particular ATx width while write failures keep decreasing with increasing ATx width

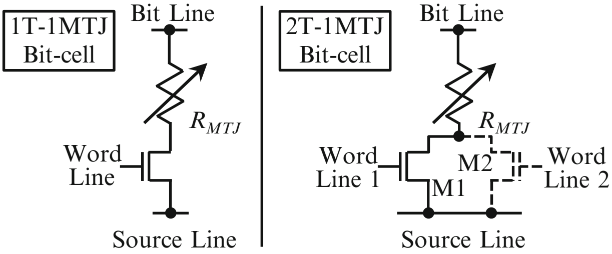


Fig. 3.11 The 2T-1MTJ bit-cell uses two access transistors with separate word lines to optimize for read failures and write failures without the need to tradeoff one for the other

M2 to be turned on. On the other hand, the read operation requires only M1 to be turned on. The size of M1 is optimized for decision failures (908 nm), while the size of M2 is as large as required to meet the write failure, array area, and array capacity requirements. Hence, the decision and the write failure probabilities of the 2T-1MTJ bit-cell may be optimized simultaneously without the need to tradeoff one for the other.

3.3.4 Stretched Write Cycle

The *stretched write cycle* (*SWC*) [23] is another optimization strategy that may be used in STT-MRAM design. SWC takes advantage of the fact that write operations

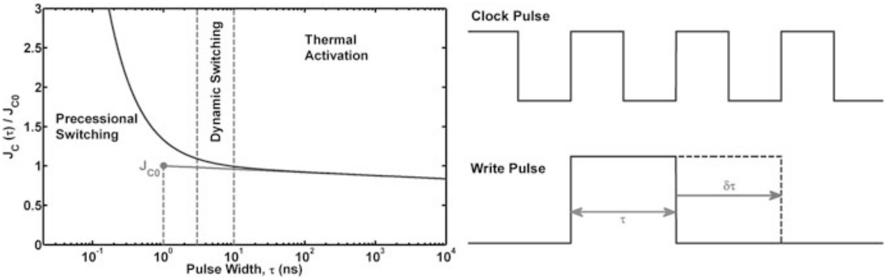


Fig. 3.12 The typical dependence of J_C on the switching time, τ, is shown on the *left*. The write pulse may be stretched by $\delta\tau$ as shown on the *right* (relative to the clock pulse) to reduce J_C

do not occur as frequently as read operations in last level caches. The critical current required for writing into the STT-MRAM bit-cell may then be lowered by allowing a longer time for write operations to complete, as shown in Fig. 3.12.

The write energy comparisons of the optimization techniques presented are shown in Fig. 3.13. The worst case design that mitigates write failures by write-voltage boosting has 18 % higher power dissipation as compared to the nominal design without process variations. The 1T-1MTJ bit-cell energy overhead is reduced to 11 % after optimization, resulting in an area overhead of 5.4 %. However, if an optimized 2T-1MTJ design is used, the energy overhead is reduced to 9 % while the area overhead is increased to 9 %. Finally, the energy dissipation becomes 3 % lower than the nominal case when SWC is used with an optimized 1T-1MTJ bit-cell. This is because the critical write current needed is significantly lower in SWC. Although the write frequency is reduced by 50 % in SWC, the throughput penalty is only 3 %. Hence, we conclude that circuit/architecture co-design can lead to ultralow power last level caches based on STT-MRAMs.

3.4 Comparisons of Cache Arrays Based on SRAM and STT-MRAM

A cache comprises of multiple arrays for storing tags and data bits. In conventional on-chip caches, both the tag and data arrays are implemented using SRAM. Since the tag array requires frequent and fast updates of status bits and history bits, the write latency of STT-MRAM may significantly impact the performance STT-MRAM based tag arrays [24]. Hence, the STT-MRAM cache we will be discussing is a hybrid cache where the tag arrays are implemented using SRAM and the data arrays are implemented using STT-MRAM. In order to estimate the overall cache latency, area and energy consumption of the STT-MRAM cache, the CACTI 6.5 simulator [25] needs to be modified to consider (a) analog read circuits in STT-

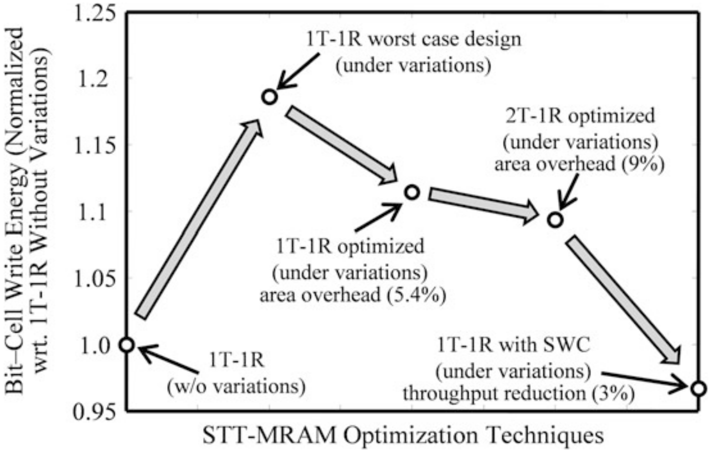


Fig. 3.13 Write energy comparison of the bit-cell optimization techniques and the overhead associated with each optimization technique

MRAM data arrays, (b) SRAM-based tag arrays along with STT-MRAM data arrays, and (c) the bit-cell layout geometries to optimize the array aspect ratio.

A comparison of caches designed with SRAM and STT-MRAM is shown in Fig. 3.14. Note that the capacity of the cache array, and not the cache area, has a more significant impact on whether caches designed with STT-MRAM outperform caches designed with SRAM. As the cache capacity increases, the wire delays in SRAM based caches increases much faster than that in STT-MRAM based caches due to the larger bit-cell footprint. Hence, high capacity caches designed with STT-MRAM have faster access time and are smaller than SRAM based caches. As Fig. 3.14 shows, an 8 MB cache designed with STT-MRAM has lower read latency than an iso-capacity cache designed with SRAM. Similarly, the write latency gap between STT-MRAM based and SRAM based caches reduces with increasing cache capacity.

A similar trend is observed in the dynamic energy consumption of the caches (Fig. 3.14). The energy dissipated in read operations in STT-MRAM based caches is higher than that of SRAM based caches due to power dissipation in the analog read circuits, despite 75 % smaller total cache area. However, the energy dissipation due to interconnects becomes dominant when cache capacity is 1 MB and higher. Therefore, read operation dynamic energy is significantly lower in STT-MRAM based caches. During write operations, STT-MRAM caches dissipate significantly larger energy than SRAM based caches. Finally, the leakage in STT-MRAM based caches is significantly lower than that in SRAM based cache because STT-MRAM bit-cells are non-volatile and have zero standby power. Only the SRAM based tag arrays and periphery dissipate leakage power in caches designed with STT-MRAM.

The total energy dissipation in a cache also depends on factors such as cache access patterns (number of read and write operations) and cache utilization (number

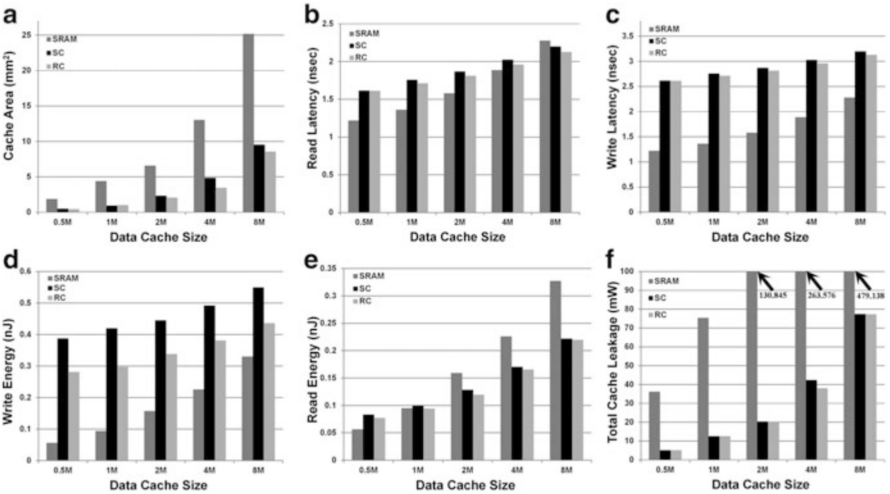


Fig. 3.14 (**a**) Array area of SRAM and STT-MRAM based caches (4-way, 64 B cache line, *B* byte, *M* mega byte), (**b**) read latency and (**c**) write latency, (**d**) read energy per operation and (**e**) write energy per operation, and (**f**) total leakage power

of times a processor accesses the cache per unit cycle). The cache utilization is lower than 30 % in today's processors [26]. Moreover, for lower levels of the cache hierarchy, the cache utilization is significantly lower than 30 %. We have measured L2 cache utilizations for various SPEC2000 benchmarks based on the Simplescalar framework [27] with a 32 KB L1 cache configuration. For a majority of the benchmarks, L2 cache utilization is lower than 3 %. The highest utilization, observed for the AMMP benchmark, is about 13 %, and the average utilization across 16 benchmarks is only 2.2 %.

As shown in Fig. 3.15, a 2 MB STT-MRAM cache shows similar or lower energy consumption than a 0.5 MB SRAM cache when the utilization is lower than 10 %. Although the STT-MRAM cache has significantly lower energy consumption at 0 % utilization (leakage only), the energy dissipation increases drastically due to excessive write energy as the utilization increases. The results are obtained using the following conditions: read and write operation ratio of 2:1, 2 GHz processor speed, and total simulation time of 1 billion processor cycles. Therefore, an STT-MRAM cache can achieve high energy-efficiency along with high capacity in comparison to an SRAM cache, especially in lower levels of the cache hierarchy due to the low cache utilization.

In a conventional SRAM array, column selection is required for storing multiple words in a single row [28]. Since set associativity is common in modern caches, column selection in SRAM arrays is imperative. Furthermore, bit-interleaving can only be achieved by employing column selection. Bit-interleaving is a commonly adopted technique in SRAM arrays (1) to mitigate soft errors [28], and (2) to increase array density by bit-line multiplexing [25]. In the column selection

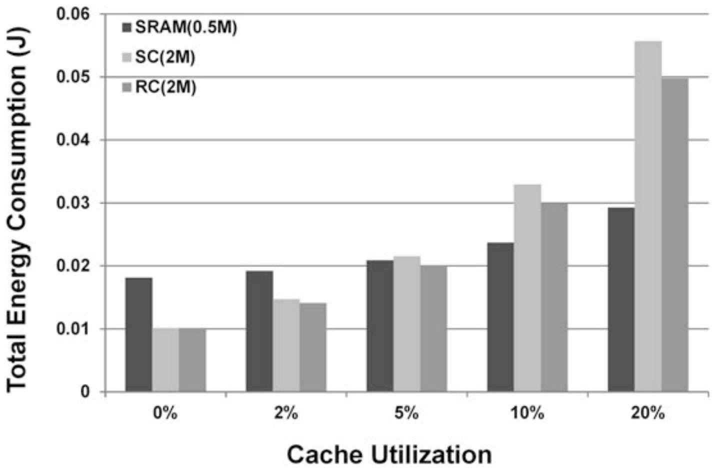


Fig. 3.15 Total energy consumption versus cache utilization for SRAM and for STT-MRAM based caches shows that when 0.1 M data is stored in cache, STT-MRAM dissipation is much lower

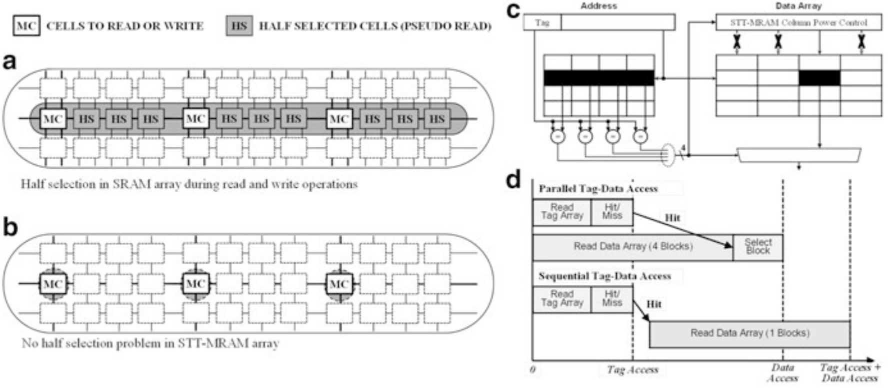

MC CELLS TO READ OR WRITE
HS HALF SELECTED CELLS (PSEUDO READ)
a
MC
HS
Half selection in SRAM array during read and write operations
b
MC
No half selection problem in STT-MRAM array
c
Address
Tag
Data Array
STT-MRAM Column Power Control
4
d
Parallel Tag-Data Access
Read Tag Array
Hit/ Miss
Hit
Read Data Array (4 Blocks)
Select Block
Sequential Tag-Data Access
Read Tag Array
Hit/ Miss
Hit
Read Data Array (1 Blocks)
0
Tag Access
Data Access
Tag Access + Data Access

Fig. 3.16 (**a**) SRAM cache access dissipates additional power in the bit-lines of unselected cells, whereas (**b**) STT-MRAM based cache do not have the half-select problem. (**c**) Tag-data access needs to be sequential to take advantage of the lack of half-select problem in STT-MRAM. (**d**) Sequential tag-data access incurs additional read latency since the cache hit needs to occur before reading data

operation of an SRAM array, all unselected bit-cells in the accessed row have to be under read mode to prevent unexpected bit flips, when a word-line is asserted. This phenomenon is commonly known as pseudo-read or half-selection [28]. Note that, in an STT-MRAM array, the non-volatility of bit-cells can eliminate the half selection problem. As presented in Fig. 3.16, the unselected bit-cells can remain in standby mode, and hence, consume no energy during both read and write column selection operations. However, a sequential tag-data access is needed in order to determine which of the columns need to be the selected prior to actual access, which increases the read latency since a cache hit must occur prior to reading the data array. Based on our simulation parameters, the average read latency penalty is about 500 ps for the 2 MB STT-MRAM based caches. However, the read energy savings is about 40–50 %.

3.5 Conclusion

Based on the simulation results presented in this chapter, we may conclude that spin-transfer torque MRAM is becoming more viable as a technology for on-chip last-level caches. Significant energy savings are achieved due to the large cache capacities enabled by the small footprint of STT-MRAM memory cells. Further reduction in the critical switching current of STT-MRAM will increase the achievable energy savings [29]. The non-volatility of STT-MRAM may also be exploited to enable a new "normally-off" computing paradigm [30]. However, crucial design issues need to be overcome for STT-MRAM to be viable for caches next to the processor and become a truly universal memory technology. For example, the lack of a self-referenced differential sensing scheme in STT-MRAMs limits the performance of its read operations and also its robustness against process variations. Hence, there is a need to explore alternative MTJ structures to improve STT-MRAM performance, and it may take some time before suitable structures become a reality. Even so, STT-MRAM offers exciting possibilities in integrating new functionality into on-chip caches in its current form [31]. This ability to integrate new functionality on-chip to complement the CMOS circuitry may be key in driving the future adoption of on-chip STT-MRAM technology.

References

1. ITRS Roadmap 2014. http://www.itrs.net.
2. Huai Y. Spin-transfer torque MRAM (STT-MRAM): challenges and prospects. AAPPS Bull. 2008;18(6):33–40.
3. Datta D, Behin-Aein B, Datta S, Salahuddin S. Voltage asymmetry of spin-transfer torques. IEEE Trans Nanotechnol. 2012;11(2):261–72.
4. Dorrance R, Ren F, Toriyama Y, Hafez AA, Yang CK, Markovic D. Scalability and design-space analysis of a 1T-1MTJ memory cell for STT-RAMs. IEEE Trans Electron Devices. 2012;59(4):878–87.
5. Fong X, Choday SH, Roy K. Bit-cell level optimization for non-volatile memories using magnetic tunnel junctions and spin-transfer torque switching. IEEE Trans Nanotechnol. 2012;11(1):172–81.
6. Slonczewski JC. Current-driven excitation of magnetic multilayers. J Magn Magn Mater. 1996;159(1–2):L1–7.
7. Berger L. Emission of spin waves by a magnetic multilayer traversed by a current. Phys Rev B. 1996;54(13):9353–8.
8. Myers EB. Current-induced switching of domains in magnetic multilayer devices. Science. 1999;285(5429):867–70.

9. Katine J, Albert F, Buhrman R, Myers E, Ralph D. Current-driven magnetization reversal and spin-wave excitations in Co/Cu/Co pillars. Phys Rev Lett. 2000;84(14):3149–52.
10. Huai Y, Albert F, Nguyen P, Pakala M, Valet T. Observation of spin-transfer switching in deep submicron-sized and low-resistance magnetic tunnel junctions. Appl Phys Lett. 2004;84(16):3118.
11. Ikeda S, Miura K, Yamamoto H, Mizunuma K, Gan HD, Endo M, Kanai S, Hayakawa J, Matsukura F, Ohno H. A perpendicular-anisotropy CoFeB-MgO magnetic tunnel junction. Nat Mater. 2010;9(9):721–4.
12. Kishi T, Yoda H, Kai T, Nagase T, Kitagawa E, Yoshikawa M, Nishiyama K, Daibou T, Nagamine M, Amano M, Takahashi S, Nakayama M, Shimomura N, Aikawa H, Ikegawa S, Yuasa S, Yakushiji K, Kubota H, Fukushima A, Oogane M, Miyazaki T, Ando K. Lower-current and fast switching of a perpendicular TMR for high speed and high density spin-transfer-torque MRAM. In: 2008 IEEE International Electron Devices Meeting, 2008, p. 1–4.
13. Sun J. Spin-current interaction with a monodomain magnetic body: a model study. Phys Rev B. 2000;62(1):570–8.
14. Salahuddin S, Datta S. Self-consistent simulation of quantum transport and magnetization dynamics in spin-torque based devices. Appl Phys Lett. 2006;89(15):153504.
15. Yuasa S, Nagahama T, Fukushima A, Suzuki Y, Ando K. Giant room-temperature magnetoresistance in single-crystal Fe/MgO/Fe magnetic tunnel junctions. Nat Mater. 2004;3(12):868–71.
16. Lin CJ, Kang SH, Wang YJ, Lee K, Zhu X, Chen WC, Li X, Hsu WN, Kao YC, Liu MT, Nowak M, Yu N. 45nm low power CMOS logic compatible embedded STT MRAM utilizing a reverse-connection 1T/1MTJ cell. In: 2009 IEEE International Electron Devices Meeting (IEDM), 2009, p. 1–4.
17. Mansuripur M, Giles R. Demagnetizing field computation for dynamic simulation of the magnetization reversal process. IEEE Trans Magn. 1988;24(6):2326–8.
18. Newell AJ, Williams W, Dunlop DJ. A generalization of the demagnetizing tensor for nonuniform magnetization. J Geophys Res. 1993;98(B6):9551–5.
19. Brown W. Thermal fluctuations of a single-domain particle. Phys Rev. 1963;130(5):1677–86.
20. Fong X, Gupta SK, Mojumder NN, Choday SH, Augustine C, Roy K. KNACK: a hybrid spin-charge mixed-mode simulator for evaluating different genres of spin-transfer torque MRAM bit-cells. In: 2011 International Conference on Simulation of Semiconductor Processes and Devices, 2011, p. 51–4.
21. Li J, Ndai P, Goel A, Salahuddin S, Roy K. Design paradigm for robust spin-torque transfer magnetic RAM (STT MRAM) from circuit/architecture perspective. IEEE Trans Very Large Scale Integr Syst. 2010;18(12):1710–23.
22. Jeong G, Cho W, Ahn S, Jeong H, Koh G, Hwang Y. A 0.24-μm 2.0-V 1T1MTTJ 16-kb nonvolatile magnetoresistance RAM with self-reference sensing scheme. IEEE J Solid-State Circuits. 2003;38(11):1906–10.
23. Augustine C, Mojumder NN, Fong X, Choday SH, Park SP, Roy K. Spin-transfer torque MRAMs for low power memories: perspective and prospective. IEEE Sens J. 2012;12(4):756–66.
24. Rasquinha M, Choudhary D, Chatterjee S, Mukhopadhyay S, Ram TSTT, Yalamanchili S. An energy efficient cache design using Spin Torque Transfer (STT) RAM. In: 2010 IEEE/ACM International Symposium on Low-Power Electronics and Design (ISLPED), 2010, p. 389–94.
25. Muralimanohar N, Balasubramonian R, Jouppi N. Optimizing NUCA organizations and wiring alternatives for large caches with CACTI 6.0. In: 40th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO 2007), 2007, p. 3–14.
26. Ramaswamy S, Yalamanchili S. An utilization driven framework for energy efficient caches. In: Proceedings of the 15th International Conference on High Performance Computing (HiPC'08), 2008, p. 583–94.
27. Simplescalar. 2011 LLC. http://www.simplescalar.com.

28. Park SP, Kim SY, Lee D, Kim J-J, Griffin WP, Roy K. Column-selection-enabled 8T SRAM array with ∼1R/1W multi-port operation for DVFS-enabled processors. In: IEEE/ACM International Symposium on Low Power Electronics and Design, 2011, p. 303–8.
29. Yoda H, Fujita S, Shimomura N, Kitagawa E, Abe K, Nomura K, Noguchi H, Ito J. Progress of STT-MRAM technology and the effect on normally-off computing systems. In: 2012 International Electron Devices Meeting, 2012, p. 11.3.1–4.
30. Kawahara T. Challenges toward gigabit-scale spin-transfer torque random access memory and beyond for normally off, green information technology infrastructure (Invited). J Appl Phys. 2011;109(7):07D325.
31. Lee D, Fong X, Roy K. R-MRAM: a ROM-embedded STT MRAM cache. IEEE Electron Device Lett. 2013;34(10):1256–8.

Chapter 4
Embedded STT-MRAM: Device and Design

Seung H. Kang and Seong-Ook Jung

Abstract Spin-transfer-torque magnetoresistive random access memory (STT-MRAM) is made of a combination of semiconductor integrated circuits (IC) and a dense array of nanometer-scale magnetic tunnel junctions (MTJ). This emerging memory is of growing technological interest due to its potential to bring disruptive device innovation to the world of electronics. STT-MRAM is capable of providing high speed, unlimited endurance, and nonvolatility simultaneously, which is often recognized as a unique advantage over conventional and other emerging memories. While the technology is at an early stage and evolving in multiple platforms, STT-MRAM is particularly compelling as an embedded memory for system-on-chip (SOC). STT-MRAM can be integrated into SOC without altering baseline logic platforms both in process and in design. This chapter overviews key device and circuit subjects from the perspective of co-designing logic and MTJ.

4.1 Introduction

Generic device scaling no longer secures the evolution of IC, causing the silicon-based technology to face unprecedented challenges in materials, devices, and processes. These challenges translate to compromises in power dissipation, performance, and cost for a wide range of IC products. While the end of physical scaling is not imminent, its value is being heavily eroded by the growing technological and economic concerns at the nanoscale. Some of promising innovations that can mitigate or overcome such problems may be found in spintronic IC. In the past few years, the spintronics community has achieved significant discoveries and breakthroughs [1]. Most recognized is the emergence of STT-MRAM [2–6]. Key discoveries and advances have triggered industry-wide R&D efforts in pursuit of an alternative memory in lieu of conventional memories that are not only facing acute

S.H. Kang (✉)
Qualcomm Technologies Inc., 5775 Morehouse Dr., San Diego, CA 92121, USA
e-mail: kang@qti.qualcomm.com

S.-O. Jung
Yonsei University, Seoul, South Korea
e-mail: sjung@yonsei.ac.kr

R.O. Topaloglu (ed.), *More than Moore Technologies for Next Generation Computer Design*, DOI 10.1007/978-1-4939-2163-8_4

tradeoffs in performance and power, but also nearing fundamental scaling limits. In parallel, various forms of MTJ-based novel logic devices and circuits have been demonstrated [7–9], opening a possible path for spintronic IC to expand beyond memory applications. Furthermore, a novel computing architecture concept, known as normally-off computer, was proposed as a way to reduce the energy consumption of modern microprocessors [10–12]. Still at an early stage in its endeavor, the global spintronics community continues to propel a plethora of innovations in materials, devices, circuits, and architectures.

STT-MRAM is particularly compelling as an embedded memory for SOC. In contrast to standalone commodity memories, each type of SOC requires a different combination of memory attributes such as speed, energy consumption, and reliability including cyclic endurance and data retention. STT-MRAM can be offered in a variety of macros whose designs are customized for application-specific SOC. In general, density requirements are found over a wide range (a few kbits to 256 Mbits). Yet, even in small densities, it can realize significant values in system performance, energy consumption, security, and cost, when device and circuit attributes are tailored at a system-architecture level. Furthermore, the memory element MTJ can be integrated in a fully logic-compatible way without altering or adversely impacting baseline logic platforms by adding two or three mask layers into a back-end-of-line (BEOL) flow [3].

Driving STT-MRAM beyond discrete devices and arrays toward SOC necessitates extensive learning cycles in device, circuit, yield, and reliability engineering. In order to produce variability- and fault-tolerant STT-MRAM, a systematic design methodology is required to assure robust functionality of STT-MRAM over a wide range of process-voltage-temperature (PVT) windows. This chapter overviews key device and circuit subjects from the perspective of co-designing logic and MTJ to enable STT-MRAM as a scalable custom embedded memory to serve advanced SOC.

4.2 Device Physics

4.2.1 Magnetic Tunnel Junction (MTJ)

A MTJ is a building block as a storage element for STT-MRAM. A MTJ consists of metallic ferromagnetic films separated by an oxide tunnel barrier, typically an ultra-thin magnesium oxide (MgO). The conductance of a ferromagnetic metal-insulator-ferromagnetic metal (FM1-I-FM2) structure is governed by tunnel magnetoresistance, a quantum mechanical phenomenon that results from spin-dependent tunneling [13]. When conduction electrons are emitted from one ferromagnetic metal electrode FM2, schematically illustrated in Fig. 4.1, they are spin-polarized to the magnetization direction of FM2 and tunnel through the thin tunnel barrier with their spin states conserved. The electron density of states in the opposite

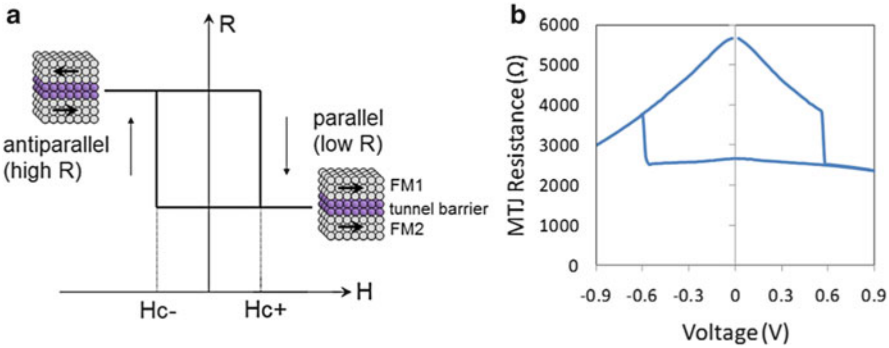


Fig. 4.1 (**a**) Conceptual illustration of a MTJ hysteresis curve. A binary resistance state is obtained through configuring the magnetization of FM1 (free layer) with respect to that of FM2 (pinned layer) either parallel or antiparallel at the switching field H_{c+} or H_{c-}, respectively; (**b**) a MTJ hysteresis curve driven by spin-transfer-torque (STT) switching

ferromagnetic metal electrode FM1 that these tunneling electrons encounter is dependent on the magnetization direction of FM1. Consequently, the electrical resistance (R) of FM1-I-FM2 structure is determined by relative orientations of the magnetizations, which is described by [14]

$$R = \frac{R_{\perp}}{1 + \frac{TMR}{2}\cos\theta} \tag{4.1}$$

where θ is the angle between the two configurations, $R_{\perp}$ is the resistance measured in the perpendicular magnetic configuration ($\theta = \pi/2$). R becomes minimum (R_p) for the parallel magnetization configuration ($\theta = 0$) and maximum (R_{ap}) for the anti-parallel configuration ($\theta = \pi$). Accordingly, a MTJ serves as a variable resistor that can be configured to have binary states (0 and 1) defined by two discrete resistance values (R_p and R_{ap}, respectively). The tunnel magnetoresistance ratio (TMR) is then defined as:

$$TMR = \frac{R_{ap} - R_p}{R_p} \times 100\,\% \tag{4.2}$$

TMR is one of critical device parameters for the design of STT-MRAM for error-free and high-speed read operations since the signal margin for sensing an array of MTJ is governed by TMR.

Figure 4.2 illustrates typical MTJ film stacks that essentially consist of metallic films separated by a tunnel barrier, most commonly MgO on the order of 1 nm in thickness. Depending on the orientation of the magnetization with respect to the film plane, two representative cases are shown here: (a) in-plane MTJ (i-MTJ); and (b) perpendicular MTJ (p-MTJ). The free layer is a soft ferromagnetic metal

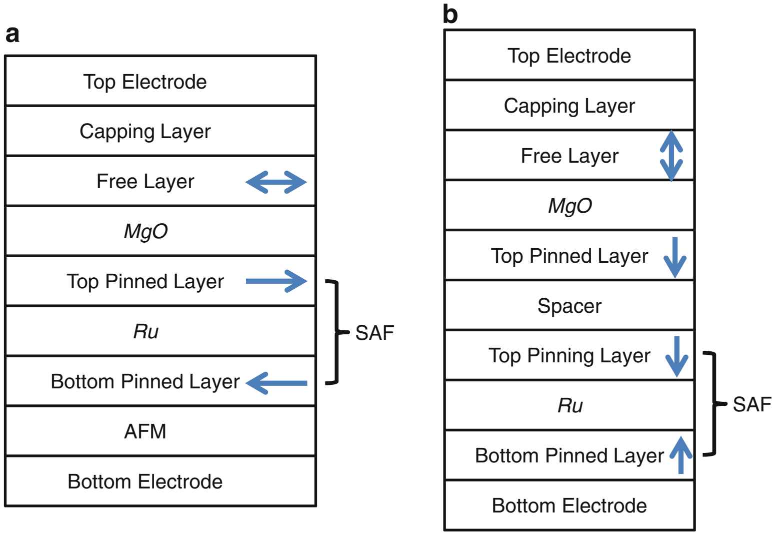


Fig. 4.2 Schematic illustration of typical MTJ film stacks: (**a**) in-plane MTJ (i-MTJ); (**b**) perpendicular MTJ (p-MTJ)

(e.g. CoFeB) whose magnetization can be switched by STT. The reference layer is commonly a synthetic structure to provide a reference magnetization fixed in one direction (e.g. top pinned layer in Fig. 4.2a) relative to the free layer magnetization. For i-MTJ, the magnetization of the bottom pinned layer is fixed by an antiferromagnet (AFM) pinning layer (e.g. PtMn) via the exchange bias effect. The top pinned layer is then antiferromagnetically coupled to the bottom pinned layer via interlayer exchange coupling with a non-magnetic spacer (e.g. Ru). This type of reference layer is called a synthetic antiferromagnet (SAF). In comparison, for p-MTJ, the out-of-plane magnetization of the bottom pinned layer can be developed inherently during film formation, hence the AFM pinning layer is not necessary (Fig. 4.2b). Ordinarily, the reference layer stack of p-MTJ is still structured in a SAF configuration. To promote high TMR, however, p-MTJ often necessitates an additional pinned layer (commonly in CoFeB) underneath the tunnel barrier which is exchange-coupled with the SAF.

The metallic films of MTJ are deposited by physical vapor deposition (PVD). The MgO barrier can be grown by PVD or a combination of PVD and oxidation. MTJ device properties are tailored through a selection of desired materials and a precise control of microstructure, film thickness, and cross-sectional feature size. Key MTJ parameters, critical to optimizing performance, energy consumption and reliability, include: TMR, resistance–area product (RA), energy barrier (E_B) and switching current density (J_c).

Experimental TMR reached ~600 % at room temperature in a CoFeB/MgO/Co FeB junction [15]. From a microstructure perspective, the most critical factor in achieving high TMR is promoting strong MgO (001) texture. For practical device applications, high TMR needs to be achieved in conjunction with relatively low RA, preferably, $<10\ \Omega\ \text{cm}^2$. TMR of 253 % at RA $= 5.9\ \Omega\ \text{cm}^2$ has been demonstrated by inserting CoFe as a crystallization template to induce preferred grain growth in MgO and then to promote crystallization of CoFeB through annealing [16]. In-situ annealing of the MgO barrier has been known to promote the (001) texture further, resulting in high TMR ($>$170 %) even for MTJ films with an ultralow RA ($\sim 1\ \Omega\ \text{cm}^2$) [17].

At a static mode, MTJ maintains its resistance state without power (i.e. non-volatile) as long as the magnetic anisotropy of its free layer is greater than the thermal excitation energy described by k_BT where k_B is the Boltzmann constant and T is temperature. For i-MTJ which is typically patterned into an elliptically shaped cell, the free layer magnetic moment can have only two energetically favorable states aligned with the long-axis (called easy-axis) of the MTJ, thereby allowing either R_p or R_{ap}. For p-MTJ, the two states are determined by out-of-plane moments. Hence p-MTJ does not require a particular shape and is typically patterned in a circular shape.

Under the simplified assumption of a single-domain free layer, the energy barrier (E_B) between the two energetically favorable states is often given by

$$E_B = \frac{M_s H_k V}{2}\left(1 - \frac{H_{ext}}{H_k}\right)^2 \tag{4.3}$$

where M_s is the saturation magnetization of the free layer, V is the free layer volume, H_k is the effective uniaxial anisotropy field, and H_{ext} is the external field present along the easy-axis (which vanishes in the absence of any stray field). For MTJ to be non-volatile, E_B must be larger than the thermal excitation energy over a range of operating and storage temperatures. For example, for a single MTJ to retain its state for 10 years, E_B must be $40\,k_BT$ (1 eV) or greater. A recent report demonstrated that E_B can be $100\,k_BT$ or greater, which is remarkable for p-MTJ on the order of 30 nm in diameter [6].

4.2.2 Spin-Transfer-Torque (STT) Switching

A traditional way of programming MTJ is to apply a magnetic field to switch the free layer magnetization. A drawback of this method is the requirement of large current to induce sufficient magnetic field. It is also well understood that this method does not provide good scalability because decreasing the MTJ size entails larger switching fields, hence, even more current.

A breakthrough in physics of MTJ switching was accomplished in 1996 by the theoretical formulation that the free layer magnetization could be modulated by the direct transfer of spin angular momentum from spin-polarized electrons [18, 19]. This phenomenon, called spin-transfer-torque (STT) magnetization reversal, delivered a new means to control the free layer magnetization by directly applying electric current through MTJ without a need of magnetic field. The magnitude of STT scales with the current density (J). This is particularly beneficial for device scalability since the critical switching current (I_c) should scale proportionally to the size of MTJ. A breakthrough demonstration of STT-MRAM at an array level was first reported in 2005, including TMR of 160 % and switching speed as fast as 1 ns [2].

For i-MTJ, the intrinsic critical switching current (I_{c0}) is given by

$$I_{c0} = \frac{2e}{\hbar}\,\frac{\alpha}{\eta} M_S V \left(H_{k||} + \frac{H_d - H_{k\perp}}{2} \right) \tag{4.4}$$

where α is the damping constant, η is the spin polarization constant, $H_{k||}$ is the uniaxial anisotropy field in the film plane, H_d is the effective perpendicular demagnetization field that corresponds to the field required to saturate the free layer moment perpendicular to the film plane, and $H_{k\perp}$ is the anisotropy field perpendicular to the plane. The H_d term, given by $4\pi M_s$, represents an additional energy term that needs to be overcome during STT switching because the shape anisotropy induces an oscillatory motion of magnetization confined in the direction perpendicular to the film plane, resulting in an elliptical precession. Undesirably, H_d only greatly increases I_{c0} without contributing to E_B. A technological challenge in building STT-MRAM is to reduce I_{c0} while maintaining sufficient E_B. Hence, an effective way of reducing I_{c0} without degrading E_B is to introduce perpendicular anisotropy $H_{k\perp}$ to cancel a substantial portion of H_d.

Considered as an essential figure of merit, the STT switching efficiency is described by the ratio of E_B and I_{c0}. For i-MTJ, it is typically on the order of 0.5–1 $k_BT/\mu A$. This allows good scalability for i-MTJ as small as approximately 40 nm (short axis). However, the success of STT-MRAM is largely dependent on whether MTJ can be scaled to deep nanoscale nodes (30 nm and below) in conjunction with low switching energy and high stability. Unless the STT efficiency is raised significantly, i-MTJ may not provide sufficient E_B for nonvolatility, which limits physical scaling of i-MTJ for future nodes.

This scalability challenge can be overcome by adopting p-MTJ which provides much greater anisotropy even in small features. E_B of p-MTJ is determined by crystalline or interface perpendicular magnetic anisotropy (PMA), not by the shape anisotropy of i-MTJ. Various PMA materials have been investigated, which include $L1_0$-ordered FePt or FePd alloys, Co-based superlattices such as Co/Pt and Co/Ni laminates, rare-earth/transition metal alloys, etc. To build useful MTJ devices, however, these materials must be engineered for an optimal combination of materials properties like M_s and H_k and device properties like TMR and J_c. A prior report addressed that the anisotropy resulting from the CoFeB-MgO interface can induce large $H_{k\perp}$ [4]. When CoFeB is sufficiently thin (typically ~1.5 nm or thinner), such interface PMA can overcome the demagnetization field, i.e., $H_{k\perp} > H_d$. The film can then become magnetized fully perpendicular to the plane. With further tuning of the stack, such interface PMA can be achieved for even thicker CoFeB. Recently, p-MTJ devices utilizing interfacial PMA of CoFeB have successfully been engineered for fully functional 8 Mbit STT-MRAM [6].

Referring to Eq. (4.4), I_{c0} and E_B pertaining to p-MTJ with interfacial PMA are described by

$$I_{c0} = \frac{e}{\hbar}\,\frac{\alpha}{\eta}\,M_S V H_{k\perp}^{eff} \tag{4.5}$$

$$E_B = \frac{M_S V H_{k\perp}^{eff}}{2} \tag{4.6}$$

where $H_{k\perp}^{eff}$ is the effective perpendicular anisotropy field. In contrast to i-MTJ described by Eqs. (4.3) and (4.4), note that I_{c0} is directly proportional to E_B. The absence of the H_d term means that STT switching is far more efficient. This leads to substantially higher STT efficiency (E_B/I_{c0}). Recently, $E_B/I_{c0} \sim 5\,k_B T/\mu A$ has been reported from an array of ~30 nm p-MTJ [6], suggesting that the STT efficiency of p-MTJ could be an order of magnitude greater than that of in-plane MTJ. This is a significant breakthrough demonstrating the scalability of p-MTJ based on interfacial PMA of CoFeB, which is a preferable material to achieve high TMR as well.

4.3 Device Engineering

4.3.1 Bitcell and Array

STT-MRAM is a hybrid IC built on a combination of semiconductor logic and MTJ. Its bitcell which represents 1 bit is commonly architected in 1 transistor plus 1 MTJ (1T-1J). As shown in Fig. 4.3, a MTJ is connected in series to an n-type metal oxide semiconductor transistor (NMOS). This transistor is called an access transistor since it controls read and write access to the connected MTJ as a digital switch.

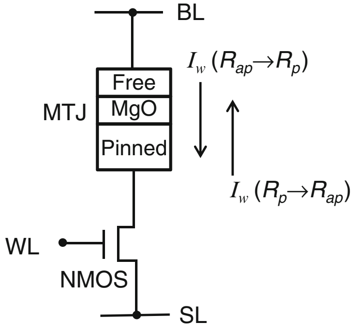


Fig. 4.3 Schematic representation of a 1T-1J bitcell

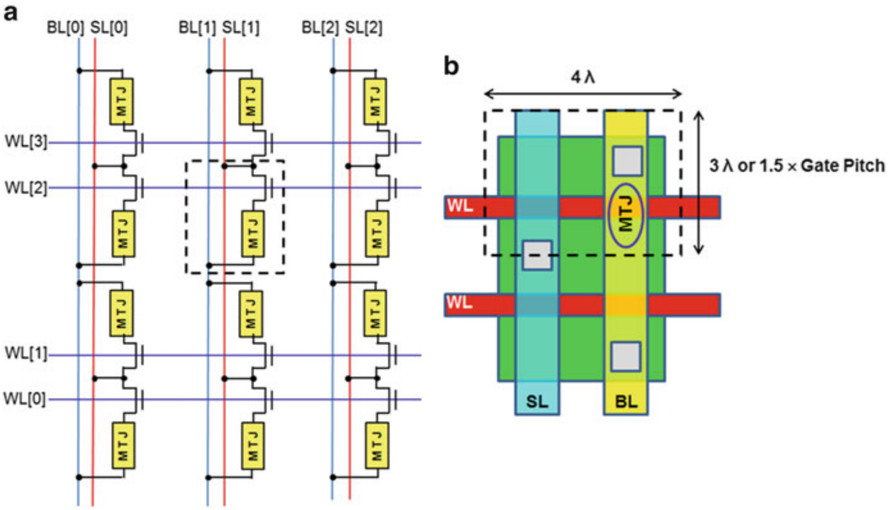


Fig. 4.4 STT-MRAM array architecture with SL parallel to BL: (**a**) 4 × 3 array; (**b**) 1T-1J bitcell layout

Figure 4.4 is a schematic representation of a typical STT-MRAM array that consists of 1T-1J bitcells. To read the information stored in a cell, the word line (WL) of the selected cell is turned on and a small read current is applied by a sensing circuit (Sect. 4.4) to either the selected bit line (BL) or the source line (SL) with the other end of the cell grounded (*GND*). A sense amplifier determines the cell state by sensing the difference between the cell resistance and the reference resistance predefined from a reference MTJ array. In comparison, the write operation requires bidirectional currents because the direction of write current determines which resistance state (R_p or R_{ap}) is programmed to MTJ. With the bitcell architecture shown in Fig. 4.4, for $R_{ap} \rightarrow R_p$, a write voltage is applied to BL ($V_{BL} = V_{DD}$) with WL turned on ($V_{WL} = V_{DD}$) and SL grounded ($V_{SL} = 0$ V, *GND*), and vice versa for $R_p \rightarrow R_{ap}$. For successful write operation, the write current (I_w) supplied to the MTJ in each bitcell must be larger than the MTJ critical switching current (I_c).

Figure 4.4b shows an example layout of 1T-1J bitcell with an array architecture illustrated in Fig. 4.4a. Provided that the minimum metal half-pitch is λ, two metal lines BL and SL running in parallel limit the minimum bitcell width to 4 λ. Then the metal plate connected to the source and the drain of the access transistor may limit the bitcell height to 3 λ or 1.5 times of the gate pitch. Assuming the metal pitch is larger than the gate pitch, the bitcell size can be as small as 12 λ^2.

The array architecture shown in Fig. 4.4 is simple to design and operate. One shortcoming of this structure is that every BL is coupled with its own SL, thereby causing a larger array footprint. A more compact array can be realized by placing SL orthogonal to BL, as shown in Fig. 4.5. SL is then parallel to WL and shared between two neighboring rows of WL. With this architecture, the bitcell size can be

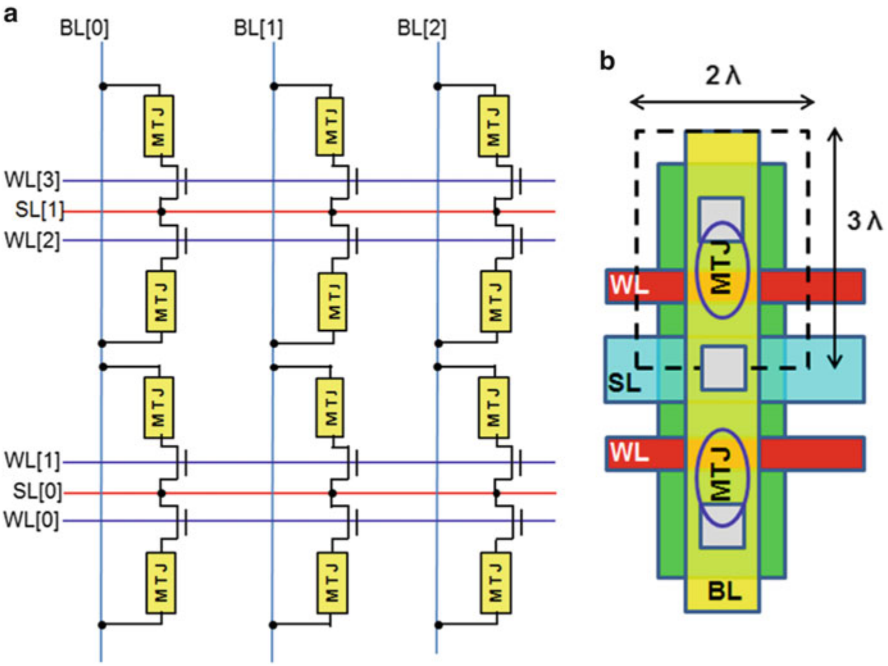


Fig. 4.5 STT-MRAM array architecture with SL orthogonal to BL: (**a**) 4 × 3 array; (**b**) 1T-1J bitcell layout

as small as 6 λ^2 (Fig. 4.5b), a half of the size in Fig. 4.4. However, this architecture results in more complex write operation for $R_p \rightarrow R_{ap}$. When SL is raised to a write voltage, the selected BL is grounded. Simultaneously, all the unselected BL associated with the selected WL must be raised to the same level of the write voltage to avoid unintentional current flows to the unselected MTJ. Consequently, this architecture consumes more power during write operation. Furthermore, this may even necessitate two separate write pulses to complete a full write cycle, since in chip-level operation each full cycle carries multiple bits (typically, 32, 64, or 128 bits) of R_p and R_{ap} concurrently. Accordingly, this architecture is not desirable for low power and high speed applications.

Table 4.1 describes an example of the attributes of a bitcell embedded for a 45-nm low-power logic platform [3]. The bitcell size is ~50 F^2, where F is 45 nm (minimum feature size of this node). When the half-metal pitch λ is used, the size is ~20 λ^2 since λ is 70 nm. This is significantly larger than that of an ideal layout of the same array architecture in Fig. 4.4b, which is attributed to the constraints of logic design rules of this particular logic technology.

Table 4.1 Key attributes of an embedded STT-MRAM bitcell demonstrated for a 45-nm low-power logic platform [3]

Feature size: F/λ	45/70 nm
V_{DD} (core/IO)	1.1/1.8 V
Cell architecture	1T-1J (reversely connected)
Bitcell size	0.1026 μm^2
Access NMOS (length/width)	40/270 nm
MTJ size	40 nm (short axis)
MTJ aspect ratio	2.5–3
TMR/RA	110 %/9 Ω μm^2
BEOL	Cu/low-k Seven metal layers

F is the minimum logic feature size, and λ is the minimum metal half-pitch

4.3.2 Writability

MTJ switching is a current-induced phenomenon, and the switching operation requires a bidirectional control of current. For 1T-1J, the currents supplied to MTJ are not symmetrical with respect to the polarity of current, owing to the phenomenon known as the source degeneration effect. This occurs when a resistive load is placed at the source side of a transistor. As a consequence, despite the same operating voltage (V_{DD}) applied to BL or SL, the transistor output currents are asymmetrical. This is illustrated in Fig. 4.6, where such asymmetry is simulated at a full circuit level. This causes a significant disadvantage which reduces the write margin of 1T-1J. Furthermore, the STT effect on a typical MTJ is also asymmetrical, which is described by I_c asymmetry (β), defined as $\left|\frac{I_c^{P\to AP}}{I_c^{AP\to P}}\right|$. Typical MTJ devices exhibit β of 1.5 or larger, presumably, due to smaller STT effect for $R_p \to R_{ap}$ (electrons flowing from the free layer to the reference layer). When these two effects are coupled in a conventional 1T-1J bitcell, it is much more difficult to switch the cell from R_p to R_{ap}, often results in increase in transistor size or operation voltage. Several approaches have been suggested to mitigate these problems: (1) I_c asymmetry reduction using dual spin polarizers [20]; (2) a "top-pinned" MTJ film stack [21]; and (3) a modified 1T-1J with a reversely connected MTJ [3].

In most cases of STT-MRAM targeted for fast switching, a primary challenge is to design for the capability of supplying sufficiently large driving current for MTJ switching. A simple alternative to 1T-1J is 2T-1J, for which one MTJ is coupled with two access transistors in parallel. The drive current can become significantly larger. Despite the fact that an additional transistor makes the effective transistor size twice as large as that of 1T-1J, the bitcell size increases only by ~33 % to 16 λ^2, as shown in Fig. 4.7. This is realized through an optimized 2T-1J layout by sharing the source line between neighboring bitcells and therefore eliminating the spacing between the active regions of neighboring bitcells. Compared with the 1T-1J bitcell (Fig. 4.4b) whose bitcell height is often 1.5 times of the gate pitch, the height of the 2T-1J bitcell is increased to 2 times of the gate pitch, thereby increasing the bitcell size by ~33 %.

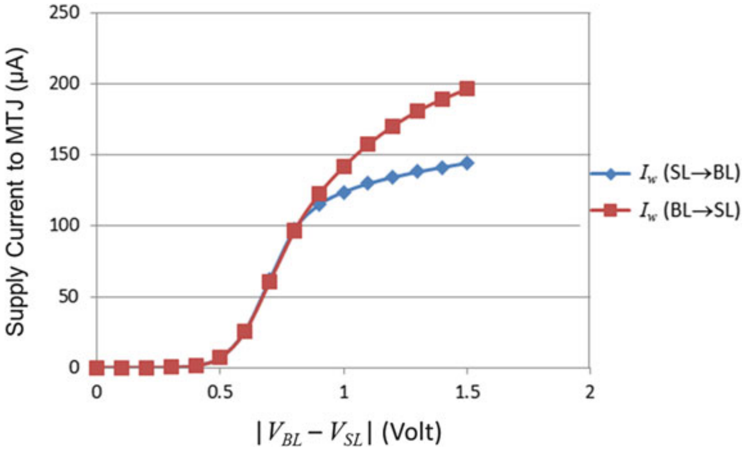
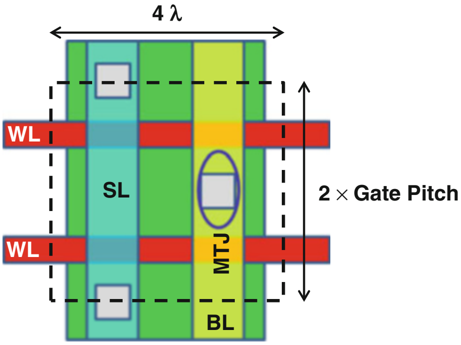


Fig. 4.6 Write supply current (I_w) simulated from a full STT-MRAM chip. Due to the source degeneration effect, I_w is asymmetrical with respect to the polarity of the current

Fig. 4.7 A 2T-1J bitcell layout. While the effective transistor width is doubled versus 1T-1J of Fig. 4.4b, the bitcell size is only ∼33 % larger



In STT-MRAM, I_c has a strong dependence on write pulse width, as illustrated in Fig. 4.8. Fast MTJ switching, often referred to as precessional switching (∼10 ns or below), requires substantially larger I_c than relatively slow switching. This leads to challenges in designing high-performance bitcells. Unless the MTJ size is substantially small, it is often difficult to realize sub-10 ns switching without enlarging the bitcell size. This is a primary reason why continuing innovations in MTJ materials engineering are still desired to reduce J_c. Recent advances have realized reliable switching in 1T-1J below 4 ns with write error rate lower than 10^{-6} [6].

Practically, it is necessary to tailor MTJ and bitcell attributes for varying write speed requirements depending on different STT-MRAM product applications. For example, for embedded Level 2 or Level 3 CPU cache memory, the MTJ switching speed is preferred to be on the order of a few nanoseconds, although this could often be relaxed significantly through various design optimization techniques. In contrast, for traditional embedded nonvolatile memory applications, the switching speed on the order of a microsecond is still compelling (a few orders of magnitudes faster than embedded Flash). An advantage of STT-MRAM is such that MTJ can be tuned for custom bitcells which can serve widely varying ranges of product applications.

4.4 Circuit Design

4.4.1 Write Circuit

The write operation of STT-MRAM is to switch the state of MTJ by supplying current higher than I_c. The polarity of the current determines the switched state, either 0 (R_p) or 1 (R_{ap}). As shown in Fig. 4.9 [22], a write driver is connected to BL and SL, respectively, which acts as a current source or a sink depending on the current polarity. Each write driver is realized by a tri-state inverter. The magnitude of the write supply current (I_w) is determined by the size of the write driver. To write 0, the current flows from the free layer to the pinned layer of MTJ, so that the write driver of BL operates as a current source and that of SL as a current sink. Accordingly, the D value of the write driver (Fig. 4.9c) is high for BL and low for SL. As shown in Fig. 4.8, I_c is a function of write pulse width. I_c becomes higher as the pulse width is shorter. Thus a write enable signal (WET, WEB) should be controlled precisely to prevent write failure (occurring when $I_w < I_c$). On the other hand, a wear-out reliability risk may arise when I_w is too high. Therefore, designing a write driver must consider two factors: precise control of write pulse width and optimal sizing of the driver.

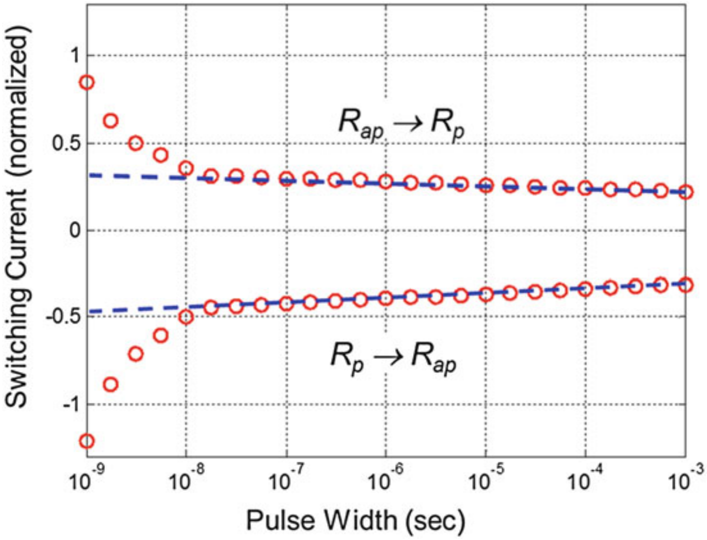


Fig. 4.8 An example characteristic of switching current (I_c) as a function of write pulse (I_w) width

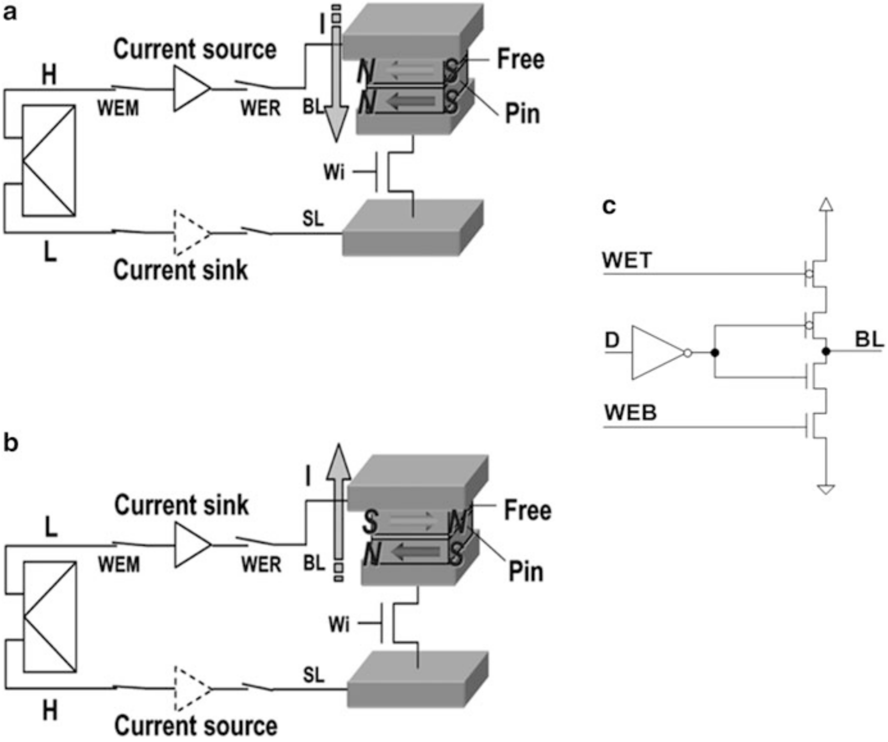


Fig. 4.9 Illustration of STT-MRAM write operation [22]: (**a**) 1 → 0; (**b**) 0 → 1; (**c**) a schematic of a tri-state write driver

Both I_w and I_c are dependent on process variation and can be modeled by Gaussian distributions. For a single cell, the write access pass yield (WAPY), expressed in sigma (standard deviation), is obtained by combining the distributions of I_w and I_c

$$WAPY_{Cell} = \frac{\mu_{I_w} - \mu_{I_c}}{\sqrt{\sigma_{I_w}^2 + \sigma_{I_c}^2}} \tag{4.7}$$

where μ_{I_W} and μ_{I_C} are the mean of I_w and I_c, respectively, and σ_w and σ_c are the standard deviation of I_w and I_c,respectively.

4.4.2 Read Circuit

4.4.2.1 Conventional Sensing Circuit

The read operation of STT-MRAM determines the resistance state of each cell with respect to the predefined state of a reference MTJ array. The operation relies on a sensing circuit and a sense amplifier which converts an output voltage of the sensing circuit to a digital signal. Figure 4.10 shows a conventional sensing circuit designed for MRAM [23]. The circuit is comprised of a data branch and two reference branches. Each branch includes a clamp NMOS (NC_D or NC_R) and a load PMOS (PL_D or PL_R). The sensing current (I_s) is controlled by the gate voltage of clamp NMOS (V_{G_clamp}). The clamp NMOS generates different currents according to the MTJ state 0 or 1. The source voltage of clamp NMOS is fixed in a saturation region. The saturation current of clamp NMOS is high at 0 and low at 1. The two clamp NMOS of the reference branches (NC_R) are designed to generate a saturation current at a medium level between 0 and 1 of the data branch, as shown in Fig. 4.11. The saturation current of NC_R is conveyed to PL_D through a current mirror circuit. Thus,

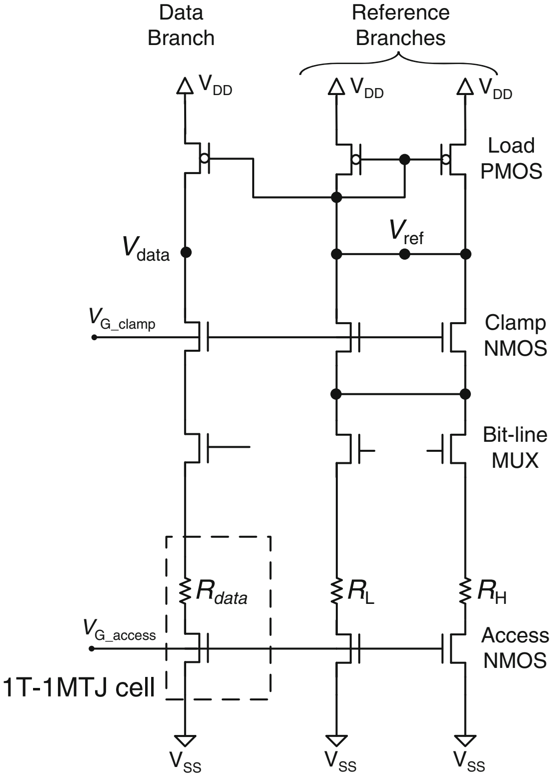


Fig. 4.10 Schematic illustration of a conventional sensing circuit for MRAM [23]

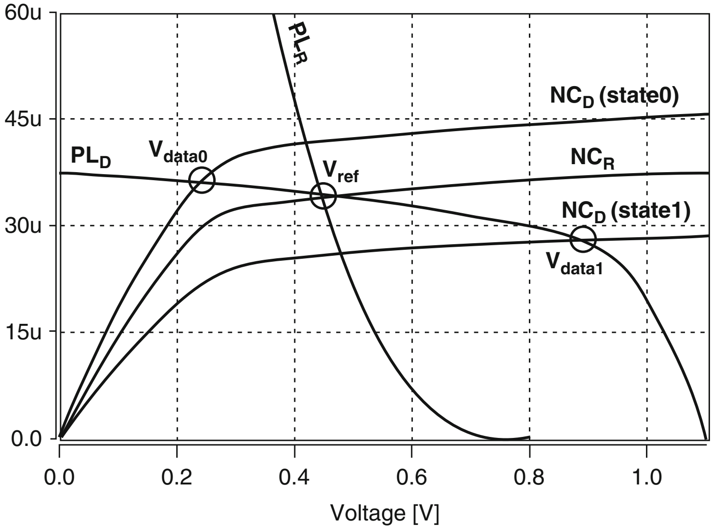


Fig. 4.11 I–V characteristics of clamp NMOS and load PMOS in a conventional sensing circuit

the saturation current of NC_D is larger than that of PL_D for 0 and accordingly V_{data0} is low, whereas the saturation current of NC_D is smaller than that of PL_D for 1 and V_{data1} is high.

The sense amplifier determines 0 or 1 of the data MTJ by comparing output voltages (V_{data0}, V_{data1}, V_{ref}) of the sensing circuit. Read is successful when the difference between V_{data} and V_{ref} ($\Delta V_0 = V_{ref} - V_{data0}$, $\Delta V_1 = V_{data1} - V_{ref}$) is larger than the offset voltage of the sense amplifier (V_{SA_OS}). Note that V_{data} is susceptible to PVT variations. Thus, it is important to design V_{ref} in a way to trace PVT variations of V_{data}.

4.4.2.2 Read Yield

A circuit designer needs to prevent two types of functional failure during read operation. Sensing failure occurs when ΔV_0 or ΔV_1 is smaller than V_{SA_OS}. Read disturbance failure is possible when I_s exceeds I_c (i.e. unintentional switching during sensing). Considering these two, a statistical read yield model can be built in the following way.

The statistical distributions of ΔV_0, ΔV_1, and V_{SA_OS} can be modeled by Gaussian distributions. For a single cell, the read access pass yield for 0 or 1 (RAPY$_{\mathrm{Cell0}}$ or RAPY$_{\mathrm{Cell1}}$) is given by [24]

$$RAPY_{Cell0,1} = \frac{\mu_{\Delta V_{0,1}} - \mu_{SAOS}}{\sqrt{\sigma^2_{\Delta V_{0,1}} + \sigma^2_{SAOS}}} \tag{4.8}$$

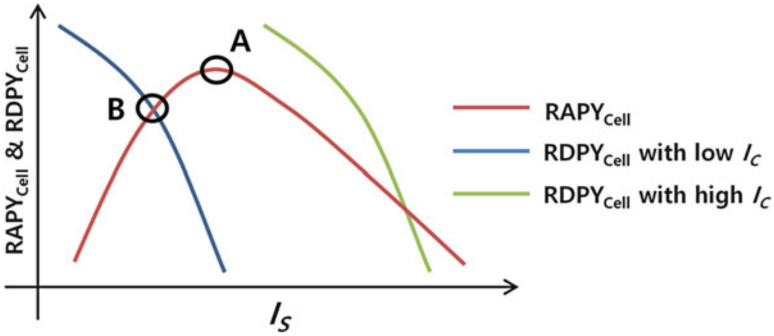


Fig. 4.12 Read yield (RAPY$_{Cell}$ and RDPY$_{Cell}$) trends as a function of I_s and I_c

where $\mu_{\Delta V_{0,1}}$ and μ_{SA_OS} are the mean of $\Delta V_{0,1}$ and V_{SA_OS}, respectively, and $\sigma_{\Delta V_{0,1}}$ and σ_{SA_OS} are the standard deviation of $\Delta V_{0,1}$ and V_{SA_OS}, respectively. RAPY$_{Cell}$ is then defined as the smaller of RAPY$_{Cell0}$ and RAPY$_{Cell1}$.

$$RAPY_{Cell} = \min\left(RAPY_{Cell0},\ RAPY_{Cell1}\right) \tag{4.9}$$

The criterion for read disturbance failure is $I_s \geq I_c$. Thus, the read disturbance pass yield (RDPY) is given by:

$$RDPY_{Cell} = \frac{\mu_{I_c} - \mu_{I_s}}{\sqrt{\sigma_{I_c}^2 - \sigma_{I_s}^2}} \tag{4.10}$$

I_s has large effects both on RAPY and on RDPY. As illustrated in Fig. 4.12, RDPY$_{Cell}$ decreases as I_s increases. In addition, RDPY$_{Cell}$ is lower when I_c is lower. On the other hand, there is an optimum I_s which maximizes RAPY$_{Cell}$ because it is difficult to achieve small $\sigma_{\Delta V_{0,1}}$ and large $\mu_{\Delta V_{0,1}}$ when I_s is too low and too high, respectively. Therefore, depending on I_c, different design strategies are applicable to maximize read yield. For high I_c, RDPY$_{Cell}$ is also high, so that I_s is tuned to maximize RAPY$_{Cell}$ (Point A). For low I_c, it is desired to find I_s to make RDPY$_{Cell}$ and RAPY$_{Cell}$ equal (Point B). In general, I_c continually scales down as the feature size shrinks, which means that controlling read disturb yield is becoming of great significance.

4.4.3 Advanced Sensing Circuits

Assuring adequate sensing margin (ΔV_0 and ΔV_1) for STT-MRAM at deeply scaled nodes necessitates extensive design efforts owing to the decrease in supply voltage and the increase in process variation. Further, STT-MRAM must be designed to avoid potential read disturbance, desiring a low-current sensing method. To solve these challenges, various types of advanced sensing circuits have been developed.

4.4.3.1 Source Degeneration PMOS

The load PMOS and the clamp NMOS shown in Fig. 4.10 can become a significant source of process variation for the sensing circuit. The clamp NMOS has a large source resistance, and it operates as a source degeneration resistance to keep the current through the clamp NMOS as constant as possible. On the other hand, the source of the load PMOS is directly connected to a voltage supply (V_{DD}), leading to large current variation. To reduce the variation effect of the load PMOS, a source degeneration scheme can be adopted by inserting a degeneration PMOS between the source of the load PMOS and the voltage supply, as shown in Fig. 4.13 [24]. This is relatively a simple method to increase read yield by minimizing the variation of ΔV_0 and ΔV_1 caused by the process variation of the load PMOS.

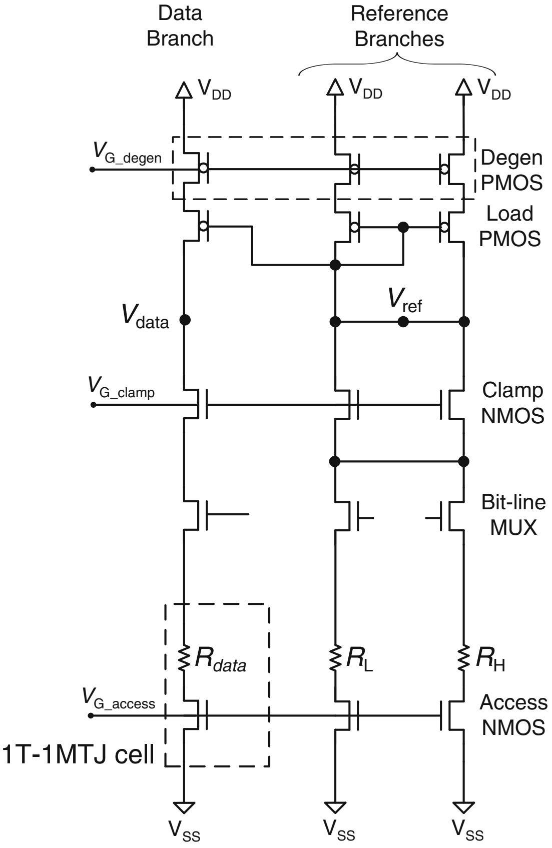


Fig. 4.13 A sensing circuit that employs degeneration PMOS [24]

4.4.3.2 Self Body Biasing

Assuring sensing margin (ΔV_0 and ΔV_1) is more challenging at lower V_{DD} and higher V_{TH} (threshold voltage), i.e., when the voltage headroom (V_{DD}–V_{TH}) is smaller. Note that high V_{TH} transistors are widely adopted for low standby power applications. Figure 4.14 describes a sensing circuit that can mitigate this challenge by utilizing self body biasing [25]. The body bias can decrease V_{TH} of the load PMOS when the sensing circuit is active. This helps secure the sensing margin during read operation while not causing high leakage current at the standby mode. In addition, it internally generates a body voltage without utilizing a body voltage generator, so that the area overhead is minimal compared with the conventional sensing circuit (Fig. 4.10). This scheme can be coupled with the degeneration PMOS described in Sect. 4.4.3.1.

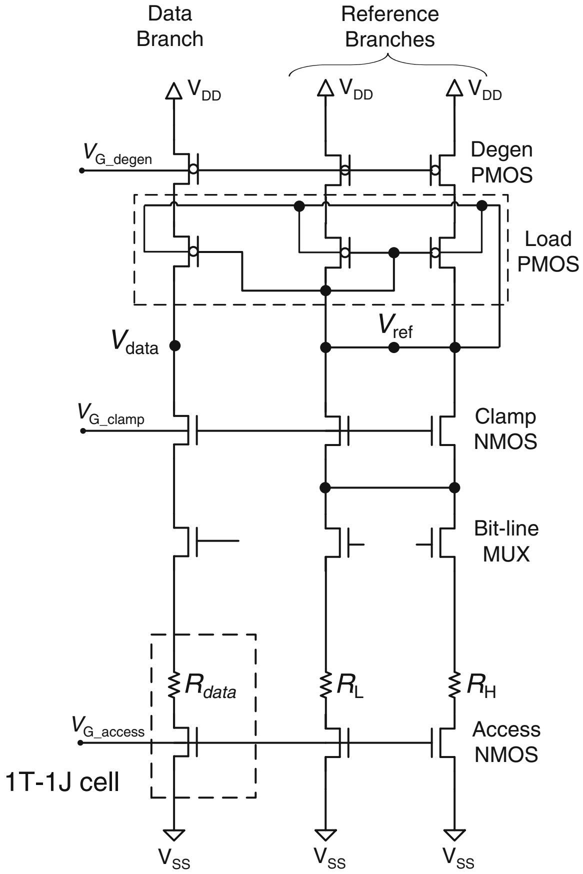


Fig. 4.14 A sensing circuit that employs self-body biasing [25]

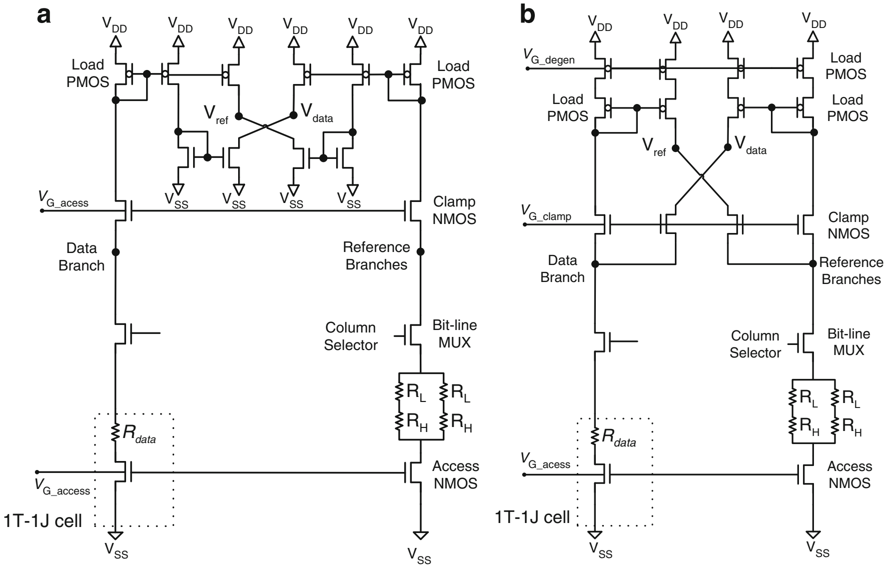


Fig. 4.15 (**a**) A sensing circuit with a symmetric cross-coupled current mirror; (**b**) a split-path sensing circuit [26]

4.4.3.3 Split-Path Sensing

A sensing circuit can adopts variable V_{ref} to increase ΔV_0 and ΔV_1. V_{ref} is modulated according to the MTJ state. At 0, V_{ref} increases, so does ΔV_0. Whereas at 1, V_{ref} decreases, hence ΔV_1 increases. Figure 4.15a shows a conventional sensing circuit with variable V_{ref} with a symmetric cross-coupled current mirror. A drawback of this circuit is that the mismatch occurred during such current mirroring increases the standard deviation of ΔV_0 and ΔV_1. An alternative scheme has been proposed by adopting a split-path sensing circuit, shown in Fig. 4.15b [26]. The split path enables variable V_{ref} while minimizing the number of current mirrors. The variable V_{ref} enhances $\mu_{\Delta V_{0,1}}$ by doubling it. The minimized number of current mirrors reduces $\sigma_{\Delta V_{0,1}}$.

4.4.3.4 Offset-Canceling Triple-Stage Sensing

One of emerging challenges for STT-MRAM sensing circuits is to reduce I_s while maintaining sensing margins. This is attributed to rapid reduction in I_c with p-MTJ (owing to high STT efficiency addressed in Sect. 4.2) and also with a demand for smaller MTJ (e.g. diameter <20 nm). Reduced I_s results in larger $\sigma_{\Delta V_{0,1}}$, which directly reduces sensing yields. The impact of $\sigma_{\Delta V_{0,1}}$ is often greater than $\mu_{\Delta V_{0,1}}$. Figure 4.16 illustrates an offset-canceling triple-stage (OCTS) sensing circuit [27].

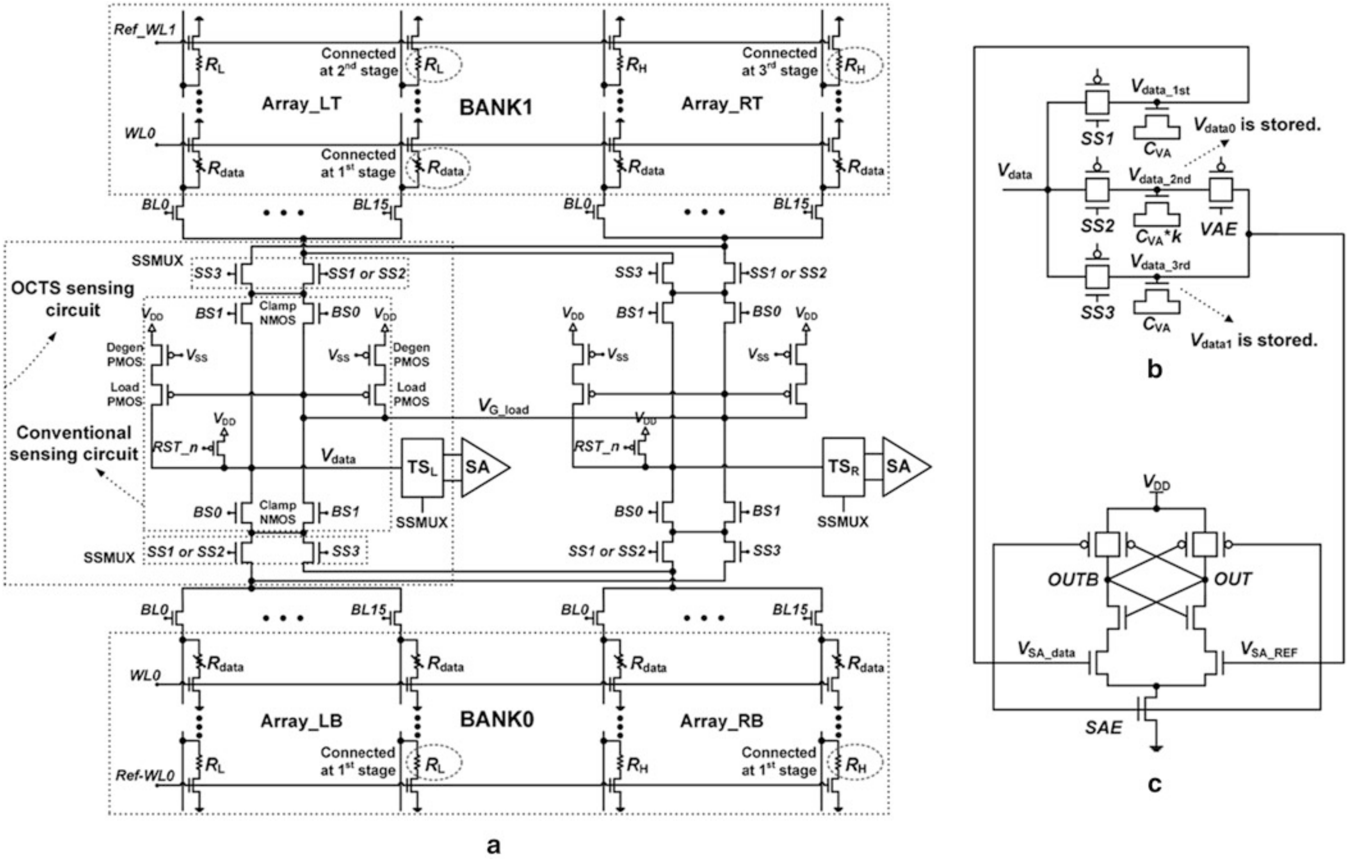


Fig. 4.16 An offset-canceling triple-stage (OCTS) sensing circuit. Diagrams of (**a**) an OCTS sensing circuit in a simplified array architecture; (**b**) the TS_L block of (**a**); (**c**) the SA block of (**a**) [27]

This is designed for reducing $\sigma_{\Delta V_{0,1}}$ by canceling the offsets of the sensing circuit caused by process variations. The principle of OCTS is to sense progressively three cells that are the data cell and the two reference cells with 0 and 1 through one sensing circuit. The output voltages are then added or subtracted to cancel out the offsets.

A drawback of OCTS is such that there is only one sensing circuit sequentially to read data and reference cells. Hence it is difficult to avoid a read speed penalty, though this may be tolerable for most applications. In addition, gate capacitors are required to store the sensed value at each stage, which increases the array size.

4.4.3.5 Self-Reference Sensing

OCTS can effectively cancel out the offsets of the sensing circuit, but cannot improve the sensing yield related to MTJ process variation. Self-reference circuits, shown in Fig. 4.17, generate V_0, V_1, and V_{ref} with only one MTJ. Such sensing circuits minimize the offsets caused not only by the sensing circuit but also by MTJ process variation. Figure 4.17a is a relatively simple self-reference scheme. It reads the MTJ cell and store V_0 and V_1 in capacitance at the first stage. At the second stage, the MTJ cell is written to 0 by a larger write current than that of the first stage. The sensing circuit reads this cell to generate V_{ref}. Because the stored information of the MTJ is removed during the read operation, this method is destructive, so that the readout value must be written back to the MTJ at the last step. This degrades the read speed and consumes more energy. In contrast, an alternative scheme, which is nondestructive, is shown in Fig. 4.17b [28]. This scheme allows maintaining the MTJ state after the first sensing. When the MTJ is at 0, the resistance of the cell is nearly constant regardless of the current flowing through the MTJ cell. The resistance change is detected by current when the MTJ turns into 1. Accordingly, this nondestructive self-reference scheme generates V_{ref} without write and write-back processes, overcoming the drawbacks of the circuit in Fig. 4.17a. However, it is difficult to secure the resistance difference when the current difference is subtle, which may cause a challenge for ever decreasing operating current requirement with scaling.

4.4.4 *Array Architecture*

In general, memory array architecture is an essential design parameter that influences performance, power consumption, yield, and chip size. Determining an optimal array architecture is therefore dependent on bitcell specification, chip specification, target yield, and even reliability. The array efficiency, a ratio of bitcell array area over total memory area including peripheral circuits, becomes higher as the array size increases. But, this leads to degradation in performance because parasitic resistances and capacitances increase owing to the increase in the number

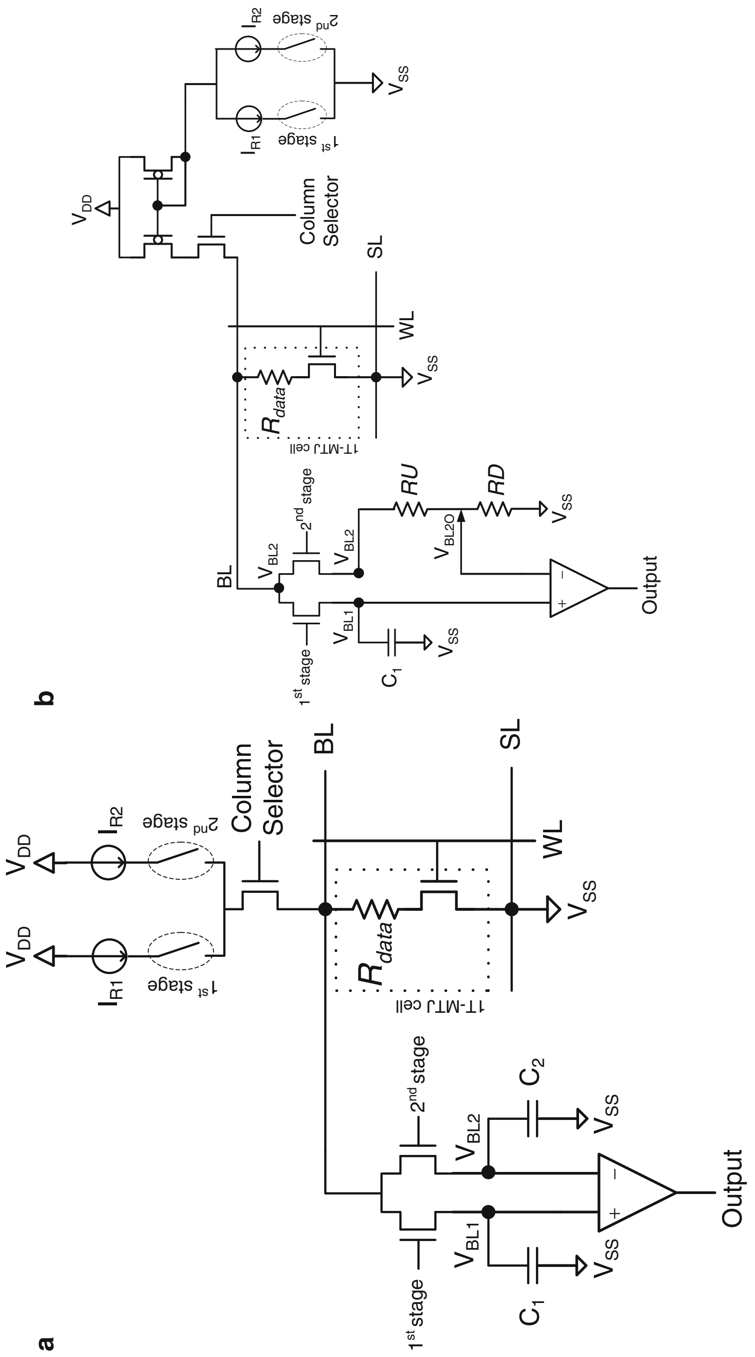


Fig. 4.17 Self-reference sensing circuits: (**a**) conventional scheme; (**b**) non-destructive scheme [28]

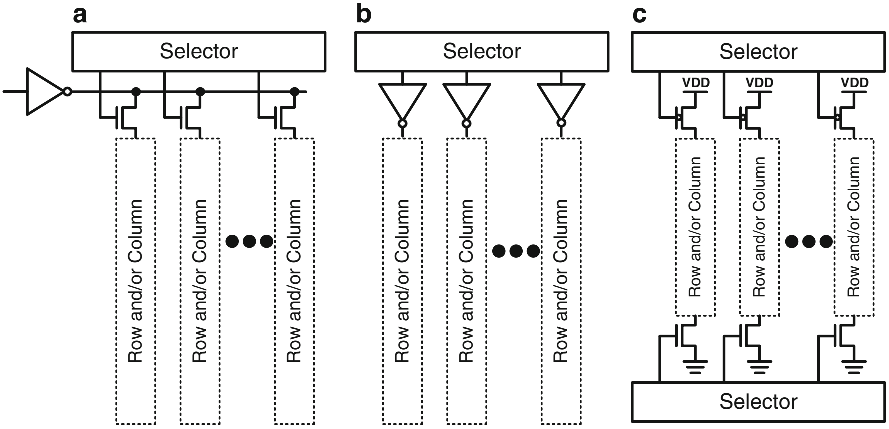


Fig. 4.18 Multiplexer architectures: (**a**) single write driver; (**b**) merged writer drivers; (**c**) separated write driver

of cells connected to BL. As shown above, read circuit design is more challenging for STT-MRAM, so that an effort to increase the array size must be considered cautiously in a way not to degrade read performance.

4.4.4.1 Multiplexer (MUX) Architecture

STT-MRAM employs a different MUX structure compared with conventional memories. For the read operation of STT-MRAM, selected BL is connected to the sensing circuit and SL to the ground (*GND*). For the write operation of 0, the write driver connected to selected BL drives V_{DD} and the driver connected to SL drives *GND*, and vice versa for 1. As shown in Fig. 4.18, these read and write operations can be enabled by different types of MUX structures. Figure 4.18a is a common architecture to achieve a small footprint since only one write driver is required for all BL and SL. However, the current originated from the write driver must pass through the MUX, hence, reducing I_w. Figure 4.18b shows a merged write driver for which selection control utilizes an independent write driver for each BL and SL. This results in a larger area, but provides higher I_w owing to the absence of MUX along the write path. Figure 4.18c illustrates selectors to drive V_{DD} and *GND* from both terminals of BL and SL by separating a write driver. The area becomes even larger, however, this MUX can realize higher yield by controlling parasitic mismatches because the lengths of BL and SL are the same for all the cells.

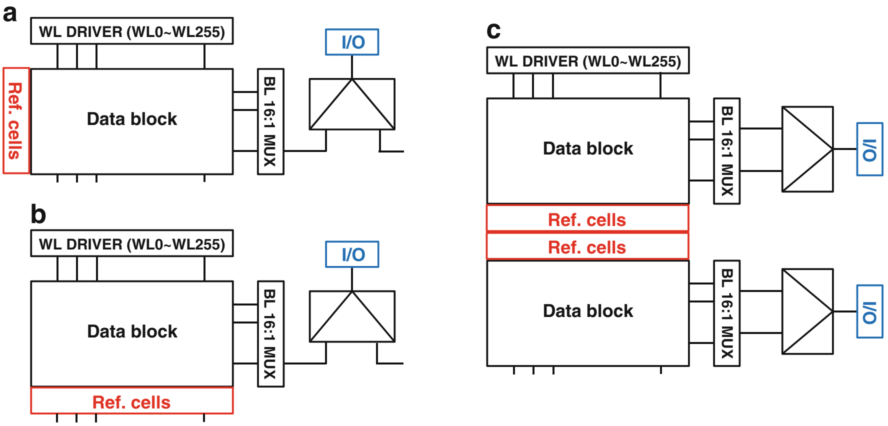


Fig. 4.19 Reference cell array architectures: (**a**) row reference; (**b**) column reference; (**c**) shared column reference

4.4.4.2 Reference Cell Architecture

STT-MRAM reference cell architectures are generally categorized according to the position of reference cells, as shown in Fig. 4.19. An array structure which places all reference cells into one WL (Fig. 4.19a) has an advantage in achieving small footprint. But, this architecture is not preferable for high-capacity STT-MRAM owing to parasitic mismatches between reference cells and data cells. In comparison, the reference architecture in Fig. 4.19b, which places all reference cells along one BL, has an advantage of achieving higher read yield. There is an area penalty associated with this, though tolerable. While each data block ordinarily has its own reference cell array, it is also possible to design a shared reference scheme for which two data blocks share two adjacent reference cells (Fig. 4.19c).

4.5 Co-design of MTJ and Logic

Designing an IC at an advanced technology node with embedded STT-MRAM requires a proven statistical circuit model which addresses systematic and random variations of MTJ and logic [29]. It is important to understand the challenges imposed by deep scaling of the logic technology. Recent work addressed a first-of-its-kind statistical circuit model and its application for designing a STT-MRAM building block and its array [30]. Figure 4.20 illustrates key components of this co-design methodology. The model can be seamlessly integrated into a common CMOS circuit simulation environment. The statistical variability-aware model fits Si data by covering PVT variations. The model is also combined with a micromagnetic physical model, hence, allowing co-optimization of MTJ physical parameters and

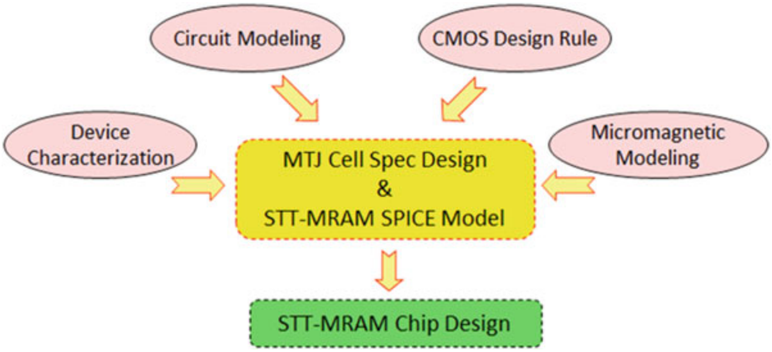


Fig. 4.20 Conceptual illustration of key components of a CMOS-MTJ co-design methodology [30]

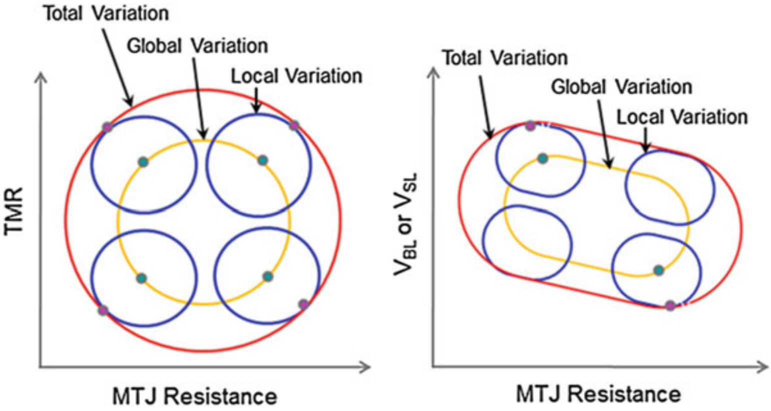


Fig. 4.21 Illustration of STT-MRAM statistical variation models which correlate key device parameters such as *R*, TMR, and V_c [29, 30]

cell circuit parameters. Hence, the model is applied to tune device parameter specifications required to meet target chip performance, yield, and reliability. Correlations among critical design parameters are systematically examined by the model. An example is shown in Fig. 4.21. By performing statistical Monte Carlo simulations, the model is capable of predicting an array functional yield. The model correlates various functional failure modes to physical cell defects or circuit-design errors. The accuracy of the model has been validated by chip-level functionality and yield data [31].

4.6 Perspective

Modern SOC memory subsystems are diverse or complicated, so that difficult to be served by one prevalent memory. It is desirable to optimize each SOC platform by a different combination of memory attributes such as speed, power consumption, reliability, and cost. In this aspect, STT-MRAM is attractively positioned since its building block MTJ can be tuned for a broad range of memory attributes which can serve largely different types of SOC applications. For example, low-power STT-MRAM can become an ideal embedded nonvolatile memory for battery-powered wireless connectivity networks pertaining to Internet-of-Things and wearable electronics, not only by storing nonvolatile codes, but also by storing and executing fast data [32, 33]. This type of STT-MRAM simplifies the conventional memory subsystem and also extends battery life. In addition, its logic-friendly design and process compatibility can realize such benefits at advanced logic nodes for which it is difficult to employ conventional embedded nonvolatile memory technology. On the other hand, high-performance ($<\sim 5$ ns) STT-MRAM can serve as an alternative to SRAM. Despite the fact that STT-MRAM is slower than SRAM at a discrete circuit level, the memory subsystem can be architected in a way that the performance can be comparable or even better at a system level. Furthermore, there is a significant range of custom SRAM for which its leakage power and cost (chip area) are critical drawbacks. One emerging case is Level-3 cache for mobile CPU. Moreover, even higher performance STT-MRAM potentially realized in custom-designed bitcells and circuits may move up to a higher level of memory hierarchy (Level-2 cache). In addition, high-throughput embedded STT-MRAM can become an attractive memory for GPU by providing higher on-chip memory density at lower energy consumption. Finally, the MTJ applied for STT-MRAM can be utilized for security and anti-tampering applications. Examples include one-time programmable memory, random number generator, and physically unclonable function.

References

1. Kang SH, Lee K. Emerging materials and devices in spintronic integrated circuits for energy-smart mobile computing and connectivity. Acta Mater. 2013;61:952–73.
2. Hosomi M, Yamagishi H, Yamamoto T, et al. A novel nonvolatile memory with spin torque transfer magnetization switching: spin-RAM. IEDM Tech Dig. 2005;2005:459–62.
3. Lin CJ, Kang SH, Wang YJ, et al. 45 nm low power CMOS logic compatible embedded STT MRAM utilizing a reverse-connection 1T/1MTJ cell. IEDM Tech Dig. 2009;2009:279–82.
4. Ikeda S, Miura K, Yamamoto H, et al. A perpendicular-anisotropy CoFeB-MgO magnetic tunnel junction. Nat Mater. 2010;9:721–4.
5. Rizzo ND, Houssameddine D, Janesky J, et al. A fully functional 64 Mb DDR3 ST-MRAM built on 90 nm CMOS technology. IEEE Trans Magn. 2013;49(7):4441–6.
6. Thomas L, Jan G, Zhu J, et al. Perpendicular STT-MRAM with high spin-torque efficiency and thermal stability for embedded memory applications. J Appl Phys. 2014;155(17):172615
7. Sekikawa M, Kiyoyama K, Hasegawa H, et al. A novel SPRAM-based reconfigurable logic block for 3D-stacked reconfigurable spin processor. IEDM Tech Dig. 2008;2008:1–3.
8. Ohno H. A hybrid CMOS/magnetic tunnel junction approach for nonvolatile integrated circuits. In: VLSI Technology Symposium, 2009, p. 122–23.

9. Ohno H, Endoh T, Hanyu T, et al. Magnetic tunnel junction for nonvolatile CMOS logic. IEDM Tech Dig. 2010;2010:9.4.1–4.
10. Ando K. Nonvolatile magnetic memory. J Fed. 2001;12:89–95.
11. Ando K, Ikegawa S, Abe K, et al. Roles of non-volatile devices in future computer systems: normally-off computers. In: Energy-aware systems and networking for sustainable initiatives. Hershey: IGI Global; 2012. p. 83–907.
12. Kawahara T. Scalable spin-transfer torque RAM technology for normally-off computing. IEEE Des Test Comput. 2011;28(1):52–63.
13. Jullière M. Tunneling between ferromagnetic films. Phys Lett A. 1975;54(3):225–6.
14. Jaffrès H, Lacour D, Nguyen Van Dau F, et al. Angular dependence of the tunnel magnetoresistance in transition-metal-based junctions. Phys Rev B. 2001;64:064427.
15. Ikeda S, Hayakawa J, Ashizawa Y, et al. Tunnel magnetoresistance of 604% at 300 K by suppression of Ta diffusion in CoFeB/MgO/CoFeB pseudo-spin-valves annealed at high temperature. Appl Phys Lett. 2008;93:082508.
16. Choi YS, Tsunematsu H, Yamagata S, et al. Novel stack structure of magnetic tunnel junction with MgO tunnel barrier prepared by oxidation methods: preferred grain growth promotion seed layers and bi-layered pinned layer. Jpn J Appl Phys. 2009;48:120214.
17. Maehara H, Nishimura K, Nagamine Y, et al. Tunnel magnetoresistance above 170% and resistance–area product of 1 Ω-μm^2 attained by in-situ annealing of ultra-thin MgO tunnel barrier. Appl Phys Express. 2011;4:033002.
18. Slonczewski JC. Current-driven excitation of magnetic multilayers. J Magn Magn Mater. 1996;159:L1–7.
19. Berger. Emission of spin waves by a magnetic multilayer traversed by a current. Phys Rev B. 1996;54:9353–8.
20. Diao Z, Panchula A, Ding Y, et al. Spin transfer switching in dual MgO magnetic tunnel junctions. Appl Phys Lett. 2007;90:132508.
21. Lee YM, Yoshida C, Tsunoda K et al. Highly scalable STT-MRAM with MTJs of top-pinned structure in 1T/1MTJ cell. In: VLSI Technology Symposium, 2010, p. 49–50.
22. Kawahara T. 2 Mb SPRAM with bit-by-bit bi-directional current write and parallelizing-direction current read. IEEE J Solid-State Circuits. 2008;43(1):109.
23. Maffitt TM. Design considerations for MRAM. IBM J Res Dev. 2006;50(1):25.
24. Kim J, Ryu K, Kang SH, et al. A novel sensing circuit for deep submicron spin transfer torque MRAM. IEEE Trans Very Large Scale Integr Syst. 2012;20(1):181–6.
25. Kim J, Ryu K, Kim JP et al. An STT-MRAM sensing circuit with self-body biasing in deep submicron technologies. IEEE Trans Very Large Scale Integr Syst. 2014;22(7):1630-4 doi:10.1109/TVLSI.2013.2272587.
26. Kim J, Na T, Kim JP. A split-path sensing circuit for spin-torque transfer MRAM. IEEE Trans Circuits Syst, 2014. doi:10.1109/TCSII.2013.2296136.
27. Na T, Kim J, Kim JP et al. An offset-canceling triple-stage sensing circuit for deep submicrometer STT-RAM. IEEE Trans Very Large Scale Integr Syst. 2014;22(7):1620-4. doi:10.1109/TVLSI.2013.2294095.
28. Chen Y. A nondestructive self-reference scheme for spin-transfer torque random access memory. In: Design, Automation and Test in Europe Conference and Exhibition (DATE), 8–12 March 2010, p. 148–53.
29. Zhu X, Kang SH. Spin-transfer-torque MRAM: device architecture and modeling. In: Wang X, editor. Metallic spintronics devices. CRC, 2014 p. 21–70.
30. Zhu X, Kang SH. Variation-aware device modeling and design for embedded STT-MRAM array. In: 55th MMM Conference HC-13, 2010
31. Kim JP, Kim T, Hao W et al. A 45nm 1Mb embedded STT-MRAM with design techniques to minimize read-disturbance. In: VLSI Circuits Symposium, 2011, p. 296–97.
32. Kang SH. Embedded STT-MRAM for energy-efficient and cost-effective mobile systems. In: VLSI Technology Symposium, 2014, p. 36–7.
33. Lee K, Kan JJ, Kang SH. Unified embedded non-volatile memory for emerging mobile markets. In: ISLPED, 2014, p. 131–6.

Chapter 5
A Thermal and Process Variation Aware MTJ Switching Model and Its Applications in Soft Error Analysis

Peiyuan Wang, Enes Eken, Wei Zhang, Rajiv Joshi, Rouwaida Kanj, and Yiran Chen

Abstract Spin-transfer torque random access memory (STT-RAM) has recently gained increased attention from circuit design and architecture societies. Although STT-RAM offers a good combination of small cell size, nanosecond access time and non-volatility for embedded memory applications, the reliability of STT-RAM is severely impacted by device variations and environmental disturbances. In this work, we develop a compact switching model for magnetic tunneling junction (MTJ), which is the data storage device in STT-RAM cells. By leveraging the capability to simulate the impact of thermal and process variations on MTJ switching, our model is able to analyze the diverse mechanisms of STT-RAM write operation failures. Besides the impacts of thermal and process variation, the soft error induced by radiation striking on the access transistor is another important threat to the MTJ reliability. It can also be analyzed by using our model. The incurred computation cost of our model is much less than the conventional macro-magnetic model, and hence, enabling its applications in comprehensive STT-RAM reliability analysis and design optimizations.

P. Wang • E. Eken • Y. Chen (✉)
Electrical and Computer Engineering Department, University of Pittsburgh,
Pittsburgh, PA 15261, USA
e-mail: wap15@pitt.edu; ene4@pitt.edu; yic52@pitt.edu

W. Zhang
Division of Computing System, School of Computer Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 639798, Singapore
e-mail: zhangwei@ntu.edu.sg

R. Joshi
IBM T.J. Watson Lab, Yorktown Heights, NY, USA
e-mail: rvjoshi@us.ibm.com

R. Kanj
Department of ECE, American University of Beirut, New York, NY, USA
e-mail: rk105@aub.edu.lb

R.O. Topaloglu (ed.), *More than Moore Technologies for Next Generation Computer Design*, DOI 10.1007/978-1-4939-2163-8_5

5.1 Introduction

The recent progress on the research of emerging nonvolatile memory (NVM) technology has attracted increased attention from circuit design and architecture societies. As one promising candidate for the future universal memory technology, spin-transfer torque random access memory (STT-RAM) offers a good combination of high cell density, nanosecond access and non-volatility [1]. Compared to the conventional memory technologies with electrically charge-based data storage mechanism, the magnetic storage mechanism of STT-RAM has better technology scalability and immunity to radiation-induced soft errors [2]. The applications of STT-RAM have been successfully demonstrated in embedded memories and reconfigurable systems [3], etc. However, similar to all nanoscale technologies, STT-RAM severely suffers from process variations when the fabrication technology enters deca-nanometer era. The STT-RAM reliability is degraded by the increased device parametric variability and the intrinsic randomness during circuit operations, i.e. thermal fluctuation.

In a STT-RAM cell, data is stored as the different resistance states of a magnetic tunneling junction (MTJ) device. '0' (low resistance) or '1' (high resistance) can be written into the MTJ by passing through a switching current with different polarizations. The MTJ switching time is directly determined by the magnitude of the MTJ switching current: increasing the MTJ switching current leads to a decrease in MTJ switching time [4]. Besides the driving ability variation of the MOS transistor connected to the MTJ, there are two major types of variations affecting the MTJ switching performance: the MTJ device parametric variability and the thermal fluctuation. The first one incurs the deviation of the magnetization switching process from the nominal one, while the second one introduces the randomness in the write process. Many studies have been performed to analyze the impact of these variations on the reliability of STT-RAM operations: Li et al. summarized the major MTJ parametric variations affecting the resistance switching and proposed a "2T1J" STT-RAM design for yield enhancement [5]. Nigam et al. developed a thermal noise model to evaluate the thermal fluctuations during the MTJ resistance switching process [6]. Joshi et al. conducted a quantitative statistical analysis on the combined impact of both CMOS/MTJ device variations and thermal fluctuations [7]. However, these works either do not construct a complete model to characterize all the variation types and their interactions, or require costly Monte-Carlo simulations with complex macro-magnetic and SPICE models.

In this work, we develop a compact MTJ switching model that is derived from the MTJ macro-magnetic modeling. The statistical electrical properties of the MTJ, i.e., the resistance variations, switching transience and switching time, can be simulated with the minimized run time cost. Experimental results show that our model is able to accurately simulate the write operation errors incurred by the device variations and/or the thermal fluctuations, or the intermittent switching current disturbance during the MTJ switching process in STT-RAM designs.

Note that radiation-induced soft error is another common threat to the electrically-charged device reliability. Although the intrinsic magnetic storage property of MTJ is immune to the radiation, the switching current of the MTJ is supplied by the MOS transistor, which is still at the risk of radiation: a soft error occurring at the MOS transistor during the write operations will cause a disturbance on the MTJ switching current. The spin-torque induced magnetization switching of the MTJ is then changed by such disturbances, leading to a delay or even a failure of write operations. However, this scenario is ignored in many research on STT-RAM analysis and designs, even in some radiation-hardness memory/circuit designs. In this work, we also comprehensively analyze the impact of the radiation on the robustness of STT-RAM, e.g., a soft error occurring at the NMOS transistor, by using the proposed model. The design methods for the enhancement of the write operation robustness of STT-RAM cell, e.g., transistor sizing, are also discussed.

The rest of this chapter is organized as the follows: Section 5.2 introduces the basics of STT-RAM technology; Sect. 5.3 describes the development details of our variation aware compact MTJ switching model; Sect. 5.4 validates our model by comparing with the macro-magnetic simulation results; Sect. 5.5 demonstrates the applications of our model in analyzing the effects of the process variations and thermal fluctuations on the MTJ switching performance; Sect. 5.6 presents the quantitative analysis on the radiation impacts on STT-RAM write operations, and possible design methods for radiation hardness optimization.

5.2 STT-RAM Basics

Figure 5.1a shows the structure of an MTJ: An oxide barrier layer, e.g., MgO, is grown between two ferromagnetic layers [1]. The MTJ resistance is determined by the relative magnetization directions of the two ferromagnetic layers: when their magnetization directions are anti-parallel (parallel), the MTJ is in high-(low-) resistance state. Usually the magnetization direction of one ferromagnetic layer (called "reference layer") is fixed by coupling to a pinned magnetization layer while the magnetization direction of the other ferromagnetic layer (called "free layer") can be changed by passing a polarized switching current: when the switching current passes through the MTJ from reference layer to free layer, the MTJ switches to high resistance state (see Fig. 5.1a); If the switching current passes through the MTJ from the other direction, MTJ switches to low resistance state (see Fig. 5.1b).

Figure 5.1c shows the popular one-transistor-one-MTJ (1T1J) STT-RAM cell design. The polarization of the MTJ switching current is controlled by the voltages applied to the bit-line (BL) and the source-line (SL). The word-line (WL) is used to control the gate of the NMOS transistor and select the STT-RAM cell.

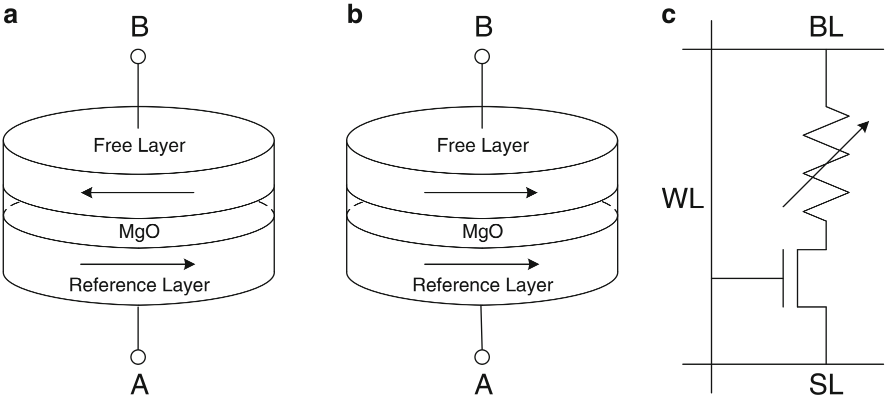


Fig. 5.1 MTJ structure. (**a**) Anti-parallel (high resistance). (**b**) Parallel (low resistance). (**c**) 1T1J STT-RAM cell structure

5.3 Compact MTJ Switching Model

In this section, we illustrate the development process of our thermal and process variation aware compact MTJ switching model which is derived from MTJ macro-magnetic modeling.

5.3.1 Model Descriptions

Because the magnetization direction of the reference layer is fixed, the magnetization switching process in an MTJ can be described by the direction change of the free layer magnetization M_f. In general, the magnetization is assumed to be a constant in magnitude. Its motion is represented by a unit direction vector $m_f = M_f / M_s$, where M_s is the saturation magnetization. At any instant of time, vector m_f makes an angle θ with the e_z axis while the plane of M_f makes an angle φ with the e_x axis, as shown in Fig. 5.2.

In other words, the motion of M_f can be uniquely described by the coordinates (θ, φ). The energy landscape experienced by M_f includes: (1) uniaxial anisotropy; (2) easy-plane anisotropy; (3) the applied magnetic field; and (4) Langevin random field torque.

During the STT-RAM operation of writing ‘1’, a spin current is injected to the magnetic body along $-e_x$ direction. We assume the spin-polarization factor is η. The spin direction is in the e_y–e_z plane, making an angle ϕ with e_z axis. Here we ignore the current-generated magnetic field by assuming spin-current effect is dominant in a small magnetic body, which is the normal working condition of the MTJ in a STT-RAM cell. The potential energy for M_f is $U_1(\theta, \varphi) = U_K + U_p$. Here

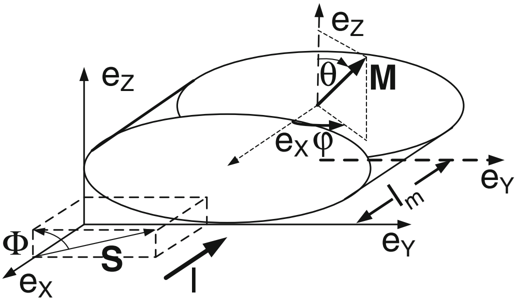


Fig. 5.2 Definition of MTJ magnetization

$U_K = K sin^2\,\theta$ is uniaxial anisotropy energy with $K = (1/2) M_f H_k$. H_k is the Stoner–Wohlfarth switching field. U_p is easy-plane anisotropy in the e_y–e_z plane while its normal direction is e_x. The environmental energy is $U_2(\theta,\ \varphi) = U_H + U_L$, where U_H is the applied field in the easy plane of e_y–e_z, making an angle of ψ with the easy axis e_z. U_L is the Langevin random field related to the temperature. The torque experienced by M_f within the free layer under the energy landscape [8] can be written as:

$$\frac{\Gamma_U}{l_m} = -m_f \times \nabla \left[U_1(\theta,\varphi) + U_2(\theta,\varphi)\right]. \tag{5.1}$$

Here Γ_U is the torque, l_m is the thickness of free layer. The potential and environmental energy introduces the four components of m_f. The first one is the uniaxial anisotropy as:

$$\frac{\Gamma_1}{l_m K} = (2sin\theta cos\theta)\left[(sin\varphi)\,e_x - (cos\varphi)\,e_y\right]. \tag{5.2}$$

The rest three terms are all normalized to the uniaxial anisotropy term. The normalized second term—easy-plane anisotropy is:

$$\frac{\Gamma_2}{l_m K} = -2h_p\left[(cos\theta sin\theta cos\varphi)\,e_y - \left(cos\varphi sin\varphi sin^2\theta\right)e_z\right]. \tag{5.3}$$

The normalized third term—applied magnetic field is:

$$\begin{aligned}\frac{\Gamma_3}{l_m K} = 2h\Big[&(sin\varphi cos\psi sin\theta - cos\theta sin\psi)\,e_x \\ &- (cos\varphi cos\psi sin\theta)\,e_y + (cos\varphi sin\theta sin\psi)\,e_z\Big].\end{aligned} \tag{5.4}$$

The normalized fourth term—Langevin random field is [9]:

$$\begin{aligned}\frac{\Gamma_4}{l_m K} = &\left(h_{L,z} sin\theta sin\varphi - h_{L,y} cos\theta\right)e_x + h_{L,x} cos\theta e_y \\ &- h_{L,z} sin\theta cos\varphi e_y + \left(h_{L,y} sin\theta cos\varphi - h_{L,x} sin\theta sin\varphi\right)e_z.\end{aligned} \tag{5.5}$$

The torque generated by the spin current can be expressed in vector form as:

$$\Gamma_5 = s m_f \times \left(m_s \times m_f\right), \tag{5.6a}$$

where $s = (\hbar/2e)\eta J$ is the spin-angular momentum deposition per unit time [8]. Here $\hbar$ is Planck constant, e is elementary charge. Similar to the components of m_f, the normalized torque can be expressed as:

$$\begin{aligned}\frac{\Gamma_5}{l_m K} = 2h_s\Big\{ &-(sin\theta cos\varphi)\,(sin\theta sin\varphi sin\phi + cos\theta cos\phi)\, e_x \\ &+ \left[cos\theta\,(sin\phi cos\theta - cos\phi sin\theta sin\varphi) + sin^2\theta cos^2\varphi sin\phi\right] e_y \\ &+ \left[sin\theta\,(sin\theta cos\phi - sin\varphi sin\phi cos\theta)\right] e_z\Big\}.\end{aligned} \tag{5.6b}$$

The dynamics of M_f can be simulated by Landau–Lifshitz–Gilbert (LLG) equation as:

$$\frac{dm_f}{dt} + \alpha\left(m_f \times \frac{dm_f}{dt}\right) = \frac{1}{2}\Omega_k \sum_{i=1}^{5}\left(\frac{\Gamma_i}{l_m K}\right), \tag{5.7}$$

where α is the Gilbert damping constant [10, 11]. By introducing a new time unit $= \frac{\Omega_k t}{1+\alpha^2}$, we obtain the ordinary differential equation of coordinates (θ, φ) to describe the motion of the magnetization vector as:

$$\begin{bmatrix}\theta' \\ \varphi'\end{bmatrix} = \sum_{i=1}^{5}\begin{bmatrix}\theta_i{}' \\ \varphi_i{}'\end{bmatrix}, \tag{5.8}$$

where the uniaxial anisotropy term is

$$\begin{bmatrix}\theta_1{}' \\ \varphi_1{}'\end{bmatrix} = -\begin{bmatrix}\alpha sin\theta cos\theta \\ cos\theta\end{bmatrix}. \tag{5.9a}$$

The easy-plane anisotropy term is

$$\begin{bmatrix}\theta_2{}' \\ \varphi_2{}'\end{bmatrix} = -h_p\begin{bmatrix}(sin\varphi + \alpha cos\theta cos\varphi)\, sin\theta cos\varphi \\ (cos\varphi cos\theta - \alpha sin\varphi)\, cos\varphi\end{bmatrix}. \tag{5.9b}$$

The applied field term is

$$\begin{bmatrix}\theta_3{}' \\ \varphi_3{}'\end{bmatrix} = -h\begin{bmatrix}\left\{\begin{matrix}cos\varphi sin\psi + \\ \alpha\,(sin\theta cos\psi - cos\theta sin\varphi sin\psi)\end{matrix}\right\} \\ \left\{\begin{bmatrix}(sin\theta cos\psi - cos\theta sin\varphi sin\psi) \\ -\alpha cos\varphi sin\psi\end{bmatrix}/sin\theta\right\}\end{bmatrix}. \tag{5.9c}$$

The Langevin random field term is

$$\begin{bmatrix} \theta_4{}' \\ \varphi_4{}' \end{bmatrix} = -h \begin{bmatrix} \left\{ \begin{matrix} \alpha \left[h_{L,z} sin\theta - cos\theta \left(h_{L,x} cos\varphi + h_{L,y} sin\varphi \right) \right] \\ + \left(h_{L,y} cos\varphi - h_{L,x} sin\varphi \right) \end{matrix} \right\} \\ \left\{ \begin{bmatrix} \alpha \left(h_{L,x} sin\varphi - h_{L,y} cos\varphi \right) + h_{L,z} sin\theta \\ - cos\theta \left(h_{L,x} cos\varphi + h_{L,y} sin\varphi \right) \end{bmatrix} / sin\theta \right\} \end{bmatrix}. \tag{5.9d}$$

And the effective spin torque term is

$$\begin{bmatrix} \theta_5{}' \\ \varphi_5{}' \end{bmatrix} = -h_s \begin{bmatrix} \alpha cos\varphi sin\phi + sin\varphi sin\phi cos\theta - cos\phi sin\theta \\ sin\phi \left(cos\varphi - \alpha sin\varphi cos\theta \right) / sin\theta + \alpha cos\phi \end{bmatrix}. \tag{5.9e}$$

By combing Eqs. (5.8) and (5.9a)–(5.9e), we obtain an analytical model with only two ordinary differential equations to simulate the magnetization changing process of an MTJ over time.

Aforementioned, all four components of m_f and the spin torque term are normalized by the Stoner–Wohlfarth uniaxial-anisotropy field H_k. For example, the easy-plane anisotropy field h_p, which is used to emulate the thin-film demagnetization field $4\pi M_s$ [12, 13], is defined as:

$$h_p = 4\pi M_s / H_k. \tag{5.10}$$

In our model, applied magnetic field term h is assumed to be zero. As the term representing the thermal noise in the MTJ switching, Langevin random field is affected by the environment temperature and device geometry size as [6]:

$$h_{L,i} = \sqrt{\frac{2\alpha k_B T}{\gamma M_S V}} \mathrm{X_i}\left(\mathrm{t}\right). \; \left(i = x, \; y, \; z\right). \tag{5.11}$$

Here T is the environment temperature, V is the geometry volume of the free layer, α is the damping constant, γ is the gyro-magnetic ratio, k_B is the Boltzmann constant. $X_i(t)$ is a Gaussian random noise with zero mean and unit variance in x, y, and z axis.

Finally, the normalized effective spin torque h_s is:

$$h_s = \frac{s}{l_m M_f H_k}. \tag{5.12}$$

The definitions of all parameters are listed in Table 5.1.

The value of θ determines the transience of MTJ resistance R_{MTJ} during the switching process as [8]:

$$R_{MTJ}(t) = R_0 / \left(1 + p^2 cos\ \theta\ (t)\right), \tag{5.13}$$

Table 5.1 Summary of parameter definitions

Parameter	Definition
1	Uniaxial-anisotropy field
h_p	Easy-plane anisotropy field
h_L	Langevin random field
h	Applied magnetic field
h_s	Effective spin current
α	Gilbert damping constant

where R_0 is the original resistance of the MTJ and p is the effective spin polarization of the ferromagnetic electrodes. The relationship between the commonly-used tunneling magnetic resistance ratio (TMR) and p is:

$$TMR = \frac{2p^2}{1 - p^2}. \tag{5.14}$$

The high (R_H) and the low (R_L) resistance state of the MTJ equal $R_0/\left(1 - p^2\right)$ and $R_0/\left(1 + p^2\right)$, respectively.

5.3.2 Impact of Parametric Variations

The major geometry variations that affect the magnetization switching process of MTJ include the variations of MTJ shape, MTJ surface area (A) and the thickness of free layer (l_m).

The MTJ shape variations incur the change of the *demagnetizing factor*, which determines the *demagnetization field*. The calculation of demagnetizing factor requires a costly Finite Element Analysis (FEA) in macro-magnetic modeling. Compared to other parameters' variation effect on switching performance, the shape induced demagnetization factor change is relatively small. In our compact model, we ignore the shape-variation-induced demagnetizing factor change and assume it is a constant for an elliptical MTJ cell in all simulations to achieve a fast computation speed.

Besides *demagnetization field*, the geometry variations of MTJ mainly affect the Langevin random field h_L in Eq. (5.11) and the effective spin torque h_s in Eq. (5.12). Based on Eq. (5.11), h_L is inversely proportional to the square root of the MTJ free layer volume V, which equals the product of the MTJ surface area A and the free layer thickness l_m. Hence, it can be represented as $h_L \propto \frac{1}{\sqrt{A \times l_m}}$. It indicates that a large MTJ surface area and/or a thick free layer can help to minimize the variation of MTJ switching performance incurred by thermal fluctuations. Based on Eq. (5.12), h_s is proportional to the MTJ switching current density J and inversely proportional to l_m. If an MTJ is driven by a constant switching current I, the switching current density is given as $J = I/A$. Therefore, h_s is proportional to $\frac{1}{A \times l_m}$. It means that the efficiency of the spin current for MTJ switching is higher in a small size MTJ than in a large one.

In addition, in a "1T1J" STT-RAM cell design, the current through the MTJ is supplied by the NMOS transistor connected to the MTJ. The variation of MTJ resistance may change the bias condition of the NMOS transistor and consequently, affect the actual MTJ switching current. The R_0 in Eq. (5.13) is inversely proportional to the MTJ surface area A and exponentially proportional to the oxide layer thickness l_{ox} as [7]:

$$R_0 \propto \frac{e^{l_{ox}}}{A}.$$

5.3.3 Temperature Dependency

The working temperature increase of the MTJ mainly increases the magnitude of Langevin random field term, or $h_L \propto \sqrt{T}$ [see Eq. (5.11)]. It means that the MTJ becomes less stable and easier to switch, while suffering from a larger thermal noise at any instant of time [14]. We also note that the reduction of MTJ free layer volume V makes the variability of MTJ switching more sensitive to temperature changes: as shown in Eq. (5.11), the magnitude of h_L varies more at a smaller V when the temperature changes. To simplify the analysis, we ignore the second-order temperature dependency of these parameters in our compact model.

The thermal stability of an MTJ is usually measured by the magnetic memorization energy $\Delta = \frac{hM_S V}{k_B T}$. $\varDelta$ also determines the long-term data stability, or data retention time: an increase of T will decrease $\varDelta$ and lead to a shorter data retention time (or non-volatility degradation).

5.4 Model Validation and Test

During the derivation of our compact MTJ switching model in Sect. 5.3.1, we simplified some pre-conditions and ignored the side effects that are usually included in a complete macro-magnetic simulation. Some examples are: (1) we ignored the current-generated magnetic field; (2) we assumed a single-domain MTJ structure, which leads to the elimination of demagnetization between two magnetic domains; (3) we ignored the second-order temperature dependency of the parameters, etc. These simplifications lead to an analytical solution of the MTJ switching model, while also incur possible inaccuracy. Therefore, we need to validate our model first before applying it to the real designs. The device parameters of MTJ used in most of our simulations are cited from [15]. The mean and standard deviations of MTJ parameters are summarized in Table 5.2.

Table 5.2 Summary of parameters

Parameter	Mean (μ)	Std. Dev. (σ)	σ/μ
lm	2.3 nm	0.115 nm	5 %
A	3,600 nm^2	180 nm^2	5 %
R_H	2,000 Ω	–	–
R_L	1,000 Ω	–	–

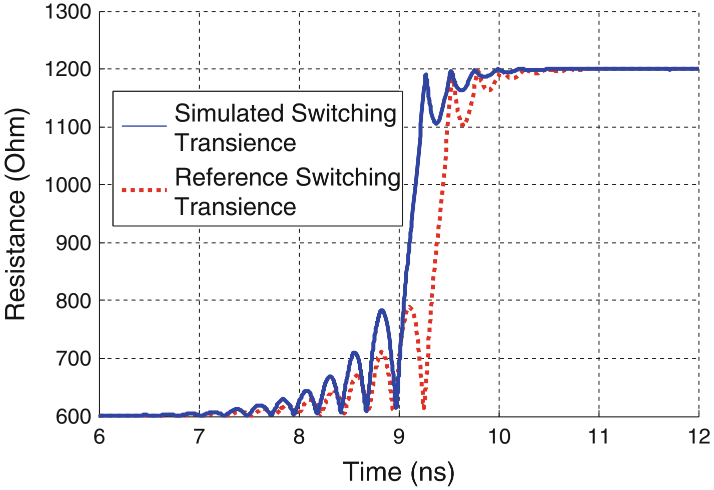


Fig. 5.3 Model validation on MTJ resistance switching transience

5.4.1 MTJ Switching Transience

In order to validate the MTJ resistance switching transience, we assume there are no external assisting magnetic fields under the normal working condition of STT-RAM cells and ignore the thermal fluctuation in our model, e.g., $h = 0$ and $h_L = 0$. A deterministic process of the magnetization switching of the MTJ free layer under a certain spin current magnitude can be obtained. Figure 5.3 shows the comparison between the corresponding result of the MTJ resistance switching transience using the same device parameter simulated by the published macro-magnetic models in [16] and the simulation result of our model.

Our model demonstrates a good agreement with the published results on the major features of the MTJ resistance switching curve, such as the damping rate and the oscillation frequency. Also, the start and the end of the transience are approximated very well by using our model.

Our compact model can also simulate the magnetic characteristics of MTJ, i.e., the precession of m_f in the free layer during the magnetization switching. Figure 5.4 shows the changes of m_f when the MTJ switches from the high-resistance state ‘1’ to the low-resistance state ‘0’ under the influence of spin current.

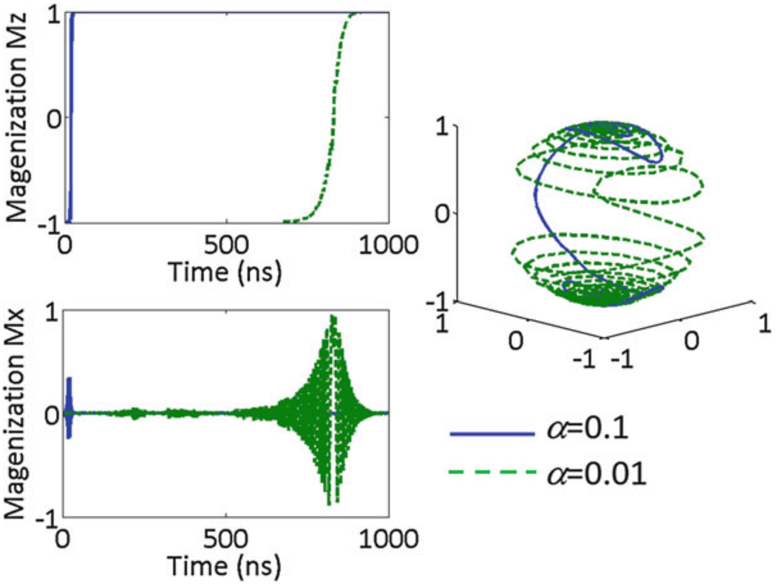


Fig. 5.4 Change of m_f during the MTJ resistance switching of 1→0

θ changes from π to 0 during the switching. The rotations of m_f vector generate the MTJ resistance oscillations in the switching transience as shown in Fig. 5.3. It can be seen that our compact model is capable of simulating both electrical and magnetic properties of the MTJ and provides a comprehensive solution for debugging of STT-RAM designs.

5.4.2 Switching Performance and Variations

As we have already known, the MTJ switching time T_{sw} is a function of the switching current I_{sw}. In general, the MTJ switching can be divided into three regions: (1) when $T_{sw} > 10$ ns, the switching mechanism is mainly a thermally activated process. The required I_{sw} increases slowly when T_{sw} decreases; (2) when $T_{sw} < 3$ ns, the switching mechanism is dominated by precessional switching. The required I_{sw} increases sharply when T_{sw} decreases; and (3) when 3 ns $< T_{sw} <$ 10 ns, a dynamic reversal that combines the precessional and thermally activated switching occurs and results in an intermediate state. Figure 5.5a shows the simulated the nominal values of T_{sw} at different I_{sw} using our model and the macro-magnetic model [2].

The device parameters in [2, 15] are used in both simulations. Figure 5.5 validates that in all three working regions, our model fits the macro-magnetic model very well.

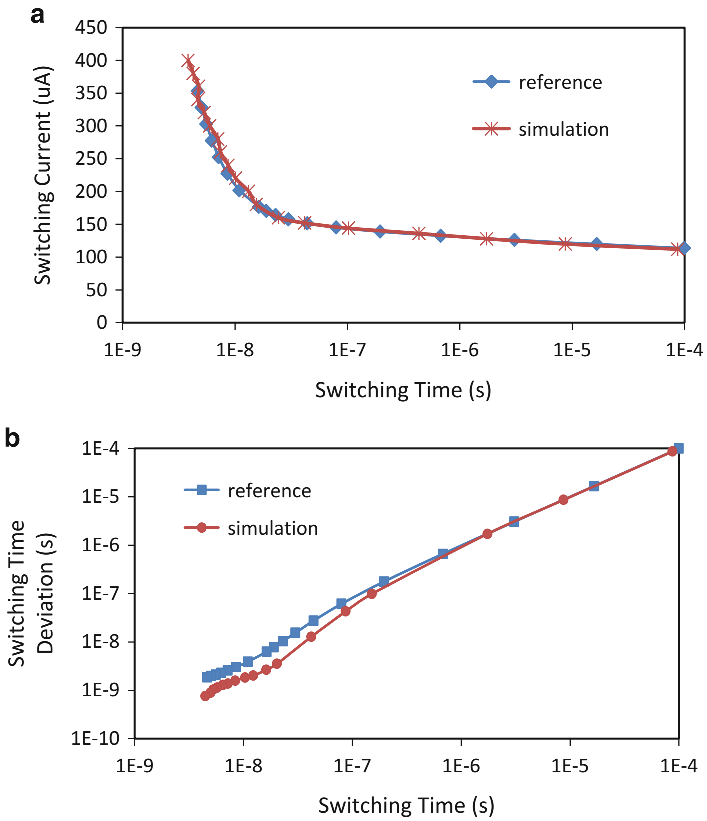


Fig. 5.5 Model validation on MTJ switching performance. (**a**) Nominal switching time vs. write current, (**b**) nominal switching time vs. switching time deviation

The MTJs working in different regions suffer from different types of thermal fluctuations: (1) when $T_{sw} > 10$ ns, the thermal fluctuation is dominated by the thermal component of internal energy, which can be modeled by a time-varying Langevin random field; (2) when $T_{sw} < 3$ ns, the thermal fluctuation is dominated by the thermally activated initial angle of precession, which is the accumulated effects of the Langevin random field over time before the write pulse applies; and (3) when 3 ns $< T_{sw} < 10$ ns, the thermal fluctuation is a mixed effect of the above two mechanisms. Figure 5.5b shows the simulated variance of T_{sw} at different switching current using our model and the macro-magnetic model under thermal fluctuations. It can be seen that our model agrees with the macro-magnetic model very well when MTJ is working in the long T_{sw} region. We note that in the short T_{sw} region, the simulated variation of MTJ switching time is more sensitive to the specific simulation environment since the same tolerance is set for a relatively wide range of switching current, e.g., the differential equation solver, the convergence

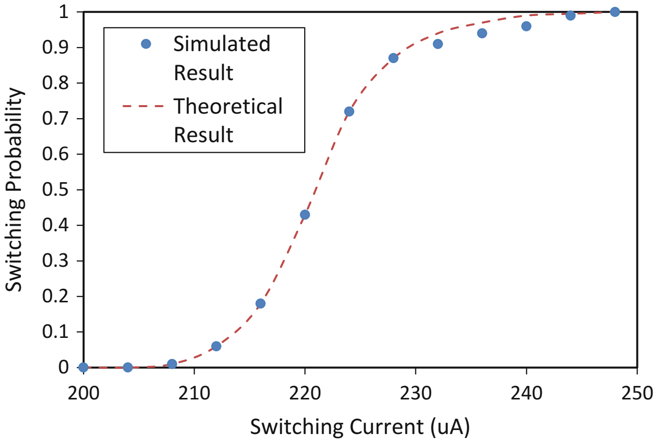


Fig. 5.6 Switching rate vs. switching current for a MTJ with a 45 × 90 nm ellipstical surface shape

threshold, and the Monte-Carlo simulation setup, etc. We believe the discrepancy between our model and the macro-magnetic model with the decrease of T_{sw} comes from the increased error accumulation in the differential equation solvers.

We also simulate the switching rate (Pr_{SW}) of a MTJ with a 45 × 90 nm elliptical shape under different switching currents. A successful switching is defined as the completion of the MTJ switching before the switching current is removed. In our simulations, the MTJ switching failures are mainly caused by the thermal-induced switching time variations. As shown in Fig. 5.6, our results fit very well with the heuristic equation [17]:

$$Pr_{\mathrm{F}} = 1 - exp\left\{-\frac{t}{\tau}exp\left[-\Delta\left(1-\frac{I_{\mathrm{sw}}}{I_{\mathrm{C}}}\right)\right]\right\} \tag{5.16}$$

Here t is the applied switching current pulse width, which is 10 ns in our simulations. Δ is the magnetic memorizing energy. τ is the inverse of the attempt frequency; I_C is the critical switching current, which is the minimum current amplitude to switch the MTJ resistance under a write pulse width of t. From the validation against both heuristic equation and macro-magnetic model, the accuracy of our model is well proven.

5.4.3 System Level Performance

In standard STT-RAM cell structure consisting of one transistor and one MTJ, the cell area is determined by the NMOS transistor: MTJ size depends on only manufacturing resolution while NMOS transistor must be wide enough in order to

Table 5.3 STT-RAM cell specification

Technology	65 nm
Write pulse duration	10 ns
Threshold current	195 μA
Cell size	40 F^2
Aspect ratio	2.5

Table 5.4 Comparison of area, access time, and energy consumption

Cache size	128 KB SRAM	512 KB STT-RAM
Area	3.62 mm^2	3.30 mm^2
Read latency	2.252 ns	2.318 ns
Write latency	2.264 ns	11.024 ns
Read energy	0.895 nJ	0.858 nJ
Write energy	0.797 nJ	4.997 nJ

provide sufficient current for MTJ switching. For example, at 65 nm technology node, threshold value of the write current will be 195 μA. Based on the assumption of W/L ratio = 10 and the width of the source region = 1.5 F (F is the minimum feature size), the STT-RAM cell area is about 10 F × 4 F = 40 F^2 [3]. The parameters for the STT-RAM model are summarized in Table 5.3 [3].

Since a 128 KB SRAM (with 146 F^2 cell size) and a 512 KB STT-RAM have similar area, the comparison of these two in terms of area, access time and access energy is given in Table 5.4 [3].

As observed from Table 5.4, STT-RAM memory has long write latency and high write energy consumption compared to the SRAM based memory. The read latencies of two memory types are also very similar. Since the capacity of the STT-RAM is three times larger than SRAM with the same area, the access miss rates of the STT-RAM cache are reduced by averagely 19.0 % [3]. However, for workloads with high write intensity, because of the long write latency of STT-RAM, system-level performance degradation is 7.52 % with respect to the SRAM counterparts [3].

The power consumption in memories can be divided in two groups—leakage power and dynamic power. While the leakage power dominates the power consumption of SRAM at the scaled technology node, STT-RAM consumes very limited leakage power that is mainly from the peripheral circuit. As a result, dynamic power dominates the STT-RAM based cache power, which is only about 68 % of the overall power of the SRAM based implementation [3].

In order to solve the long latency and high dynamic energy consumption of the STT-RAM, Guangyu et al. [3] proposed read-preemtive write buffer technique which gives the read operation priority over the write operation and prevents the critical read operation from being blocked by write operations. By using this technique, they claim that the system performance can be improved up to 9.93 % with an average of 59.3 % power reduction.

5.5 Variability Analysis in STT-RAM

Like other memory designs at scaled technology nodes, Monte-Carlo simulations are usually required in STT-RAM designs to statistically analyze the variability incurred by process variations and thermal fluctuations. However, the computation cost of Monte-Carlo on macro-magnetic models may quickly become unaffordable as the design size and design confidence level increase. In contrast to it, our compact MTJ switching model offers a good basis for such statistical designs owing to single domain assumption with compact variation settings in the model. As we shall show in this section, our model can be used to simulate the impacts of the device variations and the temperature fluctuations with much less computation cost. The MTJ specifications for the simulation are summarized in Table 5.2.

5.5.1 *Process Variations and Switching Asymmetry*

The impacts of MTJ shape variations on MTJ switching process have been discussed in Sect. 5.3.2. In this section, we perform quantitative evaluation of these impacts on STT-RAM write reliability by using our compact model. Figure 5.7 shows the simulated MTJ switching transience at different 3σ design corners (XY) of surface area A and free layer thickness l_m.

Here X denotes the corner of A, when A is small/large, the switching of MTJ is fast(F)/slow(S). Similarly, Y denotes the corner of l_m, when l_m is thin/thick, switching is fast(F)/slow(S). The mean curve indicates the switching performance of fixed MTJ. Along the discussions in Sect. 5.3.2, a smaller MTJ cell with thinner free layer (FF corner) has the fastest switching time.

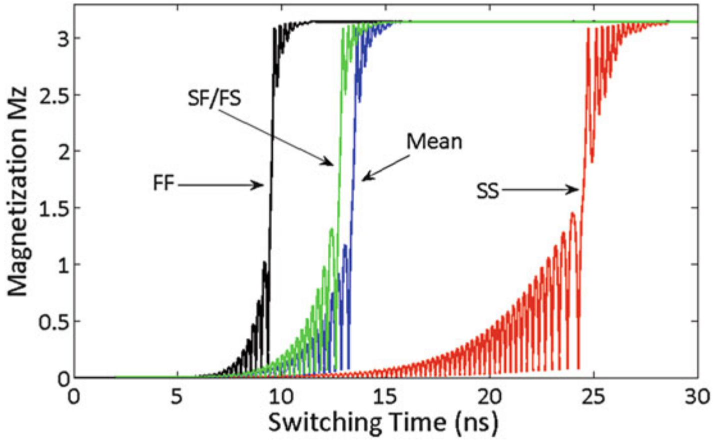


Fig. 5.7 MTJ switching transiences at different 3σ design corners

Table 5.5 MTJ write current and standard deviation under process variation

Transistor	‘0’ to ‘1’		‘1’ to ‘0’	
	I_{MTJ} (μA)	Std. Dev.	I_{MTJ} (μA)	Std. Dev.
180 nm	148.28	14.35	186.00	14.02
270 nm	194.75	18.11	263.03	15.64
360 nm	230.18	20.68	323.27	15.34
450 nm	258.18	22.76	362.77	17.15
540 nm	280.79	24.51	387.48	19.82
630 nm	299.91	26.15	404.43	21.96
720 nm	315.41	27.31	416.69	23.49

The switching time of a MTJ is also determined by the MTJ write current: when other device parameters remain constant, increasing the write current results in a reduced switching time. On the other hand, the magnitude of MTJ write current is determined by the NMOS transistor driving ability, which is mainly decided by the transistor size.

Our compact MTJ switching model can also be used to determine MTJ switching current once the required switching performance and the MTJ device parameters such as magnetization saturation, uniaxial anisotropy, free layer volume etc. are decided. After determining the switching current, the required size of the NMOS transistor can be obtained by using Table 5.5 [4].

To obtain the distributions of the MTJ write current in a 1T1J STT-RAM cell shown in Table 5.5, Monte-Carlo simulations are conducted [4]. The memory cell is designed with 45 nm PTM (predictive technology model). The device variations of both MTJ and NMOS transistors are considered in the simulations. Device parameters are already summarized in Table 5.2 and the simulated NMOS transistor size varies from 180 to 720 nm.

Placement scheme of the MTJ over the NMOS transistor and the layout of a STT-RAM cell is shown in Fig. 5.8. As can be seen in Fig. 5.8b, the MTJ has a relatively small area and the area of the STT-RAM cell is mainly determined by the NMOS transistor. Comparing with the conventional SRAM technology which has six transistors, the cell area of the STT-RAM cell is significantly reduced as only one transistor is used.

Table 5.5 also shows that MTJ retains some asymmetry in its switching process. Figure 5.9 shows the simulated MTJ switching times under different switching currents at both switching directions. Because of the asymmetry of the spin-polarization factor η at different directions, 0→1 switching requires a higher switching current than 1→0 switching for the same switching time. This result coincides with the observations of prior-art [7].

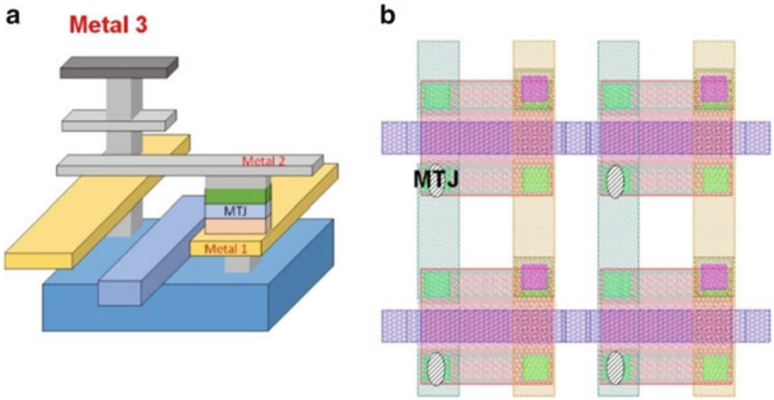


Fig. 5.8 STT-RAM cell. (**a**) 3D view and (**b**) layout design

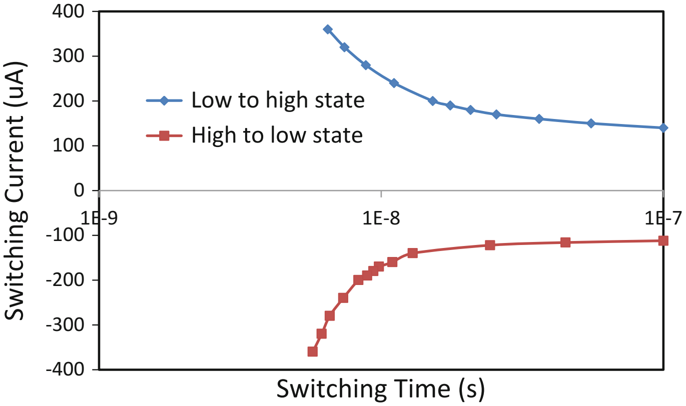


Fig. 5.9 Asymmetric switching performance of MTJ

5.5.2 *Temperature Variations*

To illustrate the impact of temperature, Figure 5.10a compares the nominal MTJ switching time (T_{sw}) at three temperatures (T): 275, 300 and 350 K, and Fig. 5.10b shows the ratios between the variance (σ) and the mean (μ) of T_{sw} at different temperatures. Raising T can improve the nominal T_{sw} because the MTJ becomes less stable and easier to switch. The temperature effect is small when the switching current is large due to the relatively strong spin-current field. It is interesting that following the increase in T, the ratio σ/μ significantly decreases when the MTJ switching current is low. It can be explained as the follows: In this long T_{sw} working region, the increase in T will reduce the MTJ switching time. At the same time, when temperature increases, the magnitude of thermal noise rises. However, since the impact of the thermal noise is an accumulative effect, the reduction in MTJ

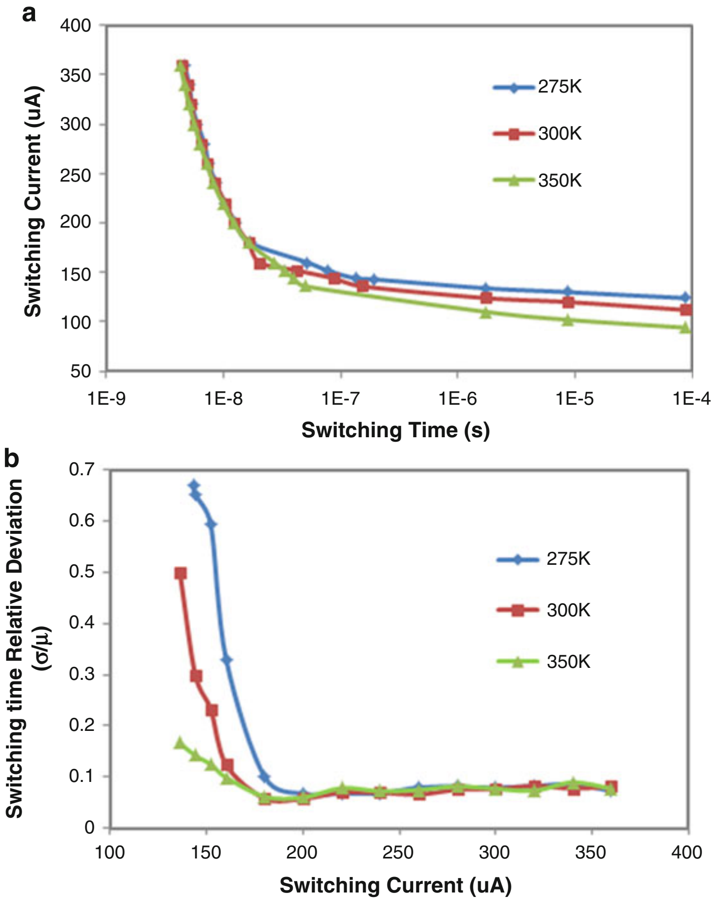


Fig. 5.10 MTJ switching time at different temperatures. (**a**) Nominal switching time vs. switching current. (**b**) Nominal switching time deviation vs. switching current

switching time in turn makes the thermal noise accumulated over the switching time reduce even further. Hence, σ decreases faster than μ, and a smaller ratio is observed in higher temperature. In other words, the impact of the thermal noise is reduced in magnetization switching process when temperature increases in the long MTJ working region.

Although for a constant switching current, MTJ switching time decreases when the temperature increases, increasing temperature has a negative impact on the driving ability of the NMOS transistor as electron mobility decreases accordingly [18]. Among the above two factors, the adverse effect of temperature on the driving ability of the NMOS transistor is more prominent than that on the MTJ switching time improvement. Thus, the required write pulse period will increase as the temperature gets higher [18].

When the STT-RAM based cache is placed over the CPU core using 3-D stacking technology, the unbalanced thermal effects across the die introduces different write pulse width requirement for different regions. In order to maintain an acceptable bit error rate for all region under temperature variations, a non-uniform write pulse width should be applied. Bi et al. [18] propose to place temperature sensors around the STT-RAM cache blocks. Before performing a write operation, different write pulse width is applied based on the temperature of the relevant region. By applying this technique, performance can be increased by 3.8 % and power consumption can be reduced by 4.8 %.

5.6 Application on Soft Error Analysis

Soft error is another important factor to be considered for device reliability, especially for space electronics. Under radiation effect, when enough charge is collected, the data stored as charging state may be flipped, causing function or data errors. For STT-RAM, the magnetic data storage mechanism makes it naturally resilient to radiation-induced soft errors after the data is written into the MTJ. However, the MTJ switching current is supplied by the connected NMOS transistor.

If a soft error occurs at the NMOS transistor during STT-RAM write operations, it generates a disturbance on the MTJ switching current and hence, affects the write process. To well describe the MTJ switching behavior under radiation, the transience of both MTJ magnetization switching and CMOS driving strength must be simulated together. In this section, we perform soft error analysis using our variation-aware compact model: Electron/hole pairs are generated by the strike of high-energy particles on the MOSFET device [19]. The electrons and holes move towards the opposite directions if there is an electric field between the source and the drain terminals. This movement generates an electrical charge noise, which leads to the random transience of MTJ switching current and consequently, affecting the MTJ switching performance. In the simulation setup, the STT-RAM is implemented with PTM 45 nm technology [20] and the MTJ is assumed in a 45×90 nm (short and long axis) elliptical shape.

5.6.1 Current Pulse Modeling

The shape of the noise generated by an ionizing particle on the MTJ switching current pulse is determined by the following factors: the exact geometry of the NMOS transistor's p-n junctions, the incident angle of the striking particle, and transistor doping profile, etc. In general, the form of the radiation-induced pulse of transistor driving current can be approximated as a double exponential shape as [21]:

$$I_{inject}(t) = I_{peak} \cdot \left(e^{(-t/\tau_a)} - e^{(-t/\tau_b)}\right). \tag{5.17}$$

Here $I_{peak} = Q/(\tau_a - \tau_b)$. Q represents the total collected electrical charge of the node. τ_a is the collection time-constant and τ_b is the ion-track establishment time-constant, respectively. The current $I_{inject}(t)$ charges or discharges the transistor gate and drain capacitors, generating a transient disturbance on the switching current or voltage on the MTJ. In transient analysis, a ramp approximation of the noise waveform may be used. The common approach is using a trapezoidal or triangular to model the noise generated at the electrical voltage or currentoutputs.

5.6.2 Radiation Effect Simulation

The cross section of a STT-RAM cell and the write paths are shown in Fig. 5.11a. At the onset of an ionizing radiation event, a cylindrical track of electron–hole pairs with a submicron radius and high carrier concentration forms in the wake of the energetic ion's passage. When the generated ionization track traverses or comes close to the depletion region, the electric field rapidly collects carriers, creating a current/voltage noise at that node. As shown in Fig. 5.11b, the radiation strike is modeled as a current source that can be easily incorporated in the STT-RAM cell circuit for fast simulations. Figure 5.12 shows the examples of the approximated undershoots of the MTJ switching current during the radiation attacks with different intensities. The corresponding MTJ resistance switching transience is also shown

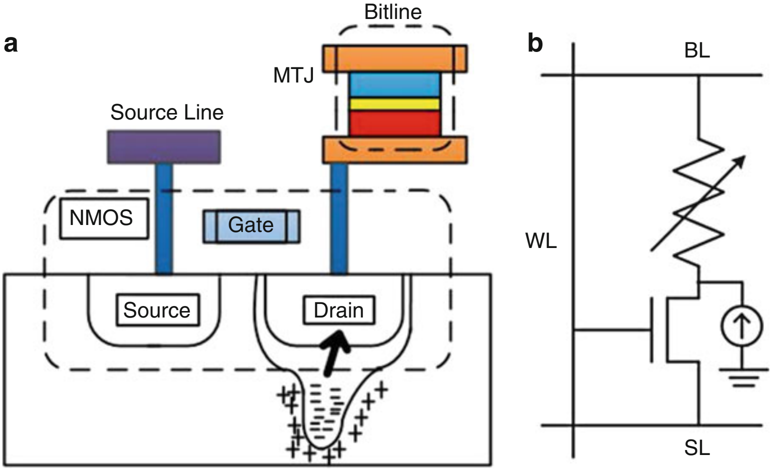


Fig. 5.11 STT-RAM cell with NMOS strike. (**a**) Cross section of the MTJ integration and ionization. (**b**) Equivalent circuit

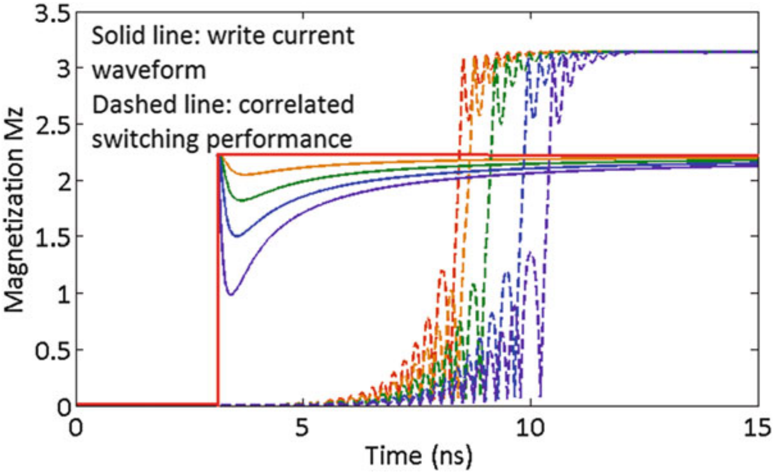


Fig. 5.12 MTJ switching transience under different strength of radiation strikes

in Fig. 5.12. Following the increase of the radiation-induce switching current noise magnitude and lasting time, the MTJ switching time is significantly delayed.

5.6.3 Radiation Impact Factors

There are three factors primarily determining the impacts of a single radiation on the reliability of MOS circuitry: the arrival time, the lasting period and the radiation intensity.

Obviously, a radiation can affect the STT-RAM cell only when it (and the incurred current noise) arrives within the switching window of the MTJ. Here the switching window is defined as the period between the beginning and the end of MTJ magnetization damping. As shown in Fig. 5.13, the MTJ switching time varies as the radiation arrival time changes.

If the radiation occurs in the middle of the MTJ switching window, the overall MTJ switching time slowly increases when the attack time approaches the end of the delayed MTJ switching. This result is expected because before the MTJ switching finishes, the magnetization state is the result of an accumulated field effect where the spin-current induced torque plays a major role in driving the magnetization into the other steady state. The interruption or disturbance of this process will delay the MTJ switching time. However, if the attack occurs around or after the end of MTJ switching, the original MTJ switching time will not be affected. As we can see in Fig. 5.13, the MTJ switching time goes back to its normal level in this scenario. Also, the increase of MTJ switching current results in a shorter switching process during which the impact of a strike is minimized. The noise on the spin-current

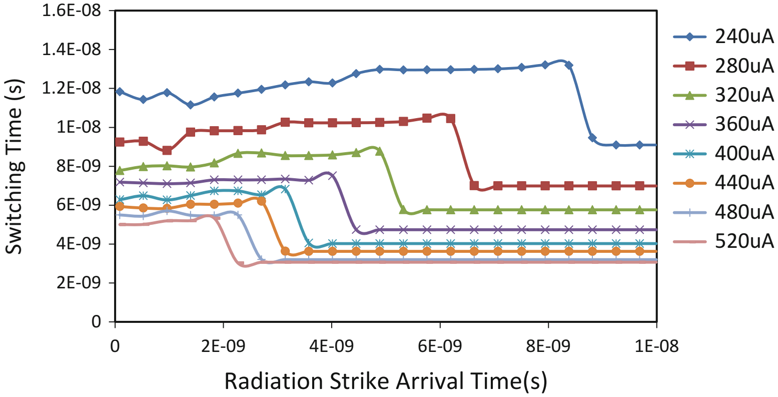


Fig. 5.13 MTJ switching time vs. radiation strike arrival time

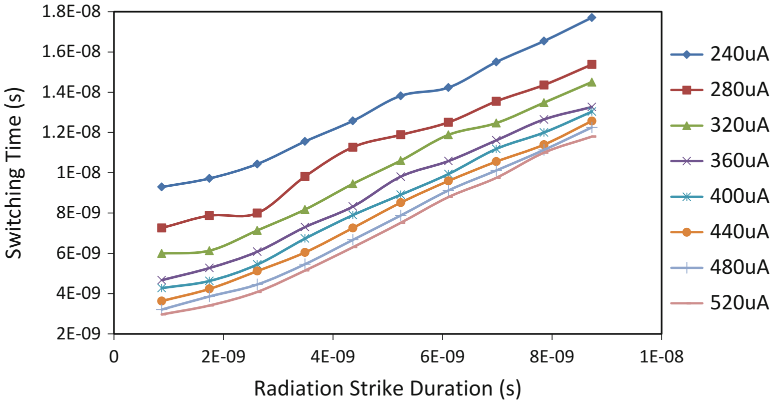


Fig. 5.14 MTJ switching time vs. radiation lasting time

induced torque has a relatively smaller impact (of delaying) on the MTJ switching time, compared to the case that the MTJ is driven by a lower switching current.

The lasting period of the radiation impact is not only determined by the forms of radiation, but also by the recovering ability of the CMOS circuit. The impact of the radiation on a larger transistor is less than the one occurs on a smaller transistor because a larger capacitor on the transistor can better tolerate the radiation charge/discharge disturbance. Also, the same radiation on the transistor with stronger driving ability (e.g., larger size) causes relatively smaller noise on the output due to its faster recovery speed. Nonetheless, longer lasting period results in more prominent disturbance on the magnetization precession and longer delay of MTJ switching, as shown in Fig. 5.14.

The MTJ switching time rises up almost proportionally when the radiation lasting time increases.

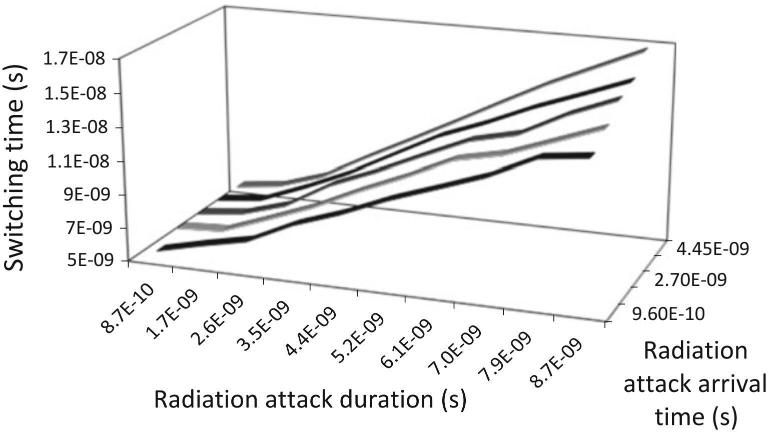


Fig. 5.15 Combination of radiation strike duration and arrival time effects on MTJ switching time

The radiation intensity mainly affects the undershoot slope. The deeper the undershooting is, more charges are needed to recover the driving ability. As a result, the lasting period will be prolonged.

Figure 5.15 shows the combined effects of radiation lasting period and arrival time during the write operation of STT-RAM. In general, MTJ switching is more sensitive to the lasting time of radiation compared to its arrival time. We note that a continuous magnetization precession is necessary to ensure a short MTJ switching time. If the prolonged MTJ switching time under the radiation exceeds the write current pulse width, i.e., the write current is removed before the MTJ switching finishes, a write error will occur.

5.6.4 *Optimization*

As discussed in the previous section, increasing the NMOS transistor size can efficiently improve the MTJ switching performance by supplying higher current. On top of that, increasing the NMOS transistor size helps the circuit to recover from the attack, leading to a shorter radiation lasting period. However, the increase of NMOS transistor size requires larger layout area, which increases the hit rate of high energy particles proportionally. The soft error rate (SER) can be estimated as [22]:

$$SER = R * \alpha \ * P\ (SE) \tag{5.18}$$

Here $R = 57$ neutrons/(m^2 s) is the particle hit rate at NYC sea level, $\alpha = 2.2*10^{-5}$ is effective particle hit ratio and $P\ (SE)$ is the probability of soft error event given an effective particle hit. In Table 5.6, $P(SE)$ is evaluated under an effective particle hit for different write current lasting time and transistor size.

Table 5.6 Probability of soft error event given an effective particle hit

NMOS width (nm)	Error rate (10 ns)	Error rate (8 ns)	Error rate (6 ns)
400	79 %	100 %	100 %
540	62 %	79 %	100 %
780	0	36.9 %	80 %
1,000	0	0	66.7 %

We can see that the larger the transistor size is and the longer the write current lasts, the smaller the soft error possibility we achieve. As expected, a large transistor size helps to secure the MTJ switching before the write current is removed and reduce the radiation effect.

However, note that if we take the relationship between the layout area and the particle hit rate into consideration, the trend of the soft error rate will be affected by multiple factors: while larger transistor supplies higher write current, it also exposes more to the particle hits due to its larger area. There is correspondingly more possibility to generate larger disturbance on the write current. Then $P(SE)$ will depend on the relationship between these two factors. Hence, the transistor sizing needs to be performed carefully with intensive simulations to achieve the radiation hardness optimization. Our model offers a way to explore potential design methods to enhance the radiation hardness of STT-RAM with less computation cost.

Conclusion

In this work, we develop a compact MTJ switching model that is derived from the macro-magnetic modeling. The statistical electrical and magnetic properties of the MTJ, i.e., the resistance switching transience and switching time, can be simulated by our model with the minimized run time cost. Moreover, the proposed model can be used to simulate soft error impacts on STT-RAM cells, which has not been well addressed in the previous works. The demonstrated applications in the static and transient analysis of STT-RAM designs show that our model offers a capable and efficient STT-RAM cell debug solution which can bridge the circuit design and magnetic device physics.

Acknowledgment This work is supported by NSF award: CNS-1116171 and CCF-1217947.

References

1. Raychowdhury A, Somasekhar D, Karnik T, De V. Design space and scalability exploration of 1T-1STTMTJ memory arrays in the presence of variability and disturbances. In: IEEE IEDM, 2009, p. 1–4.

2. Wang P, Wang X, Zhang Y, Li H, Levitan S, Chen Y. Non-persistent errors optimization in spin-MOS logic and storage circuitry. IEEE Trans Magn. 2011;47:3860–3.
3. Sun G, Dong X, Xie Y, Li J, Chen Y. A novel architecture of the 3D stacked MRAM L2 cache for CMPs. In: IEEE 15th International Symposium HPCA, 2009, p. 239–49.
4. Zhang Y, Wang X, Li Y, Jones A, Chen Y. Asymmetry of MTJ switching and its implication to STT-RAM designs. In: EDA Consortium Conference on Design, Automation and Test in Europe, 2012, p. 1313–8.
5. Li J, Augustine C, Salahuddin S, Roy K. Modeling of failure probability and statistical design of spin-torque transfer magnetic randomaccess memory (STT MRAM) array for yield enhancement. In: 45th ACM/IEEE DAC, 2008, p. 278–83.
6. Nigam A, Smullen CW, Mohan V, Chen E, Gurumurthi S, Stan MR. Delivering on the promise of universal memory for spin-transfer torque RAM (STT-RAM). In: ISLPED, 2011, p. 121–6.
7. Joshi R, Kanj R, Wang P, Li H. Universal statistical cure for predicting memory loss. In: IEEE/ACM ICCAD, 2011, p. 236–9.
8. Sun JZ. Spin-current interaction with a monodomain magnetic body: a model study. Phys Rev B Condens Matter. 2000;62:570–8.
9. Slonczewski JC. Conductance and exchange coupling of two ferromagnets separated by a tunneling barrier. Phys Rev B Condens Matter. 1989;39:6995–7002.
10. O'Handley RC. Model for strain and magnetization in magnetic shape-memory alloys. J Appl Phys. 1998;83:3263–70.
11. Koch RH, Katine JA, Sun JZ. Time-resolved reversal of spin transfer switching in a nanomagnet. Phys Rev Lett. 2004;92:088302.
12. Levitt MH. Demagnetization field effects in two-dimensional solution NMR. Concepts Magn Reson. 1996;8:77–103.
13. Beleggia M, Graef MD, Millev YT, Goode DA, Rowlands G. Demagnetization factors for elliptic cylinders. J Phys D Appl Phys. 2005;38:3333–42.
14. Liniers M, Flores J, Bermejo FJ, Gonzales JM, Vicent JL, et al. Systematic study of the temperature dependence of the saturation magnetization in Fe, Fe-Ni and Co-based amorphous alloys. IEEE Trans Magn. 1989;25:3363–5.
15. Wang X, Chen Y, Li H, Dimitrov D, Liu H. Spin torque random access memory down to 22 nm technology. IEEE Trans Magn. 2008;44:2479–82.
16. Chen Y, Wang X, Li H, Liu H, Dimitrov D. Design margin exploration of spin-torque transfer RAM (SPRAM). In: ISQED, 2008, p. 684–90.
17. Higo Y, Yamane K, Ohba K, Narisawa H, Bessho K, et al. Thermal activation effect on spin transfer switching in magnetic tunnel junctions. Appl Phys Lett. 2005;87:082502.
18. Bi X, Li H, Kim J.-J. Analysis and Optimization of Thermal Effect on STT-RAM Based 3-D Stacked Cache Design," IEEE Computer Society Annual Symposium on VLSI (ISVLSI), Aug. 2012, pp. 374–9. DOI: 10.1109/ISVLSI.2012.56.
19. Baumann RC. Radiation-induced soft errors in advanced semiconductor technologies. IEEE Trans Device Mater Reliab. 2005;5:305–16.
20. Predictive Technology Model (PTM). http://www.eas.asu.edu/ ptm/.
21. Naseer R, Boulghassoul Y, Draper J, DasGupta S, Witulski A. Critical charge characterization for soft error rate modeling in 90 nm SRAM. In: IEEE ISCAS, 2007, p. 1879–82.
22. Zhang M, Shanbhag NR. Soft-error-rate-analysis (SERA) methodology. IEEE Trans Comput Aided Des Integr Circuits Syst. 2006;25:2140–55.

Chapter 6
Main Memory Scaling: Challenges and Solution Directions

Onur Mutlu

Abstract The memory system is a fundamental performance and energy bottleneck in almost all computing systems. Recent system design, application, and technology trends that require more capacity, bandwidth, efficiency, and predictability out of the memory system make it an even more important system bottleneck. At the same time, DRAM technology is experiencing difficult *technology scaling* challenges that make the maintenance and enhancement of its capacity, energy-efficiency, and reliability significantly more costly with conventional techniques.

In this chapter, after describing the demands and challenges faced by the memory system, we examine some promising research and design directions to overcome challenges posed by memory scaling. Specifically, we describe three major solution directions: (1) enabling new DRAM architectures, functions, interfaces, and better integration of the DRAM and the rest of the system (an approach we call *system-DRAM co-design*), (2) designing a memory system that employs emerging non-volatile memory technologies and takes advantage of multiple different technologies (i.e., *hybrid memory systems*), (3) providing predictable performance and QoS to applications sharing the memory system (i.e., *QoS-aware memory systems*). We also briefly describe our ongoing related work in combating scaling challenges of NAND flash memory.

6.1 Introduction

Main memory is a critical component of all computing systems, employed in server, embedded, desktop, mobile and sensor environments. Memory capacity, energy, cost, performance, and management algorithms must scale as we scale the size of the computing system in order to maintain performance growth and enable new applications. Unfortunately, such scaling has become difficult because recent trends in systems, applications, and technology greatly exacerbate the memory system bottleneck.

O. Mutlu (✉)
Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA
e-mail: onur@cmu.edu

R.O. Topaloglu (ed.), *More than Moore Technologies for Next Generation Computer Design*, DOI 10.1007/978-1-4939-2163-8_6

6.2 Trends: Systems, Applications, Technology

In particular, on the *systems/architecture front*, energy and power consumption have become key design limiters as the memory system continues to be responsible for a significant fraction of overall system energy/power [69]. More and increasingly heterogeneous processing cores and agents/clients are sharing the memory system [4, 19, 21, 34, 45, 46, 107], leading to increasing demand for memory capacity and bandwidth along with a relatively new demand for predictable performance and quality of service (QoS) from the memory system [81, 87, 106].

On the *applications front*, important applications are usually very data intensive and are becoming increasingly so [6], requiring both real-time and offline manipulation of great amounts of data. For example, next-generation genome sequencing technologies produce massive amounts of sequence data that overwhelms memory storage and bandwidth requirements of today's high-end desktop and laptop systems [2, 109, 115] yet researchers have the goal of enabling low-cost personalized medicine. Creation of new *killer applications* and usage models for computers likely depends on how well the memory system can support the efficient storage and manipulation of data in such data-intensive applications. In addition, there is an increasing trend towards consolidation of applications on a chip to improve efficiency, which leads to the sharing of the memory system across many heterogeneous applications with diverse performance requirements, exacerbating the aforementioned need for predictable performance guarantees from the memory system [106, 108].

On the *technology front*, two major trends profoundly affect memory systems. First, there is increasing difficulty scaling the well-established charge-based memory technologies, such as DRAM [3, 38, 40, 53, 62, 77] and flash memory [8, 9, 12, 59, 76], to smaller technology nodes. Such scaling has enabled memory systems with reasonable capacity and efficiency; lack of it will make it difficult to achieve high capacity and efficiency at low cost. Second, some emerging resistive memory technologies, such as phase change memory (PCM) [62, 63, 98, 100, 113], spin-transfer torque magnetic memory (STT-MRAM) [17, 60] or resistive RAM (RRAM) [114] appear more scalable, have latency and bandwidth characteristics much closer to DRAM than flash memory and hard disks, and are non-volatile with little idle power consumption. Such emerging technologies can enable new opportunities in system design, including, for example, the unification of memory and storage subsystems [80]. They have the potential to be employed as part of main memory, alongside or in place of less scalable and leaky DRAM, but they also have various shortcomings depending on the technology (e.g., some have cell endurance problems, some have very high write latency/power, some have low density) that need to be overcome or tolerated.

6.3 Requirements: Traditional and New

System architects and users have always wanted more from the memory system: high performance (ideally, zero latency and infinite bandwidth), infinite capacity, all at zero cost! The aforementioned trends do not only exacerbate and morph the above requirements, but also add some new requirements. We classify the requirements from the memory system into two categories: *exacerbated traditional requirements* and *(relatively) new requirements*.

The traditional requirements of performance, capacity, and cost are greatly exacerbated today due to increased pressure on the memory system, consolidation of multiple applications/agents sharing the memory system, and difficulties in DRAM technology and density scaling. In terms of *performance*, two aspects have changed. First, today's systems and applications not only require low latency and high bandwidth (as traditional memory systems have been optimized for), but they also require new techniques to manage and control memory interference between different cores, agents, and applications that share the memory system [24, 81, 87, 106, 108] in order to provide high system performance as well as predictable performance (or quality of service) to different applications [106]. Second, there is a need for increased memory bandwidth for many applications as the placement of more cores and agents on chip make the memory pin bandwidth an increasingly precious resource that determines system performance [41], especially for memory-bandwidth-intensive workloads, such as GPGPUs [47, 48], heterogeneous systems [4], and consolidated workloads [43, 44, 87]. In terms of *capacity*, the need for memory capacity is greatly increasing due to the placement of multiple data-intensive applications on the same chip and continued increase in the data sets of important applications. One recent work showed that given that the core count is increasing at a faster rate than DRAM capacity, the expected memory capacity per core is to drop by 30 % every 2 years [70], an alarming trend since much of today's software innovations and features rely on increased memory capacity. In terms of *cost*, increasing difficulty in DRAM technology scaling poses a difficult challenge to building higher density (and, as a result, lower cost) main memory systems. Similarly, cost-effective options for providing high reliability and increasing memory bandwidth are needed to scale the systems proportionately with the reliability and data throughput needs of today's data-intensive applications. Hence, the three traditional requirements of performance, capacity, and cost have become exacerbated.

The relatively new requirements from the main memory system are threefold. First, *technology scalability*: there is a new need for finding a technology that is much more scalable than DRAM in terms of capacity, energy, and cost, as described earlier. As DRAM continued to scale well from the above-100–30 nm technology nodes, the need for finding a more scalable technology was not a prevalent problem. Today, with the significant circuit and device scaling challenges DRAM has been facing below the 30 nm node, it is. Second, there is a relatively new need for providing *performance predictability and QoS* in the shared main memory system. As single-core systems were dominant and memory bandwidth and capacity were much less of a shared resource in the past, the need for predictable performance was much less apparent or prevalent [81]. Today, with increasingly more cores/agents on chip sharing the memory system and increasing amounts of workload consolidation, memory fairness, predictable memory performance, and techniques to mitigate memory interference have become first-class design constraints. Third, there is a great need for much higher *energy/power/bandwidth efficiency* in the design of the main memory system. Higher efficiency in terms of energy, power, and bandwidth enables the design of much more scalable systems where main memory is shared between many agents, and can enable new applications in almost all domains where computers are used. Arguably, this is not a new need today, but we believe it is another first-class design constraint that has not been as traditional as performance, capacity, and cost.

6.4 Solution Directions

As a result of these systems, applications, and technology trends and the resulting requirements, it is our position that researchers and designers need to fundamentally rethink the way we design memory systems today to (1) overcome scaling challenges with DRAM, (2) enable the use of emerging memory technologies, (3) design memory systems that provide predictable performance and quality of service to applications and users. The rest of this chapter describes our solution ideas in these three directions, with pointers to specific techniques when possible. Since scaling challenges themselves arise due to difficulties in enhancing memory components at *solely* one level of the computing stack (e.g., the device and/or circuit levels in case of DRAM scaling), we believe effective solutions to the above challenges will require cooperation across different layers of the computing stack, from algorithms to software to microarchitecture to devices, as well as between different components of the system, including processors, memory controllers, memory chips, and the storage subsystem. As much as possible, we will give examples of such solutions and directions.

6.5 Challenge 1: New DRAM Architectures

DRAM has been the choice technology for implementing main memory due to its relatively low latency and low cost. DRAM process technology scaling has for long enabled lower cost per unit area by enabling reductions in DRAM cell size. Unfortunately, further scaling of DRAM cells has become costly [3,38,40,53,62,77] due to increased manufacturing complexity/cost, reduced cell reliability, and potentially increased cell leakage leading to high refresh rates. Several key issues to tackle include:

1. reducing the negative impact of refresh on energy, performance, QoS, and density scaling [15,71,72],
2. improving DRAM parallelism/bandwidth [15, 57], latency [68], and energy efficiency [57,68,71],
3. improving reliability of DRAM at low cost [51,58,75,90],
4. reducing the significant amount of waste present in today's main memories in which much of the fetched/stored data can be unused due to coarse-granularity management [79,94,95,110,117],
5. minimizing data movement between DRAM and processing elements, which causes high latency, energy, and bandwidth consumption [102].

Traditionally, DRAM devices have been separated from the rest of the system with a rigid interface, and DRAM has been treated as a *passive* slave device that simply responds to the commands given to it by the memory controller. We believe the above key issues can be solved more easily if we rethink the DRAM architecture and functions, and redesign the interface such that DRAM, controllers, and processors closely cooperate. We call this high-level solution approach *system-DRAM co-design*. We believe key technology trends, e.g., the 3D stacking of memory and logic [39, 74, 111] and increasing cost of scaling DRAM solely via circuit-level approaches [38,53,77], enable such a co-design to become increasingly more feasible. We proceed to provide several examples from our recent research that tackle the problems of refresh, parallelism, latency, and energy efficiency.

6.5.1 Reducing Refresh Impact and DRAM Error Management

With higher DRAM capacity, more cells need to be refreshed at likely higher rates than today. Our recent work [71] indicates that refresh rate limits DRAM density scaling: a hypothetical 64 Gb DRAM device would spend 46 % of its time and 47 % of all DRAM energy for refreshing its rows, as opposed to typical 4 Gb devices of today that spend respectively 8 % of the time and 15 % of the DRAM energy on refresh (as shown in Fig. 6.1). Today's DRAM devices refresh all rows at the same worst-case rate (e.g., every 64 ms). However, only a small number of weak rows require a high refresh rate [51,54,72] (e.g., only ~1000 rows in 32 GB DRAM require to be refreshed more frequently than every 256 ms). Retention-Aware Intelligent DRAM Refresh (RAIDR) [71] exploits this observation: it groups DRAM rows into bins (implemented as Bloom filters [5] to minimize hardware overhead) based on the retention time of the weakest cell within each row. Each row is refreshed at a rate corresponding to its retention time bin. Since few rows need high refresh rate, one can use very few bins to achieve large reductions in refresh counts: our results show that RAIDR with three bins (1.25 KB hardware cost) reduces refresh operations by ~75 %, leading to significant improvements in system performance and energy efficiency as described by Liu et al. [71].

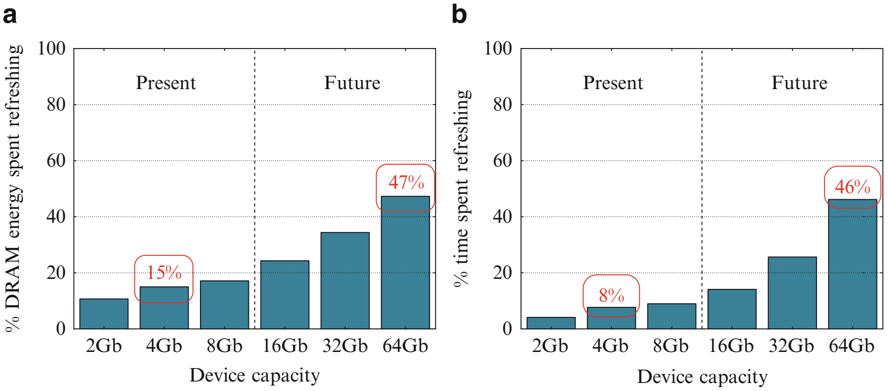


Fig. 6.1 Impact of refresh in current (DDR3) and projected DRAM devices. Reproduced from [71]. (**a**) Power consumption, (**b**) throughput loss

Like RAIDR, other approaches have also been proposed to take advantage of the retention time variation of cells across a DRAM chip. For example, some works proposed refreshing weak rows more frequently at a per-row granularity, others proposed not using memory rows with low retention times, and yet others suggested mapping critical data to cells with longer retention times such that critical data is not lost [1, 42, 52, 73, 93, 112]—see [71, 72] for a discussion of such techniques. Such approaches that exploit non-uniform retention times across DRAM require accurate retention time profiling mechanisms. Understanding of retention time as well as error behavior of DRAM devices is a critical research topic, which we believe can enable other mechanisms to tolerate refresh impact and errors at low cost. Liu et al. [72] provides an experimental characterization of retention times in modern DRAM devices to aid such understanding. Our initial results in that work, obtained via the characterization of 248 modern commodity DRAM chips from five different DRAM manufacturers, suggest that the retention time of cells in a modern device is largely affected by two phenomena: (1) Data Pattern Dependence, where the retention time of each DRAM cell is significantly affected by the data stored in other DRAM cells, (2) Variable Retention Time, where the retention time of a DRAM cell changes unpredictably over time. These two phenomena pose challenges against accurate and reliable determination of the retention time of DRAM cells, online or offline, and a promising area of future research is to devise techniques that can identify retention times of DRAM cells in the presence of data pattern dependence and variable retention time. Khan et al.'s recent work [51] provides more analysis of the effectiveness of conventional error mitigation mechanisms for DRAM retention failures and proposes *online retention time profiling* as a solution for identifying retention times of DRAM cells as a potentially promising approach in future DRAM systems.

Looking forward, we believe that increasing cooperation between the DRAM device and the DRAM controller as well as other parts of the system, including system software, is needed to communicate information about *weak* (or, unreliable) cells and the characteristics of different rows or physical memory regions from the device to the system. The system can then use this information to optimize data allocation and movement, refresh rate management, and error tolerance mechanisms. Low-cost error tolerance mechanisms are likely to be enabled more efficiently with such coordination between DRAM and the system. In fact, as DRAM technology scales and error rates increase, it might become increasingly more difficult to maintain the common illusion that DRAM is a perfect, error-free storage device. DRAM may start looking increasingly like flash memory, where the memory controller manages errors such that an acceptable specified uncorrectable bit error rate is satisfied [8, 10]. We envision a *DRAM Translation Layer (DTL)*, not unlike the *Flash Translation Layer (FTL)* of today in spirit (which is decoupled from the processor and performs a wide variety of management functions for flash memory, including error correction, garbage collection, read/write scheduling, etc.), can enable better scaling of DRAM memory into the future by not only enabling easier error management but also opening up new opportunities to perform computation and mapping close to memory. This can become especially feasible in the presence of the trend of combining the DRAM controller and DRAM via 3D stacking. What should the interface be to such a layer and what should be performed in the DTL are promising areas of future research.

6.5.2 Improving DRAM Parallelism

A key limiter of DRAM parallelism is bank conflicts. Today, a bank is the finest-granularity independently accessible memory unit in DRAM. If two accesses go to the same bank, one has to *completely* wait for the other to finish before it can be started (see Fig. 6.2). We have recently developed mechanisms, called SALP (subarray level parallelism) [57], that exploit the internal subarray structure of the DRAM bank (Fig. 6.2) to *mostly* parallelize two requests that access the same

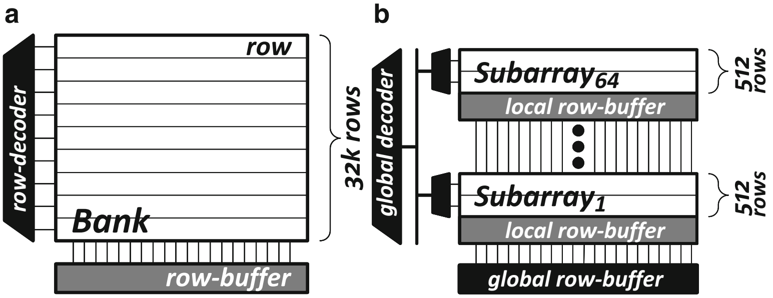


Fig. 6.2 DRAM bank organization. Reproduced from [57]. (**a**) Logical abstraction of a bank. (**b**) Implementation of a DRAM bank

DRAM bank. The key idea is to reduce the hardware sharing between DRAM subarrays such that accesses to the same bank but different subarrays can be initiated in a pipelined manner. This mechanism requires the exposure of the internal subarray structure of a DRAM bank to the controller and the design of the controller to take advantage of this structure. Our results show significant improvements in performance and energy efficiency of main memory due to parallelization of requests and improvement of row buffer hit rates (as row buffers of different subarrays can be kept active) at a low DRAM area overhead of 0.15 %. Exploiting SALP achieves most of the benefits of increasing the number of banks at much lower area and power overhead than doing so. Exposing the subarray structure of DRAM to other parts of the system, e.g., to system software or memory allocators, can enable data placement and partitioning mechanisms that can improve performance and efficiency even further.

Note that other approaches to improving DRAM parallelism especially in the presence of refresh and long write latencies are also promising to investigate. Chang et al. [15] discuss mechanisms to improve the parallelism between reads and writes, and Kang et al. [50] discuss the use of SALP as a way of tolerating long write latencies to DRAM, which they identify as one of the three key scaling challenges for DRAM, amongst refresh and variable retention time. We refer the reader to these works for more information about these parallelization techniques.

6.5.3 *Reducing DRAM Latency and Energy*

The DRAM industry has so far been primarily driven by the cost-per-bit metric: provide maximum capacity for a given cost. As shown in Fig. 6.3, DRAM chip capacity has increased by approximately 16× in 12 years while the DRAM latency reduced by only approximately 20 %. This is the result of a deliberate choice to

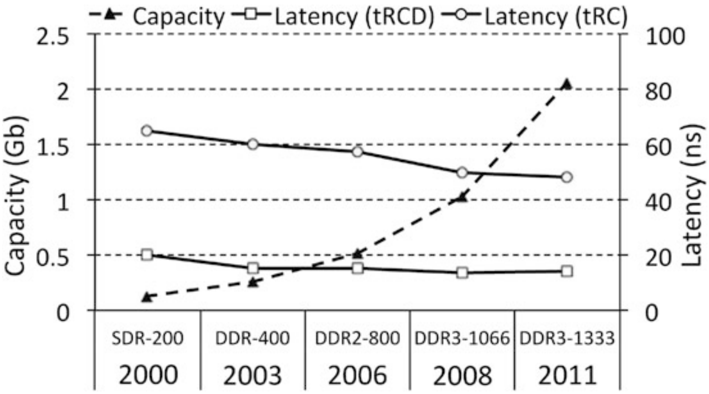


Fig. 6.3 DRAM capacity & latency over time. Reproduced from [68]

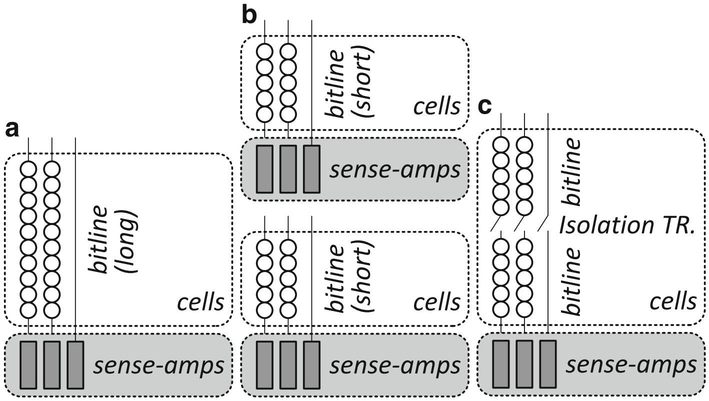


Fig. 6.4 Cost optimized commodity DRAM (**a**), latency optimized DRAM (**b**), tiered-latency DRAM (**c**). Reproduced from [68]

maximize capacity of a DRAM chip while minimizing its cost. We believe this choice needs to be revisited in the presence of at least two key trends. First, DRAM latency is becoming more important especially for response-time critical workloads that require QoS guarantees [26]. Second, DRAM capacity is becoming very hard to scale and as a result manufacturers likely need to provide new values for the DRAM chips, leading to more incentives for the production of DRAMs that are optimized for objectives other than mainly capacity maximization.

To mitigate the high area overhead of DRAM sensing structures, commodity DRAMs (shown in Fig. 6.4a) connect many DRAM cells to each sense-amplifier through a wire called a bitline. These bitlines have a high parasitic capacitance due to their long length, and this bitline capacitance is the dominant source of DRAM latency. Specialized low-latency DRAMs (shown in Fig. 6.4b) use shorter bitlines with fewer cells, but have a higher cost-per-bit due to greater sense-amplifier area overhead. We have recently shown that we can architect a heterogeneous-latency bitline DRAM, called Tiered-Latency DRAM (TL-DRAM) [68], shown in Fig. 6.4c, by dividing a long bitline into two shorter segments using an isolation transistor: a low-latency segment can be accessed with the latency and efficiency of a short-bitline DRAM (by turning off the isolation transistor that separates the two segments) while the high-latency segment enables high density, thereby reducing cost-per-bit. The additional area overhead of TL-DRAM is approximately 3 % over commodity DRAM. Significant performance and energy improvements can be achieved by exposing the two segments to the memory controller and system software such that appropriate data is cached or allocated into the low-latency segment. We expect such approaches that design and exploit heterogeneity to enable/achieve the best of multiple worlds [84] in the memory system can lead to other novel mechanisms that can overcome difficult contradictory tradeoffs in design.

Another promising approach to reduce DRAM energy is the use of dynamic voltage and frequency scaling (DVFS) in main memory [25, 27]. David et al. [25] make the observation that at low memory bandwidth utilization, lowering memory frequency/voltage does not significantly alter memory access latency. Relatively recent works have shown that adjusting memory voltage and frequency based on predicted memory bandwidth utilization can provide significant energy savings on both real [25] and simulated [27] systems. Going forward, memory DVFS can enable dynamic heterogeneity in DRAM channels leading to new customization and optimization mechanisms. Also promising is the investigation of more fine-grained power management methods within the DRAM rank and chips for both active and idle low power modes.

6.5.4 Exporting Bulk Data Operations to DRAM

Today's systems waste significant amount of energy, DRAM bandwidth and time (as well as valuable on-chip cache space) by sometimes unnecessarily moving data from main memory to processor caches. One example of such wastage sometimes occurs for bulk data copy and initialization operations in which a page is copied to another or initialized to a value. If the copied or initialized data is not immediately needed by the processor, performing such operations within DRAM (with relatively small changes to DRAM) can save significant amounts of energy, bandwidth, and time. We observe that a DRAM chip internally operates on bulk data at a row granularity. Exploiting this internal structure of DRAM can enable page copy and initialization to be performed entirely within DRAM without bringing any data off the DRAM chip, as we have shown in recent work [102]. If the source and destination page reside within the same DRAM subarray, our results show that a page copy can be accelerated by more than an order of magnitude (~11 times), leading to an energy reduction of ~74 times and *no* wastage of DRAM data bus bandwidth [102]. The key idea is to capture the contents of the source row in the sense amplifiers by activating the row, then *deactivating* the source row (using a new command which introduces very little hardware cost, amounting to less than 0.03 % of DRAM chip area), and immediately activating the destination row, which causes the sense amplifiers to drive their contents into the destination row, effectively accomplishing the page copy (shown at a high level in Fig. 6.5). Doing so reduces the latency of a 4 KB page copy operation from ~1,000 ns to less than 100 ns in an existing DRAM chip. Applications that have significant page copy and initialization lead to large system performance and energy efficiency improvements [102]. Future software can be designed in ways that can take advantage of such fast page copy and initialization operations, leading to benefits that may not be apparent in today's software that tends to minimize such operations due to their current high cost.

Going forward, we believe acceleration of other bulk data movement and computation operations in or very close to DRAM, via similar low-cost architectural

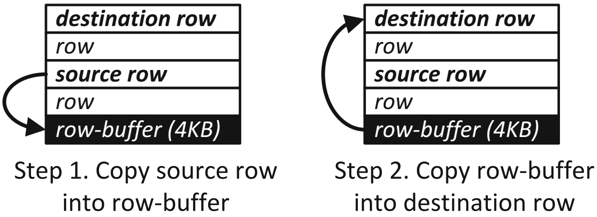


Fig. 6.5 High-level idea behind RowClone's in-DRAM page copy mechanism

support mechanisms, can enable promising savings in system energy, latency, and bandwidth. Given the trends and requirements described in Sect. 6.2, it is likely time to re-examine the partitioning of computation between processors and DRAM, treating memory as a first-class accelerator as an integral part of a heterogeneous parallel computing system [84].

6.5.5 Minimizing Capacity and Bandwidth Waste

Storing and transferring data at large granularities (e.g., pages, cache blocks) within the memory system leads to large inefficiency when most of the large granularity is not needed [49, 61, 78, 79, 96, 101, 110, 117, 118]. In addition, much of the data stored in memory has significant redundancy [33, 94, 95, 116]. Two promising research directions are to develop techniques that can (1) efficiently provide fine granularity access/storage when enough and large granularity access/storage only when needed, (2) efficiently compress data in main memory and caches without significantly increasing latency and system complexity. Our results with new low-cost, low-latency cache compression [94] and memory compression [95] techniques and frameworks are promising, providing high compression ratios at low complexity and latency. For example, the key idea of *Base-Delta-Immediate compression* [94] is that many cache blocks have low dynamic range in the values they store, i.e., the differences between values stored in the cache block are small. Such a cache block can be encoded using a base value and an array of much smaller (in size) differences from that base value, which together occupy much less space than storing the full values in the original cache block. This compression algorithm has low decompression latency as the cache block can be reconstructed using a vector addition (or even potentially vector concatenation). It reduces memory bandwidth requirements, better utilizes memory/cache space, while minimally impacting the latency to access data. Granularity management and data compression support can potentially be integrated into DRAM controllers or partially provided within DRAM, and such mechanisms can be exposed to software, which can enable higher energy savings and higher performance improvements.

6.5.6 Co-Designing DRAM Controllers and Processor-Side Resources

Since memory bandwidth is a precious resource, coordinating the decisions made by processor-side resources better with the decisions made by memory controllers to maximize memory bandwidth utilization and memory locality is a promising area of more efficiently utilizing DRAM. Lee et al. [67] and Stuecheli et al. [105] both show that orchestrating last-level cache writebacks such that dirty cache lines to the same row are evicted together from the cache improves DRAM row buffer locality of write accesses, thereby improving system performance. Going forward, we believe such coordinated techniques between the processor-side resources and memory controllers will become increasingly more effective as DRAM bandwidth becomes even more precious. Mechanisms that predict and convey slack in memory requests [23], that orchestrate the on-chip scheduling of memory requests to improve memory bank parallelism [66] and that reorganize cache metadata for more efficient bulk (DRAM row granularity) tag lookups [103] can also enable more efficient memory bandwidth utilization.

6.6 Challenge 2: Emerging Memory Technologies

While DRAM technology scaling is in jeopardy, some emerging technologies seem more scalable. These include PCM and STT-MRAM. These emerging technologies usually provide a tradeoff, and seem unlikely to completely replace DRAM (evaluated in [62–64] for PCM and in [60] for STT-MRAM), as they are not strictly superior to DRAM. For example, PCM is advantageous over DRAM because it (1) has been demonstrated to scale to much smaller feature sizes and can store multiple bits per cell [120], promising higher density, (2) is non-volatile and as such requires no refresh (which is a key scaling challenge of DRAM as we discussed in Sect. 6.5.1), and (3) has low idle power consumption. On the other hand, PCM has significant shortcomings compared to DRAM, which include (1) higher read latency and read energy, (2) *much* higher write latency and write energy, and (3) limited endurance for a given PCM cell, a problem that does not exist (practically) for a DRAM cell. As a result, an important research challenge is how to utilize such emerging technologies at the system and architecture levels such that they can augment or perhaps even replace DRAM.

Our initial experiments and analyses [62–64] that evaluated the complete replacement of DRAM with PCM showed that one would require reorganization of peripheral circuitry of PCM chips (with the goal of absorbing writes and reads before they update or access the PCM cell array) to enable PCM to get close to DRAM performance and efficiency. These initial results are reported in Lee et al. [62–64]. We have also reached a similar conclusion upon evaluation of the complete replacement of DRAM with STT-MRAM [60]: reorganization of peripheral circuitry of STT-MRAM chips (with the goal of minimizing the number of writes to the STT-MRAM cell array, as write operations are high-latency and high-energy in STT-MRAM) enables an STT-MRAM based main memory to be more energy-efficient than a DRAM-based main memory.

One can achieve more efficient designs of PCM (or STT-MRAM) chips by taking advantage of the non-destructive nature of reads, which enables simpler and narrower row buffer organizations [78] Unlike in DRAM, the entire memory row does not need to be buffered in a device where reading a memory row does not destroy the data stored in the row. Meza et al. [78] show that having narrow row buffers in emerging non-volatile devices can greatly reduce main memory dynamic energy compared to a DRAM baseline with large row sizes, without greatly affecting endurance, and for some NVM technologies, leading to improved performance. Going forward, designing systems, memory controllers and memory chips taking advantage of the specific property of non-volatility of emerging technology seems promising.

We believe emerging technologies enable at least three major system-level opportunities that can improve overall system efficiency: (1) hybrid main memory systems, (2) non-volatile main memory, (3) merging of memory and storage. We briefly touch upon each.

6.6.1 Hybrid Main Memory

A hybrid main memory system [28, 79, 98, 119] consists of multiple different technologies or multiple different types of the same technology with differing characteristics, e.g., performance, cost, energy, reliability, endurance. A key question is how to manage data allocation and movement between the different technologies such that one can achieve the best of (or close to the best of) the desired performance metrics. In other words, we would like to exercise the advantages of each technology as much as possible while hiding the disadvantages of any technology. Potential technologies include DRAM, 3D-stacked DRAM, embedded DRAM, PCM, STT-MRAM, other resistive memories, flash memory, forms of DRAM that are optimized for different metrics and purposes, etc. An example hybrid main memory system consisting of a large amount of PCM as main memory and a small amount of DRAM as its cache is depicted in Fig. 6.6.

The design space of hybrid memory systems is large, and many potential questions exist. For example, should all memories be part of main memory or should some of them be used as a cache of main memory (or should there be configurability)? What technologies should be software visible? What component of the system should manage data allocation and movement? Should these tasks be done in hardware, software, or collaboratively? At what granularity should data moved between different memory technologies? Some of these questions are tackled

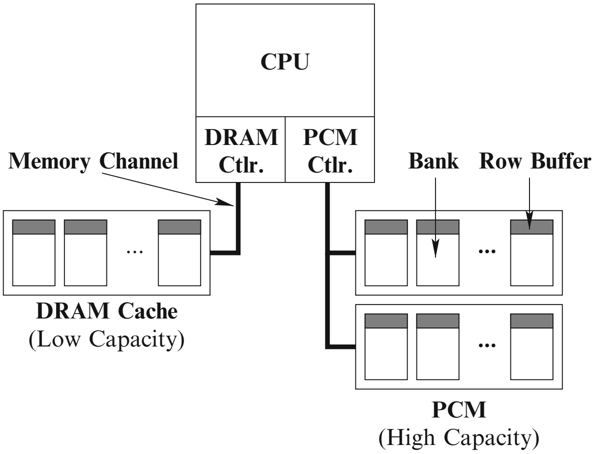


Fig. 6.6 An example hybrid main memory system organization using PCM and DRAM chips. Reproduced from [119]

in [28, 79, 98, 99, 119], among other works recently published in the computer architecture community. For example, Yoon et al. [119] make the key observation that row buffers are present in both DRAM and PCM, and they have (or can be designed to have) the same latency and bandwidth in both DRAM and PCM. Yet, row buffer misses are much more costly in terms of latency, bandwidth, and energy in PCM than in DRAM. To exploit this, we devise a policy that avoids accessing in PCM data that frequently causes row buffer misses. Hardware or software can dynamically keep track of such data and allocate/cache it in DRAM while keeping data that frequently hits in row buffers in PCM. PCM also has much higher write latency/power than read latency/power: to take this into account, the allocation/caching policy is biased such that pages that are written to more likely stay in DRAM [119].

Note that hybrid memory need not consist of completely different underlying technologies. A promising approach is to combine multiple different DRAM chips, optimized for different purposes. For example, recent works proposed the use of low-latency and high-latency DIMMs in separate memory channels and allocating performance-critical data to low-latency DIMMs to improve performance and energy-efficiency at the same time [16], or the use of highly-reliable DIMMs (protected with ECC) and unreliable DIMMs in separate memory channels and allocating error-vulnerable data to highly-reliable DIMMs to maximize server availability while minimizing server memory cost [75]. We believe these approaches are quite promising for scaling the DRAM technology into the future by specializing DRAM chips for different purposes. These approaches that exploit heterogeneity do increase system complexity but that complexity may be warranted if it is lower than the complexity of scaling DRAM chips using the same optimization techniques the DRAM industry has been using so far.

6.6.2 Exploiting and Securing Non-volatile Main Memory

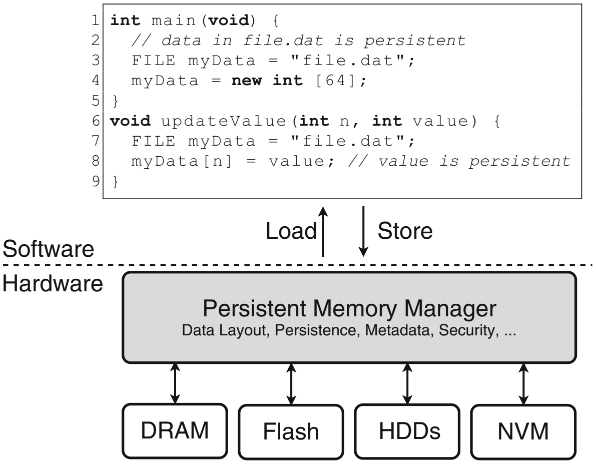
Non-volatility of main memory opens up new opportunities that can be exploited by higher levels of the system stack to improve performance and reliability/consistency (see, for example, [20, 29]). Researching how to adapt applications and system software to utilize fast, byte-addressable non-volatile main memory is an important research direction to pursue [80].

On the flip side, the same non-volatility can lead to potentially unforeseen security and privacy issues: critical and private data can persist long after the system is powered down [18], and an attacker can take advantage of this fact. Wearout issues of emerging technology can also cause attacks that can intentionally degrade memory capacity in the system [97, 104]. Securing non-volatile main memory is therefore an important systems challenge.

6.6.3 Merging of Memory and Storage

Traditional computer systems have a two-level storage model: they access and manipulate (1) volatile data in main memory (DRAM, today) with a fast load/store interface, (2) persistent data in storage media (flash and hard disks, today) with a slower file system interface. Unfortunately, such a decoupled memory/storage model managed via vastly different techniques (fast, hardware-accelerated memory management units on one hand, and slow operating/file system (OS/FS) software on the other) suffers from large inefficiencies in locating data, moving data, and translating data between the different formats of these two levels of storage that are accessed via two vastly different interfacesleading to potentially large amounts of wasted work and energy [80]. The two different interfaces arose largely due to the large discrepancy in the access latencies of conventional technologies used to construct volatile memory (DRAM) and persistent storage (hard disks and flash memory).

Today, new non-volatile memory technologies (NVM), e.g, PCM, STT-MRAM, RRAM, show the promise of storage capacity and endurance similar to or better than flash memory at latencies comparable to DRAM. This makes them prime candidates for providing applications a persistent *single-level store* with a single load/store-like interface to access all system data (including volatile and persistent data). In fact, if we keep the traditional two-level memory/storage model in the presence of these fast NVM devices as part of storage, the operating system and file system code for locating, moving, and translating persistent data from the non-volatile NVM devices to volatile DRAM for manipulation purposes becomes a great bottleneck, causing most of the memory energy consumption and degrading performance by an order of magnitude in some data-intensive workloads, as we showed in recent work [80]. With energy as a key constraint, and in light of modern high-density NVM devices, a



```
int main(void) {
  // data in file.dat is persistent
  FILE myData = "file.dat";
  myData = new int [64];
}
void updateValue(int n, int value) {
  FILE myData = "file.dat";
  myData[n] = value; // value is persistent
}
```

Load
Store
Software
Hardware
Persistent Memory Manager
Data Layout, Persistence, Metadata, Security, ...
DRAM
Flash
HDDs
NVM

Fig. 6.7 An example persistent memory manager (PMM). Reproduced from [80]

promising research direction is to unify and coordinate the management of volatile memory and persistent storage in a single level, to eliminate wasted energy and performance, and to simplify the programming model at the same time.

To this end, Meza et al. [80] describe the vision and research challenges of a persistent memory manager (PMM), a hardware acceleration unit that coordinates and unifies memory/storage management in a single address space that spans potentially multiple different memory technologies (DRAM, NVM, flash) via hardware/software cooperation. Figure 6.7 depicts an example PMM, programmed using a load/store interface (with persistent objects) and managing an array of heterogeneous devices.

The spirit of the PMM unit is much like the virtual memory management unit of a modern virtual memory system used for managing working memory, but it is fundamentally different in that it redesigns/rethinks the virtual memory and storage abstractions and unifies them in a different interface supported by scalable hardware mechanisms. The PMM: (1) exposes a load/store interface to access persistent data, (2) manages data placement, location, persistence semantics, and protection (across multiple memory devices) using both dynamic access information and hints from the application and system software, (3) manages metadata storage and retrieval, needed to support efficient location and movement of persistent data, and (4) exposes hooks and interfaces for applications and system software to enable intelligent data placement and persistence management. Our preliminary evaluations show that the use of such a unit, if scalable and efficient, can greatly reduce the energy inefficiency and performance overheads of the two-level storage model, improving both performance and energy-efficiency of the overall system, especially for data-intensive workloads [80].

We believe there are challenges to be overcome in the design, use, and adoption of such a unit that unifies working memory and persistent storage. These challenges include:

1. How to devise efficient and scalable data mapping, placement, and location mechanisms (which likely need to be hardware/software cooperative).
2. How to ensure that the consistency and protection requirements of different types of data are adequately, correctly, and reliably satisfied. How to enable the reliable and effective coexistence and manipulation of volatile and persistent data.
3. How to redesign applications such that they can take advantage of the unified memory/storage interface and make the best use of it by providing appropriate hints for data allocation and placement to the persistent memory manager.
4. How to provide efficient and high-performance backward compatibility mechanisms for enabling and enhancing existing memory and storage interfaces in a single-level store. These techniques can seamlessly enable applications targeting traditional two-level storage systems to take advantage of the performance and energy-efficiency benefits of systems employing single-level stores. We believe such techniques are needed to ease the software transition to a radically different storage interface.

6.7 Challenge 3: Predictable Performance

Since memory is a shared resource between multiple cores (or, agents, threads, or applications and virtual machines), different applications contend for memory bandwidth and capacity. As such, memory contention, or memory interference, between different cores critically affects both the overall system performance and each application's performance. Providing the appropriate bandwidth and capacity allocation to each application such that its performance requirements are satisfied is important to satisfy user expectations and service level agreements, and at the same time enable better system performance. Our past work (e.g., [81, 87, 88]) showed that application-unaware design of memory controllers, and in particular memory scheduling algorithms, leads to uncontrolled interference of applications in the memory system. Such uncontrolled interference can lead to denial of service to some applications [81], low system performance [87, 88], and an inability to satisfy performance requirements [30, 87, 106], which makes the system uncontrollable and unpredictable. In fact, an application's performance depends on what other applications it is sharing resources with: an application can sometimes have very high performance and at other times very low performance on the same system, solely depending on its co-runners. A critical research challenge is therefore how to design the memory system (including all shared resources such as main memory, caches, and interconnects) such that (1) the performance of each application is predictable and controllable, (2) while the performance and efficiency of the entire system are as high as needed or possible.

To achieve these goals, we have designed various solutions including QoS-aware memory controllers [4, 31, 55, 56, 65, 82, 83, 87, 88, 106], interconnects [14, 22–24, 36, 37, 91, 92], and entire memory systems including caches, interconnect, and memory [24, 30, 32]. These works enhanced our understanding of memory interference in multi-core and heterogeneous systems and provide viable and effective mechanisms that improve overall system performance, while also providing a fairness substrate that can enable fair memory service, which can be configured to enforce different application priorities.

A promising direction going forward is to devise mechanisms that are effective and accurate at (1) estimating and predicting application performance in the presence of interference and a dynamic system with continuously incoming and outgoing applications and (2) enforcing *end-to-end performance guarantees* within the entire shared memory system. Subramanian et al. [106] provides a simple method, called MISE (Memory-interference Induced Slowdown Estimation), for estimating application slowdowns in the presence of main memory interference. We observe that an application's memory request service rate is a good proxy for its performance, as depicted in Fig. 6.8, which shows the measured performance versus memory request service rate for three applications on a real system [106]. As such, an application's slowdown can be accurately estimated by estimating its uninterfered request service rate, which can be done by prioritizing that application's requests in the memory system during some execution intervals. Results show that average error in slowdown estimation with this relatively simple technique is approximately 8 % across a wide variety of workloads. Figure 6.9 shows the actual versus predicted slowdowns over time, for *astar*, a representative application from among the many applications examined, when it is run alongside three other applications on a simulated 4-core 1-channel system. As we can see, MISE's slowdown estimates track the actual slowdown closely. Extending such simple techniques like MISE to the entire memory and storage system is a promising area of future research in both homogeneous and heterogeneous systems. Devising memory devices and architectures that can support predictability and QoS also appears promising.

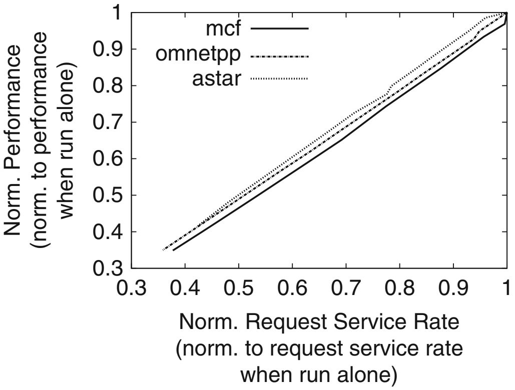


Fig. 6.8 Request service rate vs. performance. Reproduced from [106]

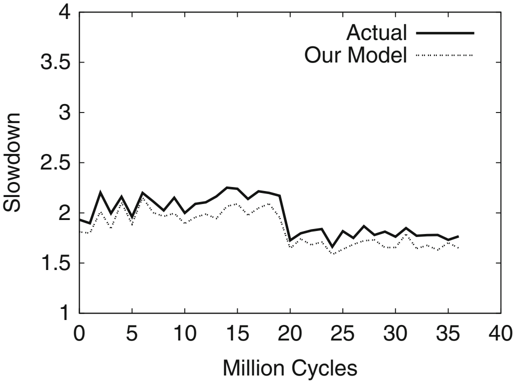


Fig. 6.9 Actual vs. predicted slowdowns. Reproduced from [106]

6.8 Aside: Flash Memory Scaling Challenges

Flash memory, another successful charge-based memory like DRAM, has been commonly employed as part of the storage system. In part of our research, we aim to develop new techniques that overcome reliability and endurance challenges of flash memory to enable its scaling beyond the 20 nm technology generations. To this end, we experimentally measure, characterize, analyze, and model error patterns that occur in existing flash chips, using an experimental flash memory testing and characterization platform [7]. Based on the understanding we develop from our experiments, we aim to develop error management techniques that mitigate the fundamental types of errors that are likely to increase as flash memory scales.

We have recently experimentally characterized complex flash errors that occur at 30–40 nm flash technologies [8], categorizing them into four types: retention errors, program interference errors, read errors, and erase errors. Our characterization shows the relationship between various types of errors and demonstrates empirically using real 3x-nm flash chips that retention errors are the most dominant error type. Our results demonstrate that different flash errors have distinct patterns: retention errors and program interference errors are program/erase-(P/E)-cycle-dependent, memory-location-dependent, and data-value-dependent. Since the observed error patterns are due to fundamental circuit and device behavior inherent in flash memory, we expect our observations and error patterns to also hold in flash memories beyond 30 nm technology.

Based on our experimental characterization results that show that the retention errors are the most dominant errors, we have developed a suite of techniques to mitigate the effects of such errors, called Flash Correct-and-Refresh (FCR) [9]. The key idea is to periodically read each page in flash memory, correct its errors using simple error correcting codes (ECC), and either remap (copy/move) the page to a different location or reprogram it in its original location by recharging the floating gates, before the page accumulates more errors than can be corrected with simple ECC. Our simulation experiments using real I/O workload traces from a variety of file system, database, and search applications show that FCR can provide 46x flash memory lifetime improvement at only 1.5 % energy overhead, with no additional hardware cost.

Recently, we have also experimentally investigated and characterized the threshold voltage distribution of different logical states in MLC NAND flash memory [12]. We have developed new models that can predict the shifts in the threshold voltage distribution based on the number of P/E cycles endured by flash memory cells. Our data shows that the threshold voltage distribution of flash cells that store the same value can be approximated, with reasonable accuracy, as a Gaussian distribution. The threshold voltage distribution of flash cells that store the same value gets distorted as the number of P/E cycles increases, causing threshold voltages of cells storing different values to overlap with each other, which can lead to the incorrect reading of values of some cells as flash cells accumulate P/E cycles. We find that this distortion can be accurately modeled and predicted as an exponential function of the P/E cycles, with more than 95 % accuracy. Such predictive models can aid the design of more sophisticated error correction methods, such as LDPC codes [35], which are likely needed for reliable operation of future flash memories.

We are currently investigating another increasingly significant obstacle to MLC NAND flash scaling, which is the increasing cell-to-cell program interference due to increasing parasitic capacitances between the cells' floating gates. Accurate characterization and modeling of this phenomenon are needed to find effective techniques to combat program interference. In recent work [11], we leverage the *read retry* mechanism found in some flash designs to obtain measured threshold voltage distributions from state-of-the-art 2Y-nm (i.e., 24-20 nm) MLC NAND flash chips. These results are then used to characterize the cell-to-cell program interference under various programming conditions. We show that program interference can be accurately modeled as additive noise following Gaussian-mixture distributions, which can be predicted with 96.8 % accuracy using linear regression models. We use these models to develop and evaluate a read reference voltage prediction technique that reduces the raw flash bit error rate by 64 % and increases the flash lifetime by 30 %. More detail can be found in Cai et al. [11].

Most recently, to improve flash memory lifetime, we have developed a mechanism called Neighbor-Cell Assisted Correction (NAC) [13], which uses the value information of cells in a neighboring page to correct errors found on a page when reading. This mechanism takes advantage of the new empirical observation that identifying the value stored in the immediate-neighbor cell makes it easier to determine the data value stored in the cell that is being read. The key idea is to re-read a flash memory page that fails error correction codes (ECC) with the set of read reference voltage values corresponding to the conditional threshold voltage distribution assuming a neighbor cell value and use the re-read values to correct the cells that have neighbors with that value. Our simulations show that NAC effectively improves flash memory lifetime by 33 % while having no (at nominal lifetime) or very modest (less than 5 % at extended lifetime) performance overhead.

Going forward, we believe more accurate and detailed characterization of flash memory error mechanisms are needed to devise models that can aid the design of more efficient and effective mechanisms to tolerate errors found in sub-20 nm flash memories. A promising direction is the design of predictive models that the system (e.g., the flash controller or system software) can use to proactively estimate the occurrence of errors and take action to prevent the error before it happens. Flash-correct-and-refresh [9], read reference voltage prediction [11], described earlier, are early forms of such predictive error tolerance mechanisms.

6.9 Conclusion

We have described several research directions and ideas to enhance memory scaling via system and architecture-level approaches. A promising approach is the co-design of memory and other system components to enable better system optimization. Enabling better cooperation across multiple levels of the computing stack, including software, microarchitecture, and devices can help scale the memory system by exposing more of the memory device characteristics to higher levels of the system stack such that the latter can tolerate and exploit such characteristics. Finally, heterogeneity in the design of the memory system can help overcome the memory scaling challenges at the device level by enabling better specialization of the memory system and its dynamic adaptation to different demands of various applications. We believe such approaches will become increasingly important and effective as the underlying memory technology nears its scaling limits at the physical level and envision a near future full of innovation in main memory architecture, enabled by the co-design of the system and main memory.

Acknowledgements I would like to thank my PhD students Rachata Ausavarungnirun and Lavanya Subramanian for logistic help in preparing this chapter and earlier versions of it. Many thanks to all my students in my SAFARI research group and collaborators at Carnegie Mellon as well as other universities, whom all contributed to the works outlined in this chapter. Thanks also go to my research group's industrial sponsors over the past 6 years, including AMD, HP Labs, IBM, Intel, Microsoft, Nvidia, Oracle, Qualcomm, Samsung. Some of the research reported here was also partially supported by GSRC, Intel URO Memory Hierarchy Program, Intel Science and Technology Center on Cloud Computing, NIH, NSF, and SRC.

This chapter is a significantly extended and revised version of an invited paper that appeared at the 5th International Memory Workshop [85], which was also presented at MemCon 2013 [86]. Part of the structure of this chapter is based on an evolving set of talks I have delivered at various venues on *Scaling the Memory System in the Many-Core Era* and *Rethinking Memory System Design for Data-Intensive Computing* between 2010–2014, including invited talks at the 2011 International Symposium on Memory Management and ACM SIGPLAN Workshop on Memory System Performance and Correctness [89] and the 2012 DAC More Than Moore Technologies Workshop. Section 6.8 of this article is a condensed and slightly revised version of the introduction of an invited article that appeared in a special issue of the Intel Technology Journal, titled *Error Analysis and Retention-Aware Error Management for NAND Flash Memory* [10].

References

1. Ahn JH, Jeong BH, Kim SH, Chu SH, Cho SK, Lee HJ, et al. Adaptive self refresh scheme for battery operated high-density mobile dram applications. In: Solid-state circuits conference (ASSCC); 2006.
2. Alkan C, Kidd JM, Marques-Bonet T, et al. Personalized copy-number and segmental duplication maps using next-generation sequencing. Nat Genet. 2009;41:1061–7.
3. Atwood, G.: Current and emerging memory technology landscape. In: Flash memory summit; 2011.
4. Ausavarungnirun R, Chang KKW, Subramanian L, Loh GH, Mutlu O. Staged memory scheduling: achieving high performance and scalability in heterogeneous systems. In: ACM SIGARCH computer architecture news (ISCA); 2012.
5. Bloom BH. Space/time trade-offs in hash coding with allowable errors. Commun ACM. 1970;13(7):422–6.
6. Bryant R. Data-intensive supercomputing: The case for DISC. CMU CS Technical Report 07-128; 2007.
7. Cai Y, Haratsch EF, McCartney M, Mai K. FPGA-based solid-state drive prototyping platform. In: IEEE 19th annual international symposium on field-programmable custom computing machines (FCCM); 2011.
8. Cai Y, Haratsch EF, Mutlu O, Mai K. Error patterns in MLC NAND flash memory: measurement, characterization, and analysis. In: Design automation & test in Europe conference & exhibition (DATE); 2012.
9. Cai Y, Yalcin G, Mutlu O, Haratsch EF, Cristal A, Unsal OS. Flash correct-and-refresh: retention-aware error management for increased flash memory lifetime. In: IEEE 30th international conference on computer design (ICCD); 2012.
10. Cai Y, Yalcin G, Mutlu O, Haratsch EF, Cristal A, Unsal O, et al. Error analysis and retention-aware error management for nand flash memory. Intel Technol J. 2013;17(1):140.
11. Cai Y, Mutlu O, Haratsch EF, Mai K. Program interference in MLC NAND flash memory: characterization, modeling, and mitigation. In: IEEE 31st international conference on computer design (ICCD); 2013.
12. Cai Y, Haratsch EF, Mutlu O, Mai K. Threshold voltage distribution in MLC NAND flash memory: characterization, analysis and modeling. In: Proceedings of the conference on design, automation and test in Europe (DATE); 2013.
13. Cai Y, Yalcin G, Mutlu O, Haratsch EF, Unsal O, Cristal A, et al. Neighbor-cell assisted error correction for MLC NAND flash memories. In: The 2014 ACM international conference on measurement and modeling of computer systems (SIGMETRICS); 2014.
14. Chang KKW, Ausavarungnirun R, Fallin C, Mutlu O. HAT: Heterogeneous adaptive throttling for on-chip networks. In: IEEE 24th international symposium on computer architecture and high performance computing (SBAC-PAD); 2012
15. Chang KKW, Lee D, Chishti Z, Alameldeen AR, Wilkerson C, Kim Y, et al. Improving DRAM performance by parallelizing refreshes with accesses. In: High performance computer architecture (HPCA); 2014.
16. Chatterjee N , Shevgoor M, Balasubramonian R, Davis A, Fang Z, Illikkal R, et al. Leveraging heterogeneity in DRAM main memories to accelerate critical word access. In: 45th annual IEEE/ACM international symposium on Microarchitecture (MICRO); 2012.
17. Chen E, Apalkov D, Diao Z, Driskill-Smith A, Druist D, Lottis D, et al. Advances and future prospects of spin-transfer torque random access memory. IEEE Trans Magn. 2010;46(6):1873–8.
18. Chhabra S, Solihin Y. i-nvmm: a secure non-volatile main memory system with incremental encryption. In: 38th annual international symposium on computer architecture (ISCA); 2011.
19. Chung E, Milder PA, Hoe JC, Mai K. Single-chip heterogeneous computing: Does the future include custom logic, FPGAs, and GPUs? In: Proceedings of the 2010 43rd annual IEEE/ACM international symposium on microarchitecture (MICRO); 2010.

20. Condit J, Nightingale EB, Frost C, Ipek E, Lee B, Burger D, et al. Better I/O through byte-addressable, persistent memory. In: Proceedings of the ACM SIGOPS 22nd symposium on operating systems principles (SOSP); 2009.
21. Craeynest VK, Jaleel A, Eeckhout L, Narvaez P, Emer J. Scheduling heterogeneous multi-cores through performance impact estimation (PIE). In: ACM SIGARCH computer architecture news (ISCA); 2012.
22. Das R, Mutlu O, Moscibroda T, Das CR. Application-aware prioritization mechanisms for on-chip networks. In: 42nd annual IEEE/ACM international symposium on microarchitecture (MICRO); 2009.
23. Das R, Mutlu O, Moscibroda T, Das CR. Aergia: Exploiting packet latency slack in on-chip networks. In: ACM SIGARCH computer architecture news (ISCA); 2010.
24. Das R, Ausavarungnirun R, Mutlu O, Kumar A, Azimi, M. Application-to-core mapping policies to reduce memory system interference in multi-core systems. In: IEEE 19th international symposium on high performance computer architecture (HPCA); 2013.
25. David H, Fallin C, Gorbatov E, Hanebutte UR, Mutlu O. Memory power management via dynamic voltage/frequency scaling. In: Proceedings of the 8th ACM international conference on autonomic computing (ICAC); 2011.
26. Dean J, Barroso LA. The tail at scale. Commun ACM. 2013;56(2):74–80.
27. Deng Q, Meisner D, Ramos L, Wenisch TF, Bianchini R.MemScale: active low-power modes for main memory. In: ACM SIGPLAN notices (ASPLOS); 2011.
28. Dhiman G, Ayoub R, Rosing T PDRAM: A hybrid PRAM and DRAM main memory system. In: 46th ACM/IEEE design automation conference (DAC); 2009.
29. Dong X, Muralimanohar N, Jouppi N , Kaufmann R, Xie Y. Leveraging 3D PCRAM technologies to reduce checkpoint overhead for future exascale systems. In: SC; 2009.
30. Ebrahimi E, Lee CJ, Mutlu O, Patt YN. Fairness via source throttling: a configurable and high-performance fairness substrate for multi-core memory systems. In: ACM sigplan notices (ASPLOS); 2010.
31. Ebrahimi E, Miftakhutdinov R, Fallin C, Lee CJ, Joao JA, Mutlu O, et al. Parallel application memory scheduling. In: Proceedings of the 44th Annual IEEE/ACM International symposium on microarchitecture (MICRO); 2011.
32. Ebrahimi E, Lee CJ, Mutlu O, Patt YN. Prefetch-aware shared-resource management for multi-core systems. In: ACM SIGARCH computer architecture news (ISCA); 2011.
33. Ekman, M.: A robust main-memory compression scheme. In: ACM SIGARCH computer architecture news (ISCA); 2005.
34. Eyerman S, Eeckhout L. Modeling critical sections in amdahl's law and its implications for multicore design. In: ACM SIGARCH computer architecture news (ISCA); 2010.
35. Gallager R. Low density parity check codes Cambridge: MIT Press; 1963.
36. Grot B, Keckler SW, Mutlu O. Preemptive virtual clock: A flexible, efficient, and cost-effective qos scheme for networks-on-chip. In: Proceedings of the 42nd annual IEEE/ACM international symposium on microarchitecture (MICRO); 2009.
37. Grot B, Hestness J, Keckler SW, Mutlu O. Kilo-NOC: A heterogeneous network-on-chip architecture for scalability and service guarantees. In: ACM SIGARCH computer architecture news (ISCA); 2011.
38. Hong S. Memory technology trend and future challenges. In: International electron devices meeting (IEDM); 2010.
39. Hybrid memory consortium. 2012. http://www.hybridmemorycube.org.
40. International technology roadmap for semiconductors (ITRS); 2011.
41. Ipek E, Mutlu O, Martinez JF, Caruana R. Self-optimizing memory controllers: a reinforcement learning approach. In: ACM SIGARCH computer architecture news (ISCA); 2008.
42. Isen C, John LK. Eskimo: Energy savings using semantic knowledge of inconsequential memory occupancy for dram subsystem. In: 42nd annual IEEE/ACM international symposium on microarchitecture (MICRO); 2009.
43. Iyer R. CQoS: a framework for enabling QoS in shared caches of CMP platforms. In: Proceedings of the 18th annual international conference on supercomputing (ICS); 2004.

44. Iyer R, Zhao L, Guo F, Illikkal R, Makineni S, Newell D, et al. QoS policies and architecture for cache/memory in cmp platforms. In: SIGMETRICS performance evaluation review; 2007.
45. Joao JA, Suleman MA, Mutlu O, Patt YN. Bottleneck identification and scheduling in multithreaded applications. In: ASPLOS; 2012.
46. Joao JA, Suleman MA, Mutlu O, Patt YN. Utility-based acceleration of multithreaded applications on asymmetric cmps. In: ACM SIGARCH computer architecture news (ISCA); 2013.
47. Jog A, Kayiran O, Mishra AK, Kandemir MT, Mutlu O, Iyer R, et al. Orchestrated scheduling and prefetching for GPGPUs. In: ACM SIGARCH computer architecture news (ISCA); 2013.
48. Jog A, Kayiran O, Chidambaram Nachiappan N, Mishra AK, Kandemir MT, Mutlu O, et al. OWL: Cooperative thread array aware scheduling techniques for improving GPGPU performance. In: ASPLOS; 2013.
49. Johnson TL, Merten MC, Hwu, WMW. Run-time spatial locality detection and optimization. In: Proceedings of the 30th annual ACM/IEEE international symposium on microarchitecture (MICRO); 1997.
50. Kang U, Yu HS, Park C, Zheng H, Halbert J, Bains K, et al. Co-architecting controllers and DRAM to enhance DRAM process scaling. In: The memory forum; 2014.
51. Khan S, Lee D, Kim Y, Alameldeen AR, Wilkerson C, Mutlu O. The efficacy of error mitigation techniques for DRAM retention failures: a comparative experimental study. In: ACM international conference on measurement and modeling of computer systems (SIGMETRICS); 2014.
52. Kim J, Papaefthymiou MC. Dynamic memory design for low data-retention power. In: PATMOS; 2000.
53. Kim K. Future memory technology: challenges and opportunities. In: International Symposium on (VLSI-TSA); 2008.
54. Kim K, Lee J. A new investigation of data retention time in truly nanoscaled DRAMs. IEEE Electron Device Lett. 2009;30(8):846–8.
55. Kim Y, Han D, Mutlu O, Harchol-Balter M. ATLAS: a scalable and high-performance scheduling algorithm for multiple memory controllers. In: IEEE 16th international symposium on high performance computer architecture (HPCA); 2010.
56. Kim Y, Papamichael M, Mutlu O, Harchol-Balter M. Thread cluster memory scheduling: exploiting differences in memory access behavior. In: 43rd annual IEEE/ACM international symposium microarchitecture (MICRO); 2010.
57. Kim Y, Seshadri V, Lee D, Liu J, Mutlu O. A case for subarray-level parallelism (SALP) in DRAM. In: ACM SIGARCH computer architecture news (ISCA); 2012.
58. Kim Y, Daly R, Kim J, Fallin C, Lee JH, Lee D, et al. Flipping bits in memory without accessing them: an experimental study of DRAM disturbance errors. In: ACM SIGARCH computer architecture news (ISCA); 2014.
59. Koh, Y.: NAND Flash Scaling Beyond 20nm. In: IMW; 2009.
60. Kultursay E, Kandemir M, Sivasubramaniam A, Mutlu O. Evaluating STT-RAM as an energy-efficient main memory alternative. In: IEEE international symposium on performance analysis of systems and software (ISPASS); 2013.
61. Kumar S, Wilkerson, C. Exploiting spatial locality in data caches using spatial footprints. In: ACM SIGARCH computer architecture news (ISCA); 1998.
62. Lee BC, Ipek E, Mutlu O, Burger D. Architecting phase change memory as a scalable DRAM alternative. In: ACM SIGARCH computer architecture news (ISCA); 2009.
63. Lee BC, Ipek E, Mutlu O, Burger, D. Phase change memory architecture and the quest for scalability. CommunACM. 2010;53(7):99–106.
64. Lee BC, Zhou P, Yang J, Zhang Y, Zhao B, Ipek E. Phase change technology and the future of main memory. IEEE Micro (Top Picks Issue). 30(1); 2010.
65. Lee CJ, Mutlu O, Narasiman V, Patt YN. Prefetch-aware DRAM controllers. In: Proceedings of the 41st annual IEEE/ACM international symposium on microarchitecture (MICRO); 2008.

66. Lee CJ, Narasiman V, Mutlu O, Patt YN. Improving memory bank-level parallelism in the presence of prefetching. In: Proceedings of the 42nd annual IEEE/ACM international symposium on microarchitecture (MICRO); 2009.
67. Lee CJ, Narasiman V, Ebrahimi E, Mutlu O, Patt YN DRAM-aware last-level cache writeback: Reducing write-caused interference in memory systems. Technical Report TR-HPS-2010-002, HPS; 2010.
68. Lee D, Kim Y, Seshadri V, Liu J, Subramanian L, Mutlu O. Tiered-latency DRAM: A low latency and low cost DRAM architecture. In: IEEE 19th international symposium on high performance computer architecture (HPCA); 2013.
69. Lefurgy C, Rajamani K, Rawson F, Felter W, Kistler M, Keller TW. Energy management for commercial servers. In: IEEE computer; 2003.
70. Lim K, Chang J, Mudge T, Ranganathan P, Reinhardt SK, Wenisch TF. Disaggregated memory for expansion and sharing in blade servers. In: ACM SIGARCH computer architecture news (ISCA); 2009.
71. Liu J, Jaiyen B, Veras R, Mutlu O. RAIDR: Retention-aware intelligent DRAM refresh. In: ACM SIGARCH computer architecture news (ISCA); 2012.
72. Liu J, Jaiyen B, Kim Y, Wilkerson C, Mutlu O. An experimental study of data retention behavior in modern DRAM devices: Implications for retention time profiling mechanisms. In: ACM SIGARCH computer architecture news (ISCA); 2013.
73. Liu S, Pattabiraman K, Moscibroda T, Zorn BG. Flikker: saving dram refresh-power through critical data partitioning. In: ACM SIGPLAN notices (ASPLOS); 2011.
74. Loh G. 3D-stacked memory architectures for multi-core processors. In: ACM SIGARCH computer architecture news (ISCA); 2008.
75. Luo Y, Govindan S, Sharma B, Santaniello M, Meza J, Kansal A, et al. Characterizing application memory error vulnerability to optimize data center cost via heterogeneous-reliability memory. In: DSN; 2014.
76. Maislos A. A new era in embedded flash memory. In: FMS; 2011.
77. Mandelman JA, Dennard RH, Bronner GB, DeBrosse JK, Divakaruni R, Li Y, et al. Challenges and future directions for the scaling of dynamic random-access memory (DRAM). In: IBM J Res Dev. 2002;46(2.3):187–212.
78. Meza J, Li J, Mutlu O. A case for small row buffers in non-volatile main memories. In: IEEE 30th international conference on computer design (ICCD); 2012.
79. Meza J, Chang J, Yoon H, Mutlu O, Ranganathan P. Enabling efficient and scalable hybrid memories using fine-granularity DRAM cache management. IEEE computer architecture letters (CAL); 2012.
80. Meza J, Luo Y, Khan S, Zhao J, Xie Y, Mutlu O. A case for efficient hardware-software cooperative management of storage and memory. In: Proceedings of 5th workshop on energy efficient design (WEED); 2013.
81. Moscibroda T, Mutlu O. Memory performance attacks: Denial of memory service in multi-core systems. In: Proceedings of 16th USENIX security symposium on USENIX security symposium (USENIX); 2007.
82. Moscibroda T, Mutlu O. Distributed order scheduling and its application to multi-core DRAM controllers. In: Proceedings of the 27th ACM symposium on principles of distributed computing (PODC); 2008.
83. Muralidhara S, Subramanian L, Mutlu O, Kandemir M, Moscibroda T. Reducing memory interference in multi-core systems via application-aware memory channel partitioning. In: Proceedings of the 44th annual IEEE/ACM international symposium on microarchitecture (MICRO); 2011.
84. Mutlu O. Asymmetry everywhere (with automatic resource management). In: CRA workshop on advance computer architecture research; 2010.
85. Mutlu, O. Memory scaling: a systems architecture perspective. In: 5th IEEE international memory workshop (IMW); 2013.
86. Mutlu O. Memory scaling: a systems architecture perspective. In: MemCon; 2013.

87. Mutlu O, Moscibroda T. Stall-time fair memory access scheduling for chip multiprocessors. In: Proceedings of the 40th annual IEEE/ACM international symposium on microarchitecture (MICRO); 2007.
88. Mutlu O, Moscibroda T. Parallelism-aware batch scheduling: Enhancing both performance and fairness of shared DRAM systems. In: ACM SIGARCH computer architecture news (ISCA); 2008.
89. Mutlu O. Memory systems in the many-core era: Challenges, opportunities, and solution directions. In: ACM SIGPLAN notices (ISMM); 2011. http://users.ece.cmu.edu/~omutlu/pub/onur-ismm-mspc-keynote-june-5-2011-short.pptx.
90. Nair PJ, Kim DH, Qureshi MK.ArchShield: architectural framework for assisting DRAM scaling by tolerating high error rates. In: Proceedings of the 40th annual international symposium on computer architecture (ISCA); 2013.
91. Nychis G, Fallin C, Moscibroda T, Mutlu O. Next generation on-chip networks: what kind of congestion control do we need? In: Proceedings of the 9th ACM SIGCOMM workshop on hot topics in networks (HotNets); 2010.
92. Nychis G, Fallin C, Moscibroda T, Mutlu O, Seshan S. On-chip networks from a networking perspective: congestion and scalability in many-core interconnects. In: ACM SIGCOMM computer communication review; 2012.
93. Ohsawa T, Kai K, Murakami K. Optimizing the DRAM refresh count for merged DRAM/logic LSIs. In: Proceedings of the 1998 international symposium on Low power electronics and design (ISLPED); 1998.
94. Pekhimenko G, Seshadri V, Mutlu O, Mowry TC, Gibbons PB, Kozuch MA. Base-delta-immediate compression: a practical data compression mechanism for on-chip caches. In: Proceedings of the 21st ACM international conference on parallel architectures and compilation techniques (PACT); 2012.
95. Pekhimenko G, Mowry TC, Mutlu O. Linearly compressed pages: a main memory compression framework with low complexity and low latency. In: Proceedings of the 21st international conference on parallel architectures and compilation technique (MICRO); 2013.
96. Qureshi MK, Suleman MA, Patt YN. Line distillation: Increasing cache capacity by filtering unused words in cache lines. In: IEEE 13th international symposium on high performance computer architecture (HPCA); 2007.
97. Qureshi MK, Karidis J, Franceschini M, Srinivasan V, Lastras L, Abali B. Enhancing lifetime and security of phase change memories via start-gap wear leveling. In: Proceedings of the 42nd annual IEEE/ACM international symposium on microarchitecture (MICRO); 2009.
98. Qureshi MK, Srinivasan V, Rivers JA. Scalable high performance main memory system using phase-change memory technology. In: ACM SIGARCH computer architecture news (ISCA); 2009.
99. Ramos LE, Gorbatov E, Bianchini R. Page placement in hybrid memory systems. In: Proceedings of the international conference on supercomputing (ICS); 2011.
100. Raoux S, Burr GW, Breitwisch MJ, Rettner CT, Chen YC, Shelby RM, et al. Phase-change random access memory: a scalable technology. IBM J Res Dev. 2008;52:465–79.
101. Seshadri V, Mutlu O, Kozuch MA, Mowry TC. The evicted-address filter: A unified mechanism to address both cache pollution and thrashing. In: Proceedings of the 21st international conference on parallel architectures and compilation techniques (PACT); 2012.
102. Seshadri V, Kim Y, Fallin C, Lee D, Ausavarungnirun R, Pekhimenko G. RowClone: Fast and efficient In-DRAM copy and initialization of bulk data. MICRO; 2013.
103. Seshadri V, Bhowmick A, Mutlu O, Gibbons PB, Kozuch MA, Mowry TC. The dirty-block index. In: ACM SIGARCH computer architecture news (ISCA); 2014.
104. Song NH, Woo DH, Lee HHS. Security refresh: prevent malicious wear-out and increase durability for phase-change memory with dynamically randomized address mapping. In: ACM SIGARCH computer architecture news (ISCA); 2010.
105. Stuecheli J, Kaseridis D, Daly D, Hunter HC, John LK. The virtual write queue: Coordinating DRAM and last-level cache policies. In: ACM SIGARCH computer architecture news (ISCA-37); 2010.

106. Subramanian L, Seshadri V, Kim Y, Jaiyen B, Mutlu O. MISE: Providing performance predictability and improving fairness in shared main memory systems. In: IEEE 19th international symposium on high performance computer architecture (HPCA); 2013.
107. Suleman MA, Mutlu O, Qureshi MK, Patt YN. Accelerating critical section execution with asymmetric multi-core architectures. In: ASPLOS; 2009.
108. Tang L, Mars J, Vachharajani N, Hundt R, Soffa ML. The impact of memory subsystem resource sharing on datacenter applications. In: ACM SIGARCH computer architecture news (ISCA); 2011.
109. Treangen T, Salzberg S. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 2012;13(1):36–46.
110. Udipi AN, Muralimanohar N, Chatterjee N, Balasubramonian R, Davis A, Jouppi NP. Rethinking DRAM design and organization for energy-constrained multi-cores. In: ACM SIGARCH computer architecture news (ISCA); 2010.
111. Udipi AN, Muralimanohar N, Balasubramonian R, Davis A, Jouppi NP. Combining memory and a controller with photonics through 3d-stacking to enable scalable and energy-efficient systems. In: ACM SIGARCH computer architecture news (ISCA); 2011.
112. Venkatesan RK, Herr S, Rotenberg E. Retention-aware placement in DRAM (RAPID): Software methods for quasi-non-volatile DRAM. In: IEEE 12th international symposium on high performance computer architecture (HPCA); 2006.
113. Wong HSP, Raoux S, Kim S, et al. Phase change memory. In: Proceedings of the IEEE; 2010.
114. Wong HSP, Lee HY, Yu S, et al. Metal-oxide rram. In: Proceedings of the IEEE metal-oxide RRAM; 2012.
115. Xin H, Lee D, Hormozdiari F, Yedkar S, Mutlu O, Alkan C. Accelerating read mapping with FastHASH. BMC Genomics. 2013;14(S13).
116. Yang J, Zhang Y, Gupta R. Frequent value compression in data caches. In: Proceedings of the 33rd annual ACM/IEEE international symposium on microarchitecture (MICRO-33); 2000.
117. Yoon DH, Jeong MK, Erez M. Adaptive granularity memory systems: A tradeoff between storage efficiency and throughput. In: ACM SIGARCH computer architecture news (ISCA); 2011.
118. Yoon DH, Jeong MK, Sullivan M, Erez M. The dynamic granularity memory system. In: ACM SIGARCH computer architecture news (ISCA); 2012.
119. Yoon H, Meza J, Ausavarungnirun R, Harding RA, Mutlu O. Row buffer locality aware caching policies for hybrid memories. In: IEEE 30th international conference on computer design (ICCD); 2012.
120. Yoon H, Muralimanohar N, Meza J, Mutlu O, Jouppi NP. Data mapping and buffering in multi-level cell memory for higher performance and energy efficiency. CMU SAFARI Technical Report; 2013.

Chapter 7
Nano-Photonic Networks-on-Chip for Future Chip Multiprocessors

Cheng Li, Paul V. Gratz, and Samuel Palermo

Abstract To meet energy-efficient performance demands, the computing industry has moved to parallel computer architectures, such as chip-multi-processors (CMPs), internally interconnected via networks-on-chip (NoC) to meet growing communication needs. Achieving scaling performance as core counts increase to the hundreds in future CMPs, however, will require high performance, yet energy-efficient interconnects. Silicon nanophotonics is a promising replacement for electronic on-chip interconnect due to its high bandwidth and low latency, however, prior techniques have required high static power for the laser and ring thermal tuning. We propose a novel nano-photonic NoC architecture, LumiNOC, optimized for high performance and power-efficiency. This paper makes three primary contributions: a novel, nanophotonic architecture which partitions the network into subnets for better efficiency; a purely photonic, in-band, distributed arbitration scheme; and a channel sharing arrangement utilizing the same waveguides and wavelengths for arbitration as data transmission. In a 64-node NoC under synthetic traffic, LumiNOC enjoys 50 % lower latency at low loads and ∼40 % higher throughput per Watt on synthetic traffic, versus other reported photonic NoCs. LumiNOC reduces latencies ∼40 % versus an electrical 2D mesh NoCs on the PARSEC shared-memory, multithreaded benchmark suite.

7.1 Introduction

Parallel architectures, such as single-chip multiprocessors (CMPs), have emerged to address power consumption and performance scaling issues in current and future VLSI process technology. Networks-on-chip (NoCs), have concurrently emerged to serve as a scalable alternative to traditional, bus-based interconnection between processor cores. Conventional NoCs in CMPs use wide, point-to-point electrical

C. Li
HP Laboratories, 1501 Page Mill Rd., Palo Alto, CA 94304, USA
e-mail: cheng.li6@hp.com

P.V. Gratz (✉) • S. Palermo
Texas A&M University, 3128 TAMU, College Station, TX 77843, USA
e-mail: pgratz@gratz1.com; spalermo@ece.tamu.edu

R.O. Topaloglu (ed.), *More than Moore Technologies for Next Generation Computer Design*, DOI 10.1007/978-1-4939-2163-8_7

links to relay cache-lines between private mid-level and shared last-level processor caches [1]. Electrical on-chip interconnect, however, is severely limited by power, bandwidth and latency constraints due to high-frequency loss of electrical traces and crosstalk from adjacent signals. These constraints are placing practical limits on the viability of future CMP scaling. For example, the efficiency of current state-of-the-art NoCs with simple CMOS inverter-based repeaters is near 2 pJ/bit [2], allowing for only near 1TB/s throughput with a typical 20 % allowance from the total 100 W processor power budget. Power in electrical interconnects has been reported as high as 12.1 W for a 48-core, 2D-mesh CMP at 2 GHz [1], a significant fraction of the system's power budget. Furthermore, achieving application performance which scales with the number of cores requires extremely low latency communication to reduce the impact of serialization points within the code. However, communication latency in a typical NoC connected multiprocessor system increases rapidly as the number of nodes increases [3]. Worst-case, no-load communication latencies in a 64-node multi-core chip can reach as high as 50 cycles, nearly 1/2 the latency of an off-chip memory access. The communication requirements of future processing systems makes traditional electrical on-chip networks prohibitive for future transformative extrascale computers.

Recently, monolithic silicon photonics have been proposed as a scalable alternative to meet future many-core systems bandwidth demands, by leveraging high-speed photonic devices [4–6], THz-bandwidth waveguides [7,8], and immense bandwidth-density via wavelength-division-multiplexing (WDM) [9, 10]. Several NoC architectures leveraging the high bandwidth of silicon photonics have been proposed. These works can be categorized into two general types: (1). Hybrid optical/electrical interconnect architecture [11–14], in which a photonic packet-switched network and an electronic circuit-switched control network are combined to respectively deliver large size data messages and short control messages; (2). Crossbar or Clos architectures, in which the interconnect is fully photonic [15–23]. Although these designs provide high and scalable bandwidth, they either suffer from relatively high latency due to the electrical control circuits for photonic path setup, or significant power/hardware overhead due to significant over-provisioned photonic channels. In future latency and power constrained CMPs, these characteristics will hobble the utility of photonic interconnect.

In this chapter, we propose LumiNOC [24], a novel PNoC architecture which addresses power and resource overheads due to channel over-provisioning, while reducing latency and maintaining high bandwidth in CMPs. LumiNoC utilizes integrated silicon waveguides that provide the potential to overcome electrical interconnect bottlenecks and greatly improve data transfer efficiency due to their flat channel loss over a wide frequency range and also relatively small crosstalk and electromagnetic noise [25]. By combining multiple data channels on a single waveguide via wavelength-division-multiplexing (WDM), LumiNoC greatly improves bandwidth density. Area-compact and energy-efficient silicon ring resonators are employed as the optical modulator and drop filter in the integrated WDM link. Silicon ring resonator modulators/filters offer advantages of small size, relative to Mach-Zehnder modulators [26], and increased filter functionality, relative to electro-absorption modulators [27]. The LumiNOC architecture makes three contributions: First, instead of conventional, globally distributed, photonic channels, requiring high laser power, we propose a novel channel sharing arrangement composed of sub-sets of cores in photonic subnets. Second, we propose a novel, purely photonic, distributed arbitration mechanism, dynamic channel scheduling, which achieves extremely low-latency without degrading throughput. Third, our photonic network architecture leverages the same wavelengths for channel arbitration and parallel data transmission, allowing efficient utilization of the photonic resources and lowering static power consumption. We show in a 64-node implementation that LumiNOC enjoys 50 % lower latency at low loads and ∼40 % higher throughput per Watt on synthetic traffic versus previous PNoCs. Furthermore, LumiNOC reduces latency ∼40 % versus an electrical 2D mesh NoCs on PARSEC shared-memory, multithreaded benchmark workloads.

7.2 Silicon Photonic Devices

Figure 7.1 shows a typical silicon photonics WDM link, where multiple wavelengths (λ_1–λ_4) generated by an off-chip continuous-wave (CW) laser are coupled into a silicon waveguide via an optical coupler. At transmit side, ring modulators insert data onto a specific wavelength through electro-optical modulation. These modulated optical signals propagate through the waveguide and arrive at the receiver side where ring filters drop the modulated optical signals of a specific wavelength at a receiver channel with photodetectors (PD) that convert the signals back to the electrical domain.

7.2.1 *Laser Source*

Laser source can either be a distributed feedback (DFB) laser bank [28], which consists of an array of DFB laser diodes, or a comb laser [29], which is able to

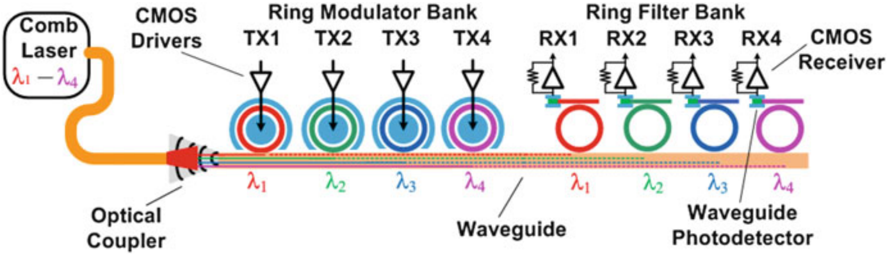


Fig. 7.1 Silicon ring resonator-based wavelength-division-multiplexing (WDM) link

generate multiple wavelengths simultaneously. Implementing a DFB laser bank for dense WDM (DWDM) photonic interconnects (e.g. 64 wavelengths) is quite challenging due to area and power budget constraints. This motivates a single broad-spectrum comb laser source, such as InAs/GaAs quantum dot comb lasers which can generate a large number of wavelengths in the 1,100–1,320 nm spectral range with typical channel spacing of 50–100 GHz and optical power of 0.2–1 mW per channel [29].

7.2.2 *Microring Resonators (MRR)*

MRRs can serve as either optical modulators for sending data or as filters for dropping and receiving data from an on-chip photonic network. A basic silicon ring modulator consists of a straight waveguide coupled with a circular waveguide with diameters on the order of tens of micrometers, as shown in Fig. 7.2a. The two terminal device contains an input port, where the light source is coupled into, and a through port, where the modulated optical signal is coupled out. When the ring circumference equals an integer number of an optical wavelength, called the resonance condition, most of the input light is coupled into the circular waveguide and only a small amount of light can be observed at the through port. As a result, the through port spectrum displays a notch-shaped characteristic, shown in Fig. 7.2b. This resonance can be shifted by changing the effective refractive index of the waveguide through the free-carrier plasma dispersion effect [30] to implement the optical modulation. For example, the ring modulator exhibits low optical output power levels at the through port when the resonance is aligned well with the laser wavelength, while high optical power levels are displayed when the resonance shifts

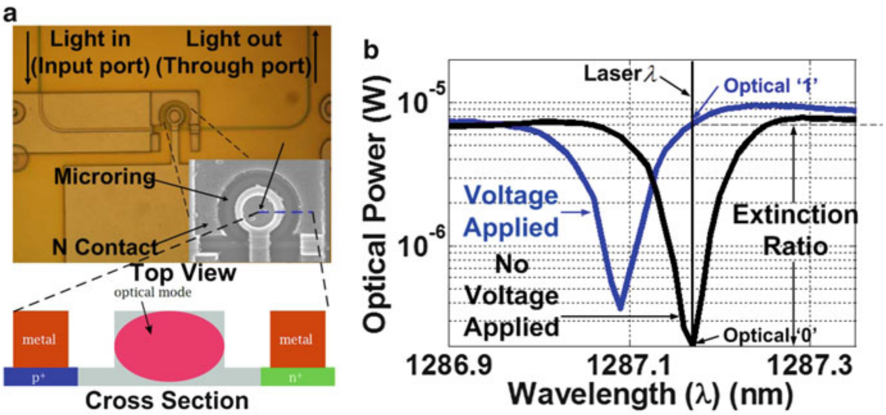


Fig. 7.2 (**a**) Top and cross section views of carrier-injection silicon ring resonator modulator, (**b**) optical spectrum at through port

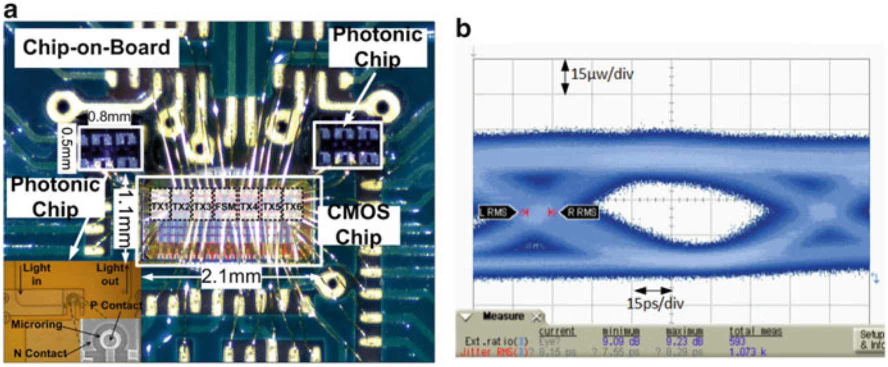


Fig. 7.3 (**a**) Optical transmitter circuit prototype bonded for optical testing, (**b**) measured ring modulator 9 Gb/s optical eye diagram

to a shorter wavelength (blue-shifts) due to the increase in the waveguide carrier density lowering the effective refractive index.

Two common implementations of silicon ring resonator modulators include carrier-injection devices [31], with an embedded p-i-n junction that is side-coupled with the circular waveguide and operating primarily in forward-bias, and carrier-depletion devices [32], with only a p-n junction side-coupled and operating primarily in reverse-bias. Although a depletion ring generally achieves higher modulation speeds relative to a carrier-injection ring due to the ability to rapidly sweep the carriers out of the junction, its modulation depth is limited due to the relatively low doping concentration in the waveguide to avoid excessive optical loss. In contrast, carrier-injection ring modulators can provide large refractive index changes and high modulation depths, but are limited by the relatively slow carrier dynamics of forward-biased p-i-n junctions. Normally, this speed limitation can be alleviated with modulation and/or equalization techniques (e.g. pre-emphasis [33]).

An example of a carrier-injection ring modulator is the 5 μm diameter device [34] shown in Fig. 7.3a, which was fabricated by HP Labs and exhibits a quality factor[1] of ~9,000. Here a chip-on-board test setup is utilized, with a 65 nm CMOS driver [31] wire-bonded to silicon ring resonator chips for optical signal characterization. The measured optical eye diagram of this prototype is show in Fig. 7.3b. It achieves an extinction ratio[2] of 9.2 dB at a modulation speed of 9 Gb/s. The modulation efficiency is 500 fJ/bit, including the electrical driver power. Adopting advanced CMOS processes (e.g. 16 nm CMOS) and photonics integration techniques (e.g. flip-chip bonding or 3D integration) will further improve the optical

[1] Quality factor characterizes a resonator's bandwidth relative to its center frequency. Higher Q indicates a lower rate of energy loss relative to the stored energy of the resonator.

[2] Extinction ratio is the ratio of two optical power levels of a modulated optical signal, expressed in dB.

modulation speed and energy efficiency. This provides strong motivation to leverage this photonic I/O architecture in a WDM system with multiple $\sim$10 Gb/s channels on a single waveguide.

However, one important issue with MRR devices is their resonance wavelength's sensitivity to temperature variation, necessitating tuning to stabilize the ring to resonate at the working wavelength. A commonly proposed resonance wavelength tuning technique is to adjust the device's temperature with a resistor implanted close to the photonic device to heat the waveguide, thus changing the refractive index [35, 36]. Thermal tuning efficiencies near 10–15 μW/GHz have been demonstrated using approaches such as substrate removal and transfer for an SOI process [37] and deep-trench isolation for a bulk CMOS process [36]. Superior efficiencies in the 1.7–2.9 μW/GHz have been achieved with localized substrate removal or undercutting [38, 39], but this comes at the cost of complex processing steps. One potential issue with this approach is that the tuning speed, which is limited by the device thermal time constant ($\sim$ms), may necessitate long calibration times. Compared with the heater-based tuning approaches, a bias-based tuning method for carrier-injection rings has advantages of fast tuning speed and flexibility in the tuning direction, while displaying comparable tuning efficiency. A recent bias-based tuning scheme was reported with a power efficiency of 6.8 μW/GHz, which includes the power of the tuning loop circuitry [31].

7.2.3 Silicon Waveguides

In photonic on-chip networks, silicon waveguides are used to carry the optical signals. In order to achieve higher aggregated bandwidth, multiple wavelengths are placed into a single waveguide in a wavelength-division-multiplexing (WDM) fashion. In this work, silicon nitride waveguides are assumed to be the primary transport strata. Similar to electrical wires, silicon nitride waveguides can be deployed into multiple strata to eliminate in-plane waveguide crossing, thus reducing the optical power loss [40].

7.2.4 Three-Dimensional Integration

In order to optimize system performance and efficiently utilize the chip area, three-dimensional integration (3DI) is emerging for the integration of silicon nanophotonic devices with conventional CMOS electronics. In 3DI, the silicon photonic on-chip networks are fabricated into a separate silicon-on-insulator (SOI) die or layer with a thick layer of buried oxide (BOX) that acts as bottom cladding to prevent light leakage into the substrate. This photonic layer stacks above the electrical layers containing the compute tiles.

7.2.5 4-Tile Photonic Crossbar Example

Figure 7.4 shows a small CMP with 4 compute tiles interconnected by a fully connected crossbar PNoC. Each tile consists of a processor core, private caches, a fraction of the shared last-level cache, and a router connecting it to the photonic network. The photonic channel connecting the nodes is shown as being composed of MMRs (small circles), integrated photodetectors [6] and silicon waveguides [7,8] (black lines connecting the circles). Transceivers (small triangles) mark the boundary between the electrical and photonic domain.

The simple crossbar architecture is implemented by provisioning four send channels, each utilizing the same wavelength in four waveguides, and four receive channels by monitoring four wavelengths in a single waveguide. Although this straightforward structure provides strictly non-blocking connectivity, it requires a large number of transceivers O(r^2) and long waveguides crossing the chip, where r is the crossbar radix, thus this style of crossbar is not scalable to a significant number of nodes. Researchers have proposed a number of PNoC architectures more scalable than fully connected crossbars, as described below.

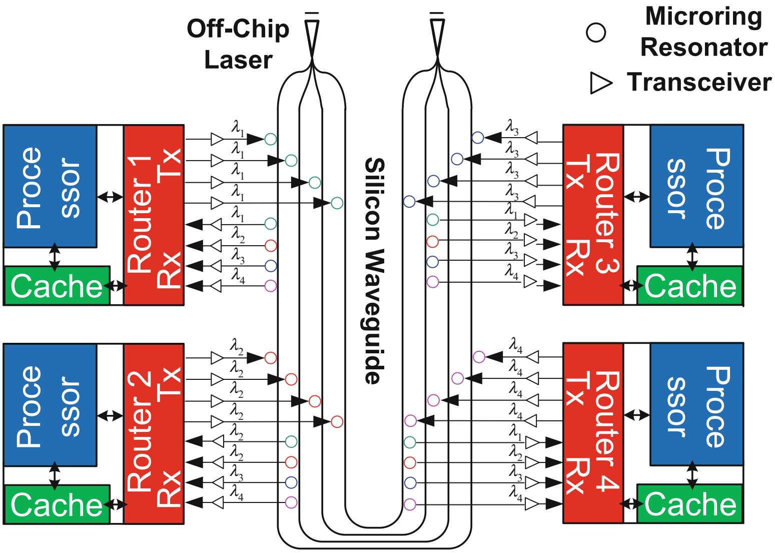


Fig. 7.4 Four-node fully connected photonic crossbar

7.3 Photonic Network-on-Chip Architecture Survey

Many PNoC architectures have been proposed which may be broadly categorized into four basic architectures: (1) Electrical-photonic (2) Crossbar (3) Multi-stage and (4) Free-space designs.

7.3.1 *Electrical-Photonic Designs*

Shacham et al. propose a hybrid electrical-photonic NoC using electrical interconnect to coordinate and arbitrate a shared photonic medium [11, 12]. These designs achieve very high photonic link utilization by effectively trading increased latency for higher bandwidth. While increased bandwidth without regard for latency is useful for some applications, it eschews a primary benefit of PNoCs over electrical NoCs, low latency. Hendry et al. addressed this issue by introducing an all optical mesh network with photonic time division multiplexing (TDM) arbitration to set up communication path. However, the simulation results show that system still suffers from relatively high average latency [41].

7.3.2 *Crossbar Designs*

Other PNoC work attempts to address the latency issue by providing non-blocking point-to-point links between nodes. In particular, several works propose crossbar topologies to improve the latency of multi-core photonic interconnect. Fully connected crossbars [17] do not scale well, but researchers have examined channel sharing crossbar architectures, called Single-Write-Multiple-Read (SWMR) or Multiple-Write-Single-Read (MWSR), with various arbitration mechanisms for coordinating shared sending and/or receiving channels. Vantrease et al. proposed Corona, a MWSR crossbar, in which each node listens on the dedicated channel, but with the other nodes competing to send data on this channel [20, 21]. To implement arbitration at sender side, the author implemented a token channel [21] or token slot [20] approach similar to token rings used in early LAN network implementations. Alternately, Pan et al. proposed Firefly, a SWMR crossbar design, with a dedicated sending channel for each node, but all the nodes in a crossbar listen on all the sending channels [19]. Pan et al. proposed broadcasting the flit-headers to specify a particular receiver.

In both SWMR and MWSR crossbar designs, over-provisioning of dedicated channels, either at the receiver (SWMR) or sender (MWSR), is required, leading to under utilization of link bandwidth and poor power efficiency. Pan et al. also proposed a channel sharing architecture, FlexiShare [18], to improve the channel utilization and reduce channel over-provisioning. The reduced number of channels, however, limit the system throughput. In addition, FlexiShare requires separated dedicated arbitration channels for sender and receiver sides, incurring additional power and hardware overhead.

Two designs propose to manage laser power consumption at runtime. Chen and Joshi propose to switch off portions of the network based on the bandwidth requirements [42]. Zhou and Kodi propose a method to predict future bandwidth needs and scale laser power appropriately [43].

7.3.3 Multi-Stage Designs

Joshi et al. proposed a photonic multi-stage Clos network with the motivation of reducing the photonic ring count, thus reducing the power for thermal ring trimming [15]. Their design explores the use of a photonic network as a replacement for the middle stage of a three-stage Clos network. While this design achieves an efficient utilization of the photonic channels, it incurs substantial latency due to the multi-stage design.

Koka et al. present an architecture consisting of a grid of nodes where all nodes in each row or column are fully connected by a crossbar [22]. To maintain full-connectivity of the network, electrical routers are used to switch packets between rows and columns. In this design, photonic "grids" are very limited in size to maintain power efficiency, since fully connected crossbars grow at $O(n^2)$ for the number of nodes connected. Kodi and Morris propose a 2-D mesh of optical MWSR crossbars to connect nodes in the x and y dimensions [44]. In a follow-on work by the same authors Morris et al. [45] proposed a hybrid multi-stage design, in which grid rows (x-dir) are subnets fully connected with a photonic crossbar, but different rows (y-dir) are connected by a token-ring arbitrated shared photonic link. Bahirat and Pasricha propose an adaptable hybrid design in which a 2-D Mesh electrical network is overlaid with a set of photonic rings [46].

7.3.4 Free-Space Designs

Xue et al. present a novel free-space optical interconnect for CMPs, in which optical free-space signals are bounced off of mirrors encapsulated in the chip's packaging [47]. To avoid conflicts and contention, this design uses in-band arbitration combined with an acknowledgment based collision detection protocol.

7.4 Power Efficiency in PNoCs

Power efficiency is an important motivation for photonic on-chip interconnect. In photonic interconnect, however, the static power consumption (due to off-chip laser, ring thermal tuning, etc.) dominates the overall power consumption, potentially leading to energy-inefficient photonic interconnects. In this section, prior photonic NoCs are examined in terms of static power efficiency. Bandwidth per watt is used as the metric to evaluate power efficiency of photonic interconnect architectures, showing that it can be improved by optimizing the interconnect topology, arbitration scheme and photonic device layout.

7.4.1 Channel Allocation

We first examine channel allocation in prior photonic interconnect designs. Several previous photonic NoC designs, from fully connected crossbars [17] to the blocking crossbar designs [16, 18–21], provide extra channels to facilitate safe arbitration between sender and receiver. Although conventional photonic crossbars achieve nearly uniform latency and high bandwidth, channels are dedicated to each node and cannot be flexibly shared by the others. Due to the unbalanced traffic distribution in realistic workloads [48], channel bandwidth cannot be fully utilized. This leads to inefficient energy usage, since the static power is constant regardless of traffic load. Over-provisioned channels also implies higher ring resonator counts, which must be maintained at the appropriate trimming temperature, consuming on-chip power. Additionally, as the network size increases, the number of channels required may increase quadratically, complicating the waveguide layout and leading to extra optical loss. An efficient photonic interconnect must solve the problem of efficient channel allocation. Our approach leverages this observation to achieve lower power consumption than previous designs.

7.4.2 Topology and Layout

Topology and photonic device layout can also cause unnecessary optical loss in the photonic link, which in turn leads to greater laser power consumption. Many photonic NoCs globally route waveguides in a bundle, connecting all the tiles in the CMP [16, 19–21]. In these designs, due to the unidirectional propagation property of optical transmission, the waveguide must be routed to reach each node twice (double-back), such that the signal being modulated by senders on the outbound path may be received by all possible receivers. The length of these double-back waveguides leads to significant laser power losses over the long distance.

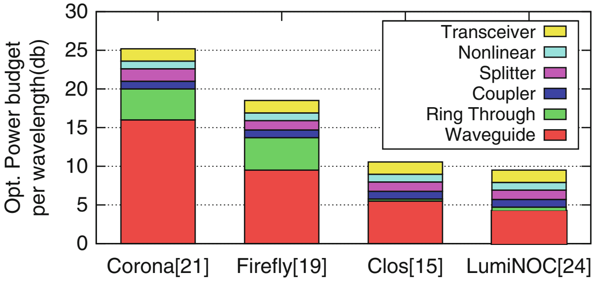


Fig. 7.5 Optical link budgets for the photonic data channels of various photonic NoCs

Figure 7.5 shows the optical link budgets for the photonic data channel of Corona [21], Firefly [19], Clos [15] and LumiNOC under same radix and chip area, based on our power model (described in Sect. 7.6.5). Flexishare [18] is not compared, since not enough information was provided in the paper to estimate the optical power budget at each wavelength. The figure shows that waveguide losses dominate power loss in all three designs. This is due to the long waveguides required to globally route all the tiles on a chip. For example, the waveguide length in Firefly and Clos network in a 400 mm^2 chip are estimated to be 9.5 and 5.5 cm, respectively. This corresponds to 9.5 and 5.5 dB loss in optical power, assuming the waveguide loss is 1 dB/cm [15]. Moreover, globally connected tiles imply a relatively higher number of rings on each waveguide, leading to higher ring through loss. Despite a single-run, bi-directional architecture, even the Clos design shows waveguide loss as the largest single component.

In contrast to other losses (e.g. coupler and splitter loss, filter drop loss and photodetector loss) which are relatively independent of interconnect architecture, waveguide and ring through loss can be reduced through layout and topology optimization. We propose a network architecture which reduces optical loss by decreasing individual waveguide length as well as the number of rings along the waveguide.

7.4.3 Arbitration Mechanism

The power and overhead introduced by the separated arbitration channels or networks in previous photonic NoCs can lead to further power efficiency losses. Corona, a MWSR crossbar design, requires a token channel or token slot arbitration at sender side [20,21]. Alternatively, Firefly [19], a SWMR crossbar design, requires head-flit broadcasting for arbitration at receiver side, which is highly inefficient in PNoCs. FlexiShare [18] requires both token stream arbitration and head-flit broadcast. These arbitration mechanisms require significant overhead in the form of dedicated channels and photonic resources, consuming extra optical laser power.

For example, the radix-32 Flexishare [18] with 16 channels requires 416 extra wavelengths for arbitration, which accounts for 16 % of the total wavelengths in addition to higher optical power for a multi-receiver broadcast of head-flits. Arbitration mechanisms are a major overhead for these architectures, particularly as network radix scales.

There is a clear need for a PNoC architecture that is energy-efficient and scalable while maintaining low latency and high bandwidth. In the following sections, we propose the LumiNOC architecture which reduces the optical loss by partitioning the global network into multiple smaller sub-networks. Furthermore, the proposed novel arbitration scheme leverages the same wavelengths for channel arbitration and parallel data transmission to efficiently utilize the channel bandwidth and photonic resources, without dedicated arbitration channels or networks which lower efficiency or add power overhead to the system.

7.5 LumiNOC Architecture

In our analysis of prior PNoC designs, we have found that a significant amount of laser power consumption was due to the waveguide length required for propagation of the photonic signal across the entire network. Based on this, the LumiNOC design breaks the network into several smaller networks (subnets), with shorter waveguides. Figure 7.6 shows three example variants of the LumiNOC architecture with different subnet sizes, in an example 16-node CMP system: the one-row, two-rows and four-rows designs (note: 16-nodes are shown to simplify explanation, in Sect. 7.6 we evaluate a 64-node design). In the one-row design, a subnet of four tiles is interconnected by a photonic waveguide in the horizontal orientation. Thus four non-overlapping subnets are needed for the horizontal interconnection. Similarly four subnets are required to vertically interconnect the 16 tiles. In the two-row design, a single subnet connects 8 tiles while in the four-row design a single

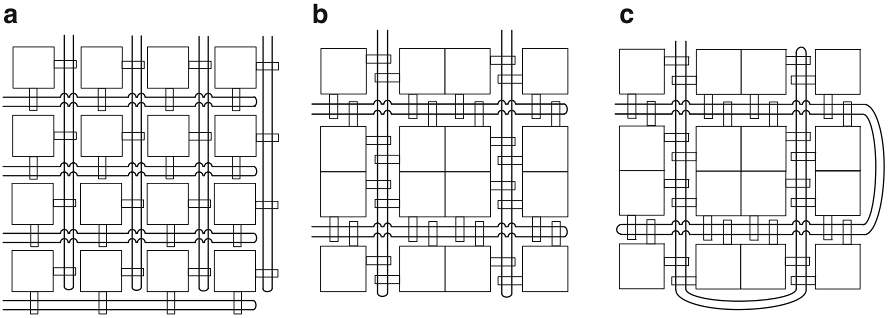


Fig. 7.6 LumiNOC interconnection of CMP with 16 tiles. (**a**) One- (**b**) two- and (**c**) four-rows interconnection

subnet touches all 16 tiles. In general, all tiles are interconnected by two different subnets, one horizontal and one vertical. If a sender and receiver do not reside in the same subnet, transmission requires a hop through an intermediate node's electrical router. In this case, transmission experiences longer delay due to the extra O/E–E/O conversions and router latency. To remove the overheads of photonic waveguide crossings required by the orthogonal set of horizontal and vertical subnets, the waveguides can be deposited into two layers with orthogonal routing [40].

Another observation from prior photonic NoC designs is that channel sharing and arbitration have a large impact on design power efficiency. Efficient utilization of the photonic resources, such as wavelengths and ring resonators, is required to yield the best overall power efficiency. To this end, we leverage the same wavelengths in the waveguide for channel arbitration and parallel data transmission, avoiding the power and hardware overhead due to the separated arbitration channels or networks. Unlike the over-provisioned channels in conventional crossbar architectures, channel utilization in LumiNOC is improved by multiple tiles sharing a photonic channel.

A final observation from our analysis of prior photonic NoC design is that placing many wavelengths within each waveguide through deep wavelength-division multiplexing (WDM) leads to high waveguide losses. This is because the number of rings that each individual wavelength encounters as it traverses the waveguide is proportional to the number of total wavelengths in the waveguide times the number of waveguide connected nodes, and each ring induces some photonic power losses. We propose to limit LumiNOC's waveguides to a few frequencies per waveguide and increase the count of waveguides per subnet, to improve power efficiency with no cost to latency or bandwidth, a technique we call "ring-splitting". Ring-splitting is ultimately limited by the tile size and optical power splitting loss. Assuming a reasonable waveguide pitch of 15 μm required for layout of microrings which have a diameter of 5 μm [31], this leaves 5 μm clearance to avoid optical signal interference between two neighboring rows of rings.

7.5.1 LumiNOC Subnet Design

Figure 7.7 details the shared channel for a LumiNOC one-row subnet design. Each tile contains W modulating "Tx rings" and W receiving "Rx Rings", where W is the number of wavelengths multiplexed in the waveguide. Since the optical signal unidirectionally propagates in the waveguide from its source at off-chip laser, each node's Tx rings are connected in series on the "Data Send Path", shown in a solid line from the laser, prior to connecting each node's Rx rings on the "Data Receive Path", shown in a dashed line. In this "double-back" waveguide layout, modulation by any node can be received by any other node; furthermore, the node which modulates the signal may also receive its own modulated signal, a feature that is leveraged in our collision detection scheme in the arbitration phase. The same wavelengths are leveraged for arbitration and parallel data transmission.

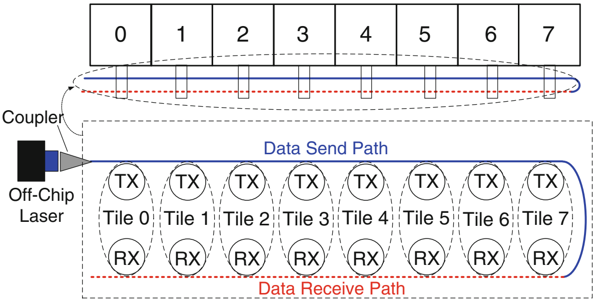


Fig. 7.7 One-row subnet of eight nodes. *Circles* (TX and RX) represent groups of rings; one *dotted oval* represents a tile

During data transmission, only a single sender is modulating on all wavelengths and only a single receiver is tuned to all wavelengths. However, during arbitration (i.e. any time data transfer is not actively occurring) the Rx rings in each node are tuned to a specific, non-overlapping set of wavelengths. Up to half of the wavelengths available in the channel are allocated to this arbitration procedure. with the other half available for credit packets as part of credit-based flow control. This particular channel division is designed to prevent optical broadcasting, the state when any single wavelength must drive more than one receiver, which if allowed would severely increase laser power [49]. Thus, at any given time a multi-wavelength channel with N nodes may be in one of three states: **Idle**—All wavelengths are un-modulated and the network is quiescent. **Arbitration**—One more sender nodes are modulating N copies of the arbitration flags; one copy to each node in the subnet (including itself) with the aim to gain control of the channel. **Data Transmission**—Once a particular sender has established ownership of the channel, it modulates all channel wavelengths in parallel with the data to be transmitted.

In the remainder of this section, we detail the following: *Arbitration*—the mechanism by which the photonic channel is granted to one sender, avoiding data corruption when multiple senders wish to transmit, including *Dynamic Channel Scheduling*, the means of sender conflict resolution, and *Data Transmission*—the mechanism by which data is transmitted from sender to receiver. *Credit Return* is also discussed.

7.5.1.1 Arbitration

We propose an optical collision detecting and dynamic channel scheduling technique to coordinate access of the shared photonic channel. This approach achieves efficient channel utilization without the latency of electrical arbitration schemes [11, 12], or the overhead of wavelengths and waveguides dedicated to standalone arbitration [18,19,21]. In this scheme, a sender works together with its own receiver to ensure message delivery in the presence of conflicts.

Receiver

Once any receiver detects an arbitration flag, it will take one of three actions: if the arbitration flag is uncorrupted (i.e. the sender flag has a 0 in only one location indicating single-sender) and the forthcoming message is destined for this receiver, it will enable all its Rx rings for the indicated duration of the message, capturing it. If the arbitration flags are uncorrupted, but the receiver is not the intended destination, it will detune all of its Rx rings for the indicated duration of the message to allow the recipient sole access. Finally, if a collision is detected, the receiver circuit will enter the **Dynamic Channel Scheduling** phase (described below).

Sender

To send a packet, a node first waits for any on-going messages to complete. Then, it modulates a copy of the arbitration flags to the appropriate arbitration wavelengths for each of the N nodes. The arbitration flags for an example 4-node subnet are depicted in Fig. 7.8. The arbitration flags are a t_{arb} cycle long header (2 in this example) made up of the destination node address (D0–D1), a bimodal packet size indicator (Ln) for the two supported payload lengths (64-bit and 576-bit), and a “1-hot” encoded source address (S0–S3) (i.e. the source address is coded so that each valid encoding for a given source will have exactly one bit set) which serves as a guard band or collision detection mechanism: since the subnet is operated

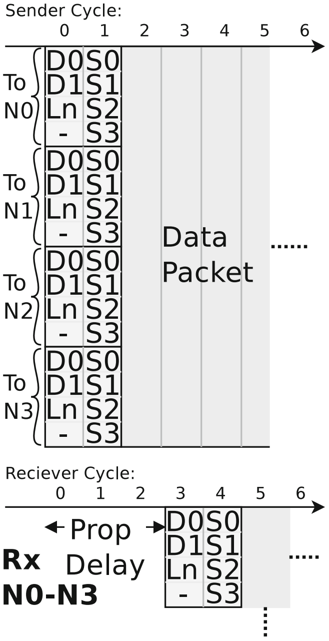


Fig. 7.8 Arbitration on a 4-node subnet

synchronously, any time multiple nodes send overlapping arbitration flags, the "1-hot" precondition is violated and all nodes are aware of the collision. We leverage self-reception of the arbitration flag to detect collision. Right after sending, the node monitors the incoming arbitration flags. If they are uncorrupted (i.e. only one bit is set in the source address), then the sender succeeded in arbitrating the channel and the two nodes proceed to the **Data Transmission** phase. If the arbitration flags are corrupted (i.e. more than one bit is set in the source address), then a conflict has occurred. Any data already sent is ignored and the conflicting senders enter the **Dynamic Channel Scheduling** regime (described below).

The physical length of the waveguide incurs a propagation delay, t_{pd} (cycles), on the arbitration flags traversing the subnet. The "1-hot" collision detection mechanism will only function if the signals from all senders are temporally aligned, so if nodes are physically further apart than the light will travel in 1 cycle, they will be in different clocking domains to keep the packet aligned as it passes the final sending node. Furthermore, the arbitration flags only start on cycles that are an integer multiple of the $t_{pd} + 1$ to assure that no nodes started arbitration during the previous t_{slot} and that all possibly conflicting arbitration flags are aligned. This means that conflicts only occur on arbitration flags, not with data.

Note that a node will not know if it has successfully arbitrated the channel until after $t_{pd} + t_{arb}$ cycles, but will begin data transmission after t_{arb}. In the case of an uncontested link, the data will be captured by the receiver without delay. Upon conflict, senders cease sending (unusable) data.

As as an example, say that the packet in Fig. 7.8 is destined for node 2 with no conflicts. At cycle 5, Nodes 1, 3, and 4 would detune their receivers, but node 2 would enable them all and begin receiving the data flits.

If the subnet size were increased without proportionally increasing the available wavelengths per subnet, then the arbitration flags will take longer to serialize as more bits will be required to encode the source and destination address. If, however, additional wavelengths are provisioned to maintain the bandwidth/node, then the additional arbitration bits are sent in parallel. Thus the general formula for $t_{arb} = ceil(1 + N + log_2(N)/\lambda)$ where N is the number of nodes and λ is the number of wavelengths per arbitration flag.

7.5.1.2 Dynamic Channel Scheduling

Upon sensing a conflicting source address, all nodes identify the conflicting senders and a dynamic, fair schedule for channel acquisition is determined using the sender node index and a global cycle count (synchronized at startup): senders transmit in $(n + cycle) \mod N$ order. Before sending data in turn, each sender transmits an abbreviated version of the arbitration flags: the destination address and the packet size. All nodes tune in to receive this, immediately followed by the **Data Transmission** phase with a single sender and receiver for the duration of the packet. Immediately after the first sender sends its last data flit, the next sender repeats this process, keeping the channel occupied until the last sender completes. After the dynamic schedule completes, the channel goes idle and any node may attempt a new arbitration to acquire the channel as previously described.

7.5.1.3 Data Transmission

In this phase the sender transmits the data over the photonic channel to the receiving node. All wavelengths in the waveguide are used for bit-wise parallel data transmission, so higher throughput is expected when more wavelengths are multiplexed into the waveguide. Two packet payload lengths, 64-bit for simple requests and coherence traffic and 576-bit for cache line transfer, are supported.

7.5.1.4 Credit Return

At the beginning of any arbitration phase (assuming the channel is not in use for Data Transmission), 1/2 of the wavelengths of the channel are reserved for credit return from the credit return transmitter (i.e. the router which has credit to return) to the credit return receiver (i.e. the node which originally sent the data packet and now must be notified of credit availability). Similar to the arbitration flags, the wavelengths are split into N different sub-channels, each one dedicated to a particular credit return receiver. Any router which has credit to send back may then modulate its credit return flag onto the sub-channel to the appropriate credit return receiver. The credit return flag is encoded similarly to the arbitration flag. In the event of a collision between two credit return senders returning credit to the same receiver, no retransmission is needed as the sender part of the flag will uniquely identify all nodes sending credit back to this particular credit return receiver. Credit is returned on a whole-packet basis, rather than a flit basis to decrease overheads. The packet size bit Ln is not used in the credit return flag; credit return receivers must keep a history of the packet sizes transmitted so that the appropriate amount of credit is returned.

7.5.2 Router Microarchitecture

The electrical router architecture for LumiNOC is shown in Fig. 7.9. Each router serves both as an entry point to the network for a particular core, as well as an intermediate node interconnecting horizontal and vertical subnets. If a processor must send data to another node on the same vertical or horizontal subnet, packets are switched from the electrical input port to the vertical photonic output port with one E/O conversion. Packets which are destined for a different subnet must be first routed to an intermediate node via the horizontal subnet before being routed on the vertical subnet. Each input port is assigned with a particular virtual-channel (VCs) to

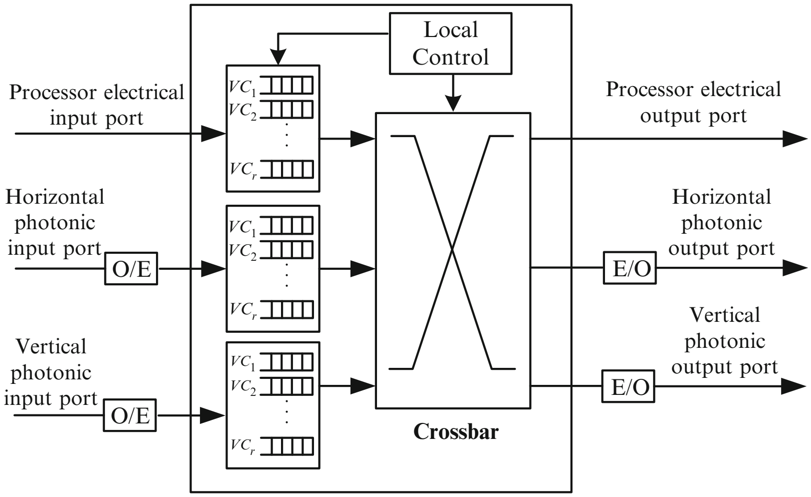

Local Control
Processor electrical input port
Horizontal photonic input port
O/E
Vertical photonic input port
O/E
VC_1
VC_2
VC_r
Crossbar
Processor electrical output port
Horizontal photonic output port
E/O
Vertical photonic output port
E/O

Fig. 7.9 Router microarchitecture

hold the incoming flits for a particular sending node. The local control unit performs routing computation, virtual-channel allocation and switching allocation in crossbar. The LumiNOC router's complexity is similar to that of an electrical, bi-directional, 1-D ring network router, with the addition of the E/O-O/E logic.

7.6 Evaluation

In this section, we describe a particular implementation of the LumiNOC architecture and analyze its performance and power efficiency.

7.6.1 64-Core LumiNOC Implementation

Here we develop a baseline physical implementation of the general LumiNOC architecture specified in Sect. 7.5 for the evaluation of LumiNOC against competing PNOC architectures. We assume a 400 mm^2 chip implemented in a 22 nm CMOS process and containing 64 square tiles that operate at 5 GHz, as shown in Fig. 7.10. A 64-node LumiNOC design point is chosen here as a reasonable network size which could be implemented in a 22 nm process technology. Each tile contains a processor core, private caches, a fraction of the shared last-level cache, and a router connecting it to one horizontal and one vertical photonic subnet. Each router input port contains seven virtual channels (VCs), each five flits deep. Credit based flow

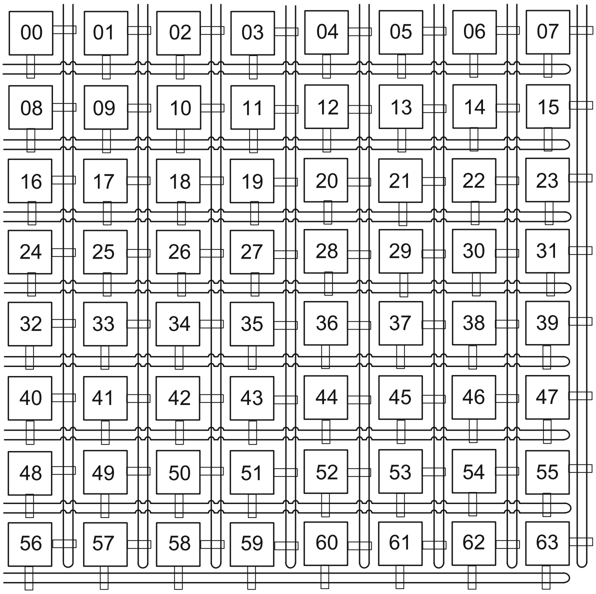


Fig. 7.10 One-row LumiNOC with 64 tiles

control is implemented via the remainder of the photonic spectrum not used for arbitration during arbitration periods in the network.

A 64-node LumiNOC may be organized into three different architectures: the one-row, two-row and four-row designs (shown in Fig. 7.6), which represent a trade-off between interconnect power, system throughput and transmission latency. For example, power decreases as row number increases from one-row to two-row, since the single waveguide is roughly with the same length, but fewer waveguides are required. The low-load latency is also reduced due to more nodes residing in the same subnet, reducing the need for intermediate hops via an electrical router. The two-row subnet design, however, significantly reduces throughput due to the reduced number of transmission channels. As a result, we choose the "one-row" subnet architecture of Fig. 7.6a, with 64-tiles arranged as shown in Fig. 7.10 for the remainder of this section. In both the horizontal and vertical axes there are 8 subnets which are formed by 8 tiles that share a photonic channel, resulting in all tiles being redundantly interconnected by two subnets. Silicon nitride waveguides are assumed to be the primary transport strata. Similar to electrical wires, silicon nitride waveguides can be deployed into multiple strata to eliminate in-plane waveguide crossing, thus reducing the optical power loss [40]. In order to optimize system performance and efficiently utilize the chip area, three-dimensional integration (3DI) is emerging for the integration of silicon nanophotonic devices with conventional CMOS electronics. In 3DI, the silicon photonic on-chip networks are fabricated into a separate silicon-on-insulator (SOI) die or layer with a thick layer of buried oxide (BOX) that acts as bottom cladding to prevent light leakage into the substrate. This photonic layer stacks above the electrical layers containing the compute tiles.

As a general trend, multirow designs tend to decrease power consumption in the router as fewer router hops are required to cover more of the network. Because of the diminishing returns in terms of throughput as channel width increases, however, congestion increases and the bandwidth efficiency drops. Further, the laser power grows substantially for a chip as large as the one described here. For smaller floorplans, however, multi-row LumiNOC would be an interesting design point.

We assume a 10 GHz network modulation rate, while the routers and cores are clocked at 5 GHz. Muxes are placed on input and output registers such that on even network cycles, the photonic ports will interface with the lower half of a given flit and on odd, the upper half. With a 400 mm^2 chip, the effective waveguide length is 4.0 cm, yielding a propagation delay of $t_{pd} = 2.7$ 10 GHz network cycles.

When sender and receiver reside in the same subnet, data transmission is accomplished with a single hop, i.e. without a stop in an intermediate electrical router. Two hops are required if sender and receiver reside in different subnets, resulting in a longer delay due to the extra O/E–E/O conversion and router latency. The "one-row" subnet based network implies that for any given node 15 of the 63 possible destinations reside within one hop, the remaining 48 destinations require two hops.

7.6.1.1 Link Width Versus Packet Size

Considering the link width, or the number of wavelengths per logical subnet, if the number of wavelengths and thus channel width is increased, it should raise ideal throughput and theoretically reduce latency due to serialization delay. We are constrained, however, by the 2.7 network cycle propagation delay of the link (t_{pd} above), and the small packet size of single cache line transfers in typical CMPs. There is no advantage to sending the arbitration flags all at once in parallel when additional photonic channels are available; the existing bits would need to be replaced with more guard bits to provide collision detection. Thus, the arbitration flags would represent an increasing overhead. Alternately, if the link were narrower, the 2.7 cycle window would be too short to send all the arbitration bits and a node would waste time broadcasting arbitration bits to all nodes after it effectively "owns" the channel. Thus, the optimal link width is 64 wavelengths under our assumptions for clock frequency and waveguide length.

If additional spectrum or waveguides are available, then we propose to implement multiple parallel, independent **Network Layers**. Instead of one network with a 128-bit data path, there will be two parallel 64-bit networks. This allows us to exploit the optimal link width while still providing higher bandwidth. When a node injects into the network, it round-robins through the available input ports for each layer, dividing the traffic amongst the layers evenly.

7.6.1.2 Ring-Splitting

Given a 400 mm^2 64-tile PNoC system, each tile is physically able to contain 80 double-back waveguides. However, the ring-splitting factor is limited to 4 (32 wavelengths per waveguide) in this design to avoid the unnecessary optical splitting loss due to the current technology. This implies a trade off of waveguide area for lower power. The splitting loss has been included in the power model in Sect. 7.6.5.

7.6.1.3 Scaling to Larger Networks

We note, it is likely that increasing cores connected in a given subnet will yield increased contention. A power-efficient means to cover the increase in bandwidth demand due to more nodes would be to increase the number of layers. We find the degree of subnet partitioning is more dependent upon the physical chip dimensions than the number of nodes connected, as the size of the chip determines the latency and frequency of arbitration phases. For this reason our base implementation assumes a large, 400 mm^2 die. Increasing nodes while retaining the same physical dimensions will cause a sub-linear increase in arbitration flag size with nodes-per-subnet (the Source ID would increase linearly, the Destination ID would increase as $log(n)$), and hence more overhead than in a smaller sub-net design.

7.6.2 Experimental Methodology

To evaluate this implementation's performance, we use a cycle-accurate, micro-architectural network simulator, ocin _tsim [50]. The network was simulated under both synthetic and realistic workloads. LumiNOC designs with 1, 2, and 4 **Network Layers** are simulated to show results for different bandwidth design points.

7.6.2.1 Photonic Networks

The baseline, 64-node LumiNOC system, as described in Sect. 7.6, was simulated for all evaluation results. Synthetic benchmark results for the Clos LTBw network are presented for comparison against the LumiNOC design. We chose the Clos LTBw design as the most competitive in terms of efficiency and bandwidth as discussed in Sect. 7.6. Clos LTBw data points were extracted from the paper by Joshi et al [15].

7.6.2.2 Baseline Electrical Network

In the results that follow, our design is compared to a electrical 2-D mesh network. Traversing the dimension order network consumes three cycles per hop; one cycle for link delay and two within each router. The routers have two virtual channels per port, each ten flits deep, and implement wormhole flow control.

7.6.2.3 Workloads

Both synthetic and realistic workloads were simulated. The traditional synthetic traffic patterns, *uniform random* and *bit-complement* represent nominal and worst-case traffic for this design. These patterns were augmented with the *P8D* pattern, proposed by Joshi et al. [15], designed as a best-case for staged or hierarchical networks where traffic is localized to individual regions. In P8D, nodes are assigned to one of 8 groups, made up of topologically adjacent nodes and nodes only send random traffic within the group. In these synthetic workloads, all packets contain data payloads of 512-bits, representing four flits of data in the baseline electrical NoC.

Realistic workload traces were captured for a 64-core CMP running PARSEC benchmarks with the sim-large input set [51]. The Netrace trace dependency tracking infrastructure was used to ensure realistic packet interdependencies are expressed as in a true, full-system CMP system [52]. The traces were captured from a CMP composed of 64 in-order cores with 32-KB, private L1I and L1D caches and a shared 16 MB LLC. Coherence among the L1 caches was maintained via a MESI protocol. A 150 million cycle segment of the PARSEC benchmark "region of interest" was simulated. Packet sizes for realistic workloads vary bimodally between 64 and 576 bits for miss request/coherence traffic and cache line transfers.

7.6.3 Synthetic Workload Results

In Fig. 7.11, the LumiNOC design is compared against the electrical and Clos networks under *uniform random*, *bit complement*, and *P8D*. The figure shows the low-load latencies of the LumiNOC design are much lower than the competing designs. This is due primarily to the lower diameter of the LumiNOC topology; destinations within one subnet are one "hop" away while those in a second subnet are two. The 1-layer network saturates at 4 Tbps realistic throughput as determined by analyzing the offered vs. accepted rate.

The different synthetic traffic patterns bring out interesting relationships. On the *P8D* pattern, which is engineered to have lower hop counts, all designs have universally lower latency than on other patterns. However, while both the electrical and LumiNOC network have around 25 % lower low-load latency than uniform random, Clos only benefits by a few percent from this optimal traffic pattern.

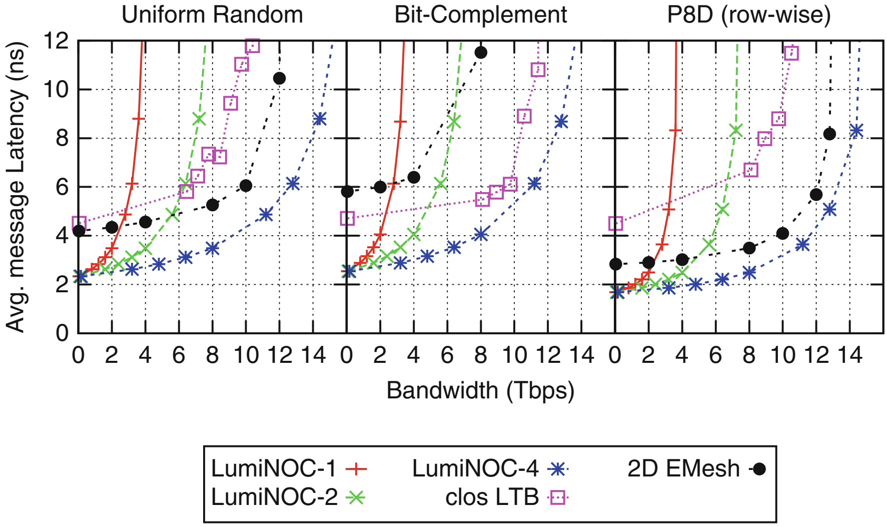


Fig. 7.11 Synthetic workloads showing LumiNOC vs. Clos LTBw and electrical network. LumiNOC-1 refers to the 1-layer LumiNOC design, LumiNOC-2 the 2-layer, and LumiNOC-4 the 4-layer

At the other extreme, the electrical network experiences a 50 % increase in no-load latency under the bit-complement pattern compared to uniform random while both Clos and the LumiNOC network are only marginally affected. This is due to the LumiNOC having a worst-case hop count of two and not all routes go through the central nodes as in the electrical network. Instead, the intermediate nodes are well distributed through the network under this traffic pattern. However, as the best-case hop count is also two with this pattern, the LumiNOC network experiences more contention and the saturation bandwidth is decreased as a result.

7.6.4 Realistic Workload Results

Figure 7.12 shows the performance of the LumiNOC network in 1-, 2- and 4-layers, normalized against the performance of the baseline electrical NoC. Even with one layer, the average message latency is about 10 % lower than the electrical network. With additional network layers, LumiNOC has approximately 40 % lower average latency. These results are explained by examining the bandwidth-latency curves in Fig. 7.11. The average offered rates for the PARSEC benchmarks are of the order of 0.5 Tbps, so these applications benefit from LumiNOC's low latency while being well under even the 1-layer LumiNOC throughput.

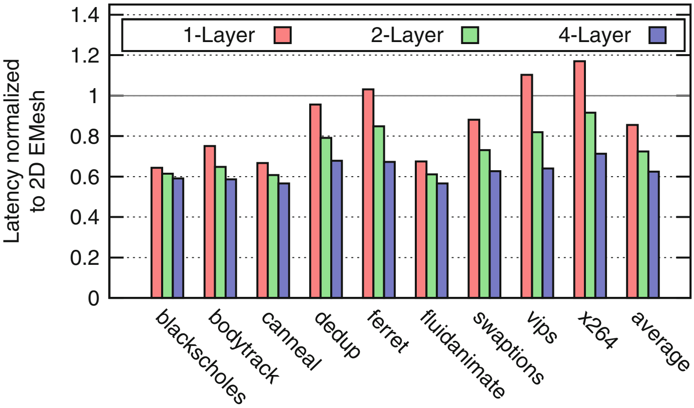


Fig. 7.12 Message Latency in PARSEC benchmarks for LumiNOC compared to electrical network

Table 7.1 Components of optical loss

Loss component	Value (dB)	Loss component	Value
Coupler	1	Waveguide	1 dB/cm
Splitter	0.2	Waveguide crossing	0.05 dB
Non-linearity	1	Ring through	0.001 dB
Modulator insertion	0.001	Filter drop	1.5 dB
Photodetector	0.1		

7.6.5 PNoC Power Model

In this section, we describe our power model and compare the baseline LumiNOC design against prior work PNoC architectures. In order for a fair comparison versus other reported PNoC architectures, we refer to the photonic loss of various photonic devices reported by Joshi et al. [15] and Pan et al. [18], shown in Table 7.1. Equation (7.1) shows the major components of our total power model.

$$TP = ELP + TTP + ERP + EO/OE \tag{7.1}$$

TP = Total Power, ELP = Electrical Laser Power, TTP = Thermal Tuning Power, ERP = Electrical Router Power and EO/OE = Electrical to Optical/Optical to Electrical conversion power. Each components is described below.

7.6.5.1 ELP

Electrical laser power is converted from the calculated optical power. Assuming a 10 µW receiver sensitivity, the minimum static optical power required at each

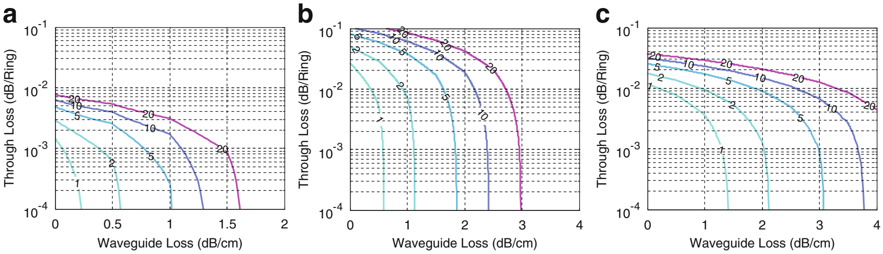


Fig. 7.13 Contour plots of the Electrical Laser Power (ELP) in Watts for networks with the same aggregate throughput. *Each line* represents a constant power level (Watts) at a given ring through loss and waveguide loss combination (assuming 30 % efficient electrical to optical power conversion). (**a**) Crossbar, (**b**) Clos, (**c**) LumiNOC

Table 7.2 Configuration comparison of various photonic NoC architectures

Literature		N_{core}	N_{node}	N_{rt}	N_{wg}	N_{wv}	N_{ring} (K)
EMesh [1]		128	64	64	NA	NA	NA
Corona [21]		256	64	64	388	24,832	1,056
FlexiShare [18]		64	32	32	NA	2,464	550
Clos [15]		64	8	24	56	3,584	14
LumiNOC	1-L	64	64	64	32	1,024	16
	2-L	64	64	64	64	2,048	32
	4-L	64	64	64	128	4,096	65

N_{core} number of cores in the CMP, N_{node} number of nodes in the NoC, N_{rt} total number of routers, N_{wg} total number of waveguides, N_{wv} total number of wavelengths, N_{ring} total number of rings

wavelength to activate the last photodetector at the end of a waveguide in the PNoC system is estimated based on Eq. (7.2). This optical power is then converted to electrical laser power using 30 % efficiency.

$$P_{optical} = N_{wg} \cdot N_{wv} \cdot P_{th} \cdot K \cdot 10^{(\frac{1}{10} \cdot l_{channel} \cdot P_{WGloss})} \cdot 10^{(\frac{1}{10} \cdot N_{ring} \cdot P_{tloss})} \quad (7.2)$$

In Eq. (7.2), N_{wg} is the number of waveguides in the PNoC system, N_{wv} is the number of wavelength per waveguide, P_{th} is receiver sensitivity power, $l_{channel}$ is waveguide length, P_{wgloss} is optical signal propagation loss in waveguide (dB/cm), N_{ring} is the number of rings attached on each waveguide, P_{tloss} is modulator insertion and filter ring through loss (dB/ring) (assume they are equal), K accounts for the other loss components in the optical path including P_c, coupling loss between the laser source and optical waveguide, P_b, waveguide bending loss, and $P_{splitter}$, optical splitter loss. Figure 7.13 shows electrical laser power contour plot, derived from Eq. (7.2) and the configurations of Table 7.2, showing the photonic device power requirements at a given electrical laser power, for a SWMR photonic crossbar (Corona) [21], Clos [15] and LumiNOC with equivalent throughput

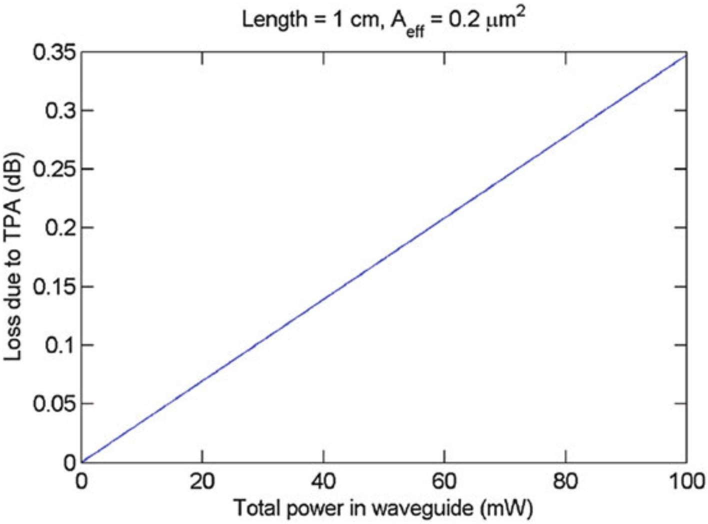


Fig. 7.14 Nonlinear optical loss in the silicon waveguide vs optical power in waveguide; waveguide length equals 1 cm with effective area of 0.2 μm^2. Figure produced by Jason Pelc of HP labs with permission

(20 Tbps), network radix and chip area. In Fig. 7.13, the x and y-axis represent two major optical loss components, waveguide propagation loss and ring through loss, respectively. A larger x- and y-intercept implies relaxed requirements for the photonic devices. As shown, given a relatively low 1W laser power budget, the two-layer LumiNOC can operate with a maximum 0.012 dB ring through loss and waveguide loss of 1.5 dB/cm.

We note that optical non-linear loss also effects the optical interconnect power. At telecom wavelengths, two-photon absorption (TPA) in the silicon leads to a propagation loss that increases linearly with the power sent down the waveguide. TPA is a nonlinear optical process and is several orders of magnitude weaker than linear absorption. This nonlinear loss, however, also has significant impact on the silicon-photonic link power budget if a high level of optical power (e.g. >1 W) is injected into silicon waveguide. Figure 7.14 shows the computed nonlinear loss of a 1 cm waveguide versus the optical power in the waveguide. It shows a nonlinear loss of ∼0.35 dB for up to ∼100 mW waveguide optical power. In LumiNoC, the non-linear effect has been included in the optical power calculation.

7.6.5.2 TTP

Thermal tuning is required to maintain microring resonant at the work wavelength. In the calculation, a ring thermal tuning power of 20 μW is assumed for a 20 K temperature tuning range [15, 18]. In a photonic NoC, total thermal tuning power (TTP) is proportional to ring count.

7.6.5.3 ERP

The baseline electrical router power is estimated by the power model reported by Kim et al. [53]. We synthesized the router using TSMC 45 nm library. Power is measured via Synopsis Power Compiler, using simulated traffic from a PARSEC [51] workload to estimate its dynamic component. Results are analytically scaled to 22 nm (dynamic power scaled according to the CMOS dynamic power equation and static power linearly with voltage).

7.6.5.4 EO/OE

The power for conversion between the electrical and optical domains (EO/OE) is based on the model reported by Joshi et al. [15], which assumes a total transceiver energy of 40 fJ/bit data-traffic dependent energy and 10 fJ/bit static energy. Since previous photonic NoCs consider different traffic loads, it is unfair to compare the EO/OE power by directly using their reported figures. Therefore, we compare the worst-case power consumption when each node was arbitrated to get a full access on each individual channel. For example, Corona is a MWSR 64 × 64 crossbar architecture. At the worst-case, 64 nodes are simultaneously writing on 64 different channels. This is combined with a per-bit activity factor of 0.5 to represent random data in the channel.

While this approach may not be 100 % equitable for all designs, we note that EO/OE power does not dominate in any of the designs (see Table 7.3). Even if EO/OE power is removed entirely from the analysis, the results would not change significantly. Further, LumiNOC experiences more EO/OE dynamic power than the other designs due hops through the middle routers.

Table 7.3 Power efficiency comparison of different photonic NoC architectures

		ELP	TTP	ERP	EO/OE	ITP	RTP	TP	RTP/W
Literature		(W)	(W)	(W)	(W)	(Tbps)	(Tpbs)	(W)	(Tbps/W)
EMesh [1]		NA	NA	NA	NA	10	3.0	26.7	**0.1**
Corona [21]		26.0	21.00	0.52	4.92	160	73.6	52.4	**1.4**
FlexiShare [18]		5.80	11.00	0.13	0.60	20	9.0	17.5	**0.5**
Clos [15]		3.30	0.14	0.10	0.54	18	10.0	4.1	**2.4**
LumiNOC	1-Layer	0.35	0.33	0.13	0.30	10	4.0	1.1	**3.6**
	2-Layers	0.73	0.65	0.26	0.61	20	8.0	2.3	**3.4**
	4-Layers	1.54	1.31	0.52	1.22	40	16.0	4.6	**3.4**

ELP electrical laser power, *TTP* thermal tuning power, *ERP* electrical router power, *EO/OE* electrical to optical/optical to electrical conversion power, *ITP* ideal throughput, *RTP* realistic throughput, *TP* total power

7.6.6 Power Comparison

Table 7.2 lists the photonic resource configurations for various photonic NoC architectures, including one-layer, two-layer and four-layer configurations of the LumiNOC. While the crossbar architecture of Corona has a high ideal throughput, the excessive number of rings and waveguides results in degraded power efficiency. In order to support equal 20 Tbps aggregate throughput, LumiNOC requires less than $\frac{1}{10}$ the number of rings of FlexiShare and almost the same number of wavelengths. Relative to the Clos architecture, LumiNOC requires around $\frac{4}{7}$ wavelengths, though approximately double number of rings.

The power and efficiency of the network designs is compared in Table 7.3. Where available/applicable, power and throughput numbers for competing PNoC designs are taken from the original papers, otherwise they are calculated as described in Sect. 7.6.5. **ITP** is the ideal throughput of the design, while **RTP** is the maximum throughput of the design under a uniform random workload as shown in Fig. 7.11. A 6×4 2 GHz electrical 2D-mesh [1] was scaled to 8×8 nodes operating at 5 GHz, in a 22 nm CMOS process (dynamic power scaled according to the CMOS dynamic power equation and static power linearly with voltage), to compare against the photonic networks.

The table shows that LumiNOC has the highest power efficiency of all designs compared in RTP/Watt, increasing efficiency by ∼40 % versus the nearest competitor, Clos [15]. By reducing wavelength multiplexing density, utilizing shorter waveguides, and leveraging the data channels for arbitration, LumiNOC consumes the least ELP among all the compared architectures. A 4-layer LumiNOC consumes ∼1/4th the ELP of a competitive Clos architecture, of nearly the same throughput. Corona [21] contains 256 cores with 4 cores sharing an electrical router, leading to a 64-node photonic crossbar architecture; however, in order to achieve throughput of 160 Gbps, each channel in Corona consists of 256 wavelengths, 4× the wavelengths in a 1-layer LumiNOC. In order to support the highest ideal throughput, Corona consumes the highest electrical router power in the compared photonic NoCs.

Although FlexiShare attempts to save laser power with its double-round waveguide, which reduces the overall non-resonance ring through-loss (and it is substantially more efficient than Corona), its RTP/W remains somewhat low for several reasons. First, similar to other PNoC architectures, FlexiShare employs a global, long waveguide bus instead of multiple short waveguides for the optical interconnects. The global long waveguides cause relatively large optical loss and overburden the laser. Second, FlexiShare is particularly impacted by the high number of ring resonators ($N_{ring} = 550$ K—Table 7.2), each of these rings need to be heated to maintain its proper frequency response and the power consumption of this heating dominates its RTP/W. Third, the dedicated physical arbitration channel in FlexiShare costs extra optical power. Finally, similar to an SWMR crossbar network (e.g. Firefly [19]), FlexiShare broadcasts to all the other receivers for receiver-side arbitration. Although the authors state that, by only broadcasting the head flit, the cost of broadcast in laser power is avoided, we would argue this would be impractical in practice. Since the turn-around time for changing off-die laser power is so high, a constant laser power is needed to support the worst-case power consumption.

7.7 Conclusions

Photonic NoCs are a promising replacement for electrical NoCs in future many-core processors. In this work, we analyze prior photonic NoCs, with an eye towards efficient system power utilization and low-latency. The analysis of prior photonic NoCs reveals that power inefficiencies are mainly caused by channel over-provisioning, unnecessary optical loss due to topology and photonic device layout and power overhead from the separated arbitration channels and networks. LumiNOC addresses these issues by adopting a shared-channel, photonic on-chip network with a novel, in-band arbitration mechanism to efficiently utilize power, achieving a high performance and scalable interconnect with extremely low latency. Simulations show under synthetic traffic, LumiNOC enjoys 50 % lower latency at low loads and ∼40 % higher throughput per Watt on synthetic traffic, versus other reported photonic NoCs. LumiNOC also reduces latencies ∼40 % versus an electrical 2D mesh NoCs on the PARSEC shared-memory, multithreaded benchmark suite.

References

1. Howard J, Dighe S, Vangal SR, Ruhl G, Borkar N, Jain S, Erraguntla V, Konow M, Riepen M, Gries M, Droege G, Lund-Larsen T, Steibl S, Borkar S, De VK, Wijngaart RVD. A 48-Core IA-32 processor in 45 nm CMOS using on-die message-passing and DVFS for performance and power scaling. IEEE J Solid-State Circuits. 2011;46:173–83.
2. Anders M, Kaul H, Hsu S, Agarwal A, Mathew S, Sheikh F, Krishnamurthy R, Borkar S. A 4.1 Tb/s bisection-bandwidth 560 Gb/s/W streaming circuit-switched mesh network-on-chip in 45 nm cmos. In: 2010 IEEE international solid-state circuits conference digest of technical papers (ISSCC), 2010. pp. 110–1.
3. Kim J, Park D, Theocharides T, Vijaykrishnan N, Das CR. A Low Latency Router Supporting Adaptivity for On-Chip Interconnects. In: 2005 Design automation conference, 2005. pp. 559–64.
4. Lipson M. Compact electro-optic modulators on a silicon chip. IEEE J Sel Top Quantum Electron. 2006;12(6):1520–6.
5. Liu A, Liao L, Rubin D, Nguyen H, Ciftcioglu B, Chetrit Y, Izhaky N, Paniccia M. High-speed optical modulation based on carrier depletion in a silicon waveguide. Opt Express. 2007;15(2):660–8.
6. Reshotko M, Block B, Jin B, Chang P. Waveguide coupled Ge-on-oxide photodetectors for integrated optical links. In: The 2008 5th IEEE international conference on group IV photonics, 2008. pp. 182–4.
7. Holzwarth C, Orcutt J, Li H, Popovic M, Stojanovic V, Hoyt J, Ram R, Smith H. Localized substrate removal technique enabling strong-confinement microphotonics in bulk Si CMOS processes. In: Conference on lasers and electro-optics, 2008. pp. 1–2.

8. Kimerling LC, Ahn D, Apsel A, Beals M, Carothers D, Chen Y-K, Conway T, Gill DM, Grove M, Hong C-Y, Lipson M, Michel J, Pan D, Patel SS, Pomerene AT, Rasras M, Sparacin DK, Tu K-Y, White AE, Wong CW. Electronic-photonic integrated circuits on the CMOS platform. In: Silicon photonics, 2006. pp. 6–15.
9. Narasimha A, Analui B, Liang Y, Sleboda T, Gunn C. A fully integrated 4–10-Gb/s DWDM optoelectronic transceiver implemented in a standard 0.13 m CMOS SOI technology. In: The IEEE international solid-state circuits conference, 2007. pp. 42–586.
10. Young I, Mohammed E, Liao J, Kern A, Palermo S, Block B, Reshotko M, Chang P. Optical I/O technology for tera-scale computing. In: The IEEE international solid-state circuits conference, 2009. pp. 468–9.
11. Hendry G, Kamil S, Biberman A, Chan J, Lee B, Mohiyuddin M, Jain A, Bergman K, Carloni L, Kubiatowicz J, Oliker L, Shalf J. Analysis of photonic networks for a chip multiprocessor using scientific applications. In: The 3rd ACM/IEEE international symposium on networks-on-chip (NOCS), 2009. pp. 104–13.
12. Shacham A, Bergman K, Carloni LP. On the design of a photonic network-on-chip. In: The first international symposium on networks-on-chip (NOCS), 2007. pp. 53–64.
13. Shacham A, Bergman K, Carloni LP. Photonic NoC for DMA communications in chip multiprocessors. In: The 15th annual IEEE symposium on high-performance interconnects, 2007. pp. 29–38.
14. Shacham A, Bergman K, Carloni LP. Photonic networks-on-chip for future generations of chip multiprocessors. IEEE Trans Comput. 2008;57(9):1246–60.
15. Joshi A, Batten C, Kwon Y-J, Beamer S, Shamim I, Asanovic K, Stojanovic V. Silicon-photonic clos networks for global on-chip communication. In: The 2009 3rd ACM/IEEE international symposium on networks-on-chip (NOCS), 2009. pp. 124–33.
16. Kirman N, Kirman M, Dokania R, Martinez J, Apsel A, Watkins M, Albonesi D. Leveraging optical technology in future bus-based chip multiprocessors. In: The 39th annual IEEE/ACM international symposium on microarchitecture (Micro), 2006. pp. 492–503.
17. Krishnamoorthy A, Ho R, Zheng X, Schwetman H, Lexau J, Koka P, Li G, Shubin I, Cunningham J. Computer systems based on silicon photonic interconnects. Proc IEEE. 2009;97(7):1337–61.
18. Pan Y, Kim J, Memik G. FlexiShare: channel sharing for an energy-efficient nanophotonic crossbar. In: The 16th IEEE international symposium on high performance computer architecture (HPCA), 2010. pp. 1–12.
19. Pan Y, Kumar P, Kim J, Memik G, Zhang Y, Choudhary A. Firefly: illuminating future network-on-chip with nanophotonics. In: 36th International symposium on computer architecture (ISCA), 2009.
20. Vantrease D, Binkert N, Schreiber R, Lipasti MH. Light speed arbitration and flow control for nanophotonic interconnects. In: 42nd Annual IEEE/ACM international symposium on microarchitecture, 2009. pp. 304–15.
21. Vantrease D, Schreiber R, Monchiero M, McLaren M, Jouppi NP, Fiorentino M, Davis A, Binkert N, Beausoleil RG, Ahn JH. Corona: system implications of emerging nanophotonic technology. In: 35th International symposium on computer architecture (ISCA), 2008. pp. 153–64.
22. Koka P, McCracken MO, Schwetman H, Zheng X, Ho R, Krishnamoorthy AV. Silicon-photonic network architectures for scalable, power-efficient multi-chip systems. In: 37th International symposium on computer architecture (ISCA), 2010. pp. 117–28.
23. Kao YH, Chao HJ, BLOCON: a bufferless photonic clos network-on-chip architecture. In: 5th ACM/IEEE international symposium on networks-on-chip (NoCS), 2011. pp. 81–8.
24. Li C, Browning M, Gratz PV, Palermo S. Luminoc: a power-efficient, high-performance, photonic network-on-chip for future parallel architectures. In: Proceedings of the 21st international conference on parallel architectures and compilation techniques, PACT '12, (New York, NY, USA), ACM, 2012. pp. 421–2.
25. Young I, Mohammed E, Liao J, Kern A, Palermo S, Block B, Reshotko M, Chang P. Optical I/O technology for tera-scale computing. IEEE J Solid State Circuits. 2010;45:235–48.

26. Lee BG, Rylyakov AV, Green WMJ, Assefa S, Baks CW, Rimolo-Donadio R, Kuchta DM, Khater MH, Barwicz T, Reinholm C, Kiewra E, Shank SM, Schow CL, Vlasov YA. Four- and eight-port photonic switches monolithically integrated with digital CMOS logic and driver circuits. In: IEEE-OSA optical fiber communications conference, 2013. pp. 1–3.
27. Roth JE, Palermo S, Helman NC, Bour DP, Miller DAB, Horowitz M. An optical interconnect transceiver at 1550 nm using low-voltage electroabsorption modulators directly integrated to CMOS. IEEE-OSA J Lightwave Technol. 2007;25:3739–47.
28. Liu A, Liao L, Rubin D, Basak J, Nguyen H, Chetrit Y, Cohen R, Izhaky N, Paniccia M. High-speed silicon modulator for future vlsi interconnect. In: Integrated photonics and nanophotonics research and applications/slow and fast light, 2007. p. IMD3, Optical Society of America.
29. Wojcik GL, Yin D, Kovsh AR, Gubenko AE, Krestnikov IL, Mikhrin SS, Livshits DA, Fattal DA, Fiorentino M, Beausoleil RG. A single comb laser source for short reach WDM interconnects. In: Society of photo-optical instrumentation engineers (SPIE) conference series. Society of photo-optical instrumentation engineers (SPIE) conference series, vol. 7230, 2009.
30. Soref RA, Bennett B. Electrooptical effects in silicon. IEEE J Quantum Electron. 1987;23:hbox123–9.
31. Li C, Bai R, Shafik A, Tabasy E, Tang G, Ma C, Chen C-H, Peng Z, Fiorentino M, Chiang P, Palermo S. A ring-resonator-based silicon photonics transceiver with bias-based wavelength stabilization and adaptive-power-sensitivity receiver. In 2013 IEEE international solid-state circuits conference digest of technical papers (ISSCC), 2013. pp. 124–5.
32. Li G, Zheng X, Yao J, Thacker H, Shubin I, Luo Y, Raj K, Cunningham JE, Krishnamoorthy AV. High-efficiency 25 Gb/s CMOS ring modulator with integrated thermal tuning. In: 8th IEEE Intentional Conference on Group IV Photonics (GFP), vol. 4, 2011. pp. 8–10.
33. Xu Q, Manipatruni S, Schmidt B, Shakya J, Lipson M. 12.5 Gbit/s carrier-injection-based silicon micro-ring silicon modulators. Opt. Express. 2007;15:430–6.
34. Chen C-H, Li C, Shafik A, Fiorentino M, Chiang P, Palermo S, Beausoleil R. A wdm silicon photonic transmitter based on carrier-injection microring modulators. In: 2014 IEEE optical interconnects conference, 2014.
35. Liu F, Patil D, Lexau J, Amberg P, Dayringer M, Gainsley J, Moghadam H, Zheng X, Cunningham J, Krishnamoorthy A, Alon E, Ho R. 10-Gbps, 5.3-mW optical transmitter and receiver circuits in 40-nm cmos. IEEE J Solid State Circuits. 2012;47(9):2049–67.
36. Sun C, Timurdogan E, Watts M, Stojanovic V. Integrated microring tuning in deep-trench bulk cmos. In: 2013 IEEE optical interconnects conference, 2013. pp. 54–5.
37. Orcutt JS, Moss B, Sun C, Leu J, Georgas M, Shainline J, Zgraggen E, Li H, Sun J, Weaver M, Urošević S, Popović M, Ram RJ, Stojanović V. Open foundry platform for high-performance electronic-photonic integration. Opt. Express. 2012;20:12222–32.
38. Dong P, Qian W, Liang H, Shafiiha R, Feng D, Li G, Cunningham JE, Krishnamoorthy AV, Asghari M. Thermally tunable silicon racetrack resonators with ultralow tuning power. Opt. Express. 2010;18:20298–304.
39. Orcutt JS, Khilo A, Holzwarth CW, Popović MA, Li H, Sun J, Bonifield T, Hollingsworth R, Kärtner FX, Smith HI, Stojanović V, Ram RJ. Nanophotonic integration in state-of-the-art cmos foundries. Opt. Express. 2011;19:2335–46.
40. Biberman A, Preston K, Hendry G, Sherwood-droz N, Chan J, Levy JS, Lipson M, Bergman K. Photonic network-on-chip architectures using multilayer deposited silicon materials for high-performance chip multiprocessors. ACM J Emerg Technol Comput Syst. 2011;7(2):1305–15.
41. Hendry G, Robinson E, Gleyzer V, Chan J, Carloni LP, Bliss N, Bergman K. Time-division-multiplexed arbitration in silicon nanophotonic networks-on-chip for high-performance chip multiprocessors. J Parallel Distrib Comput. 2011;71:641–50.
42. Chen C, Joshi A. Runtime management of laser power in silicon-photonic multibus noc architecture. IEEE J Sel Top Quantum Electron. 2013;19(2):338.
43. Zhou L, Kodi A. Probe: prediction-based optical bandwidth scaling for energy-efficient nocs. In: 2013 Seventh IEEE/ACM international symposium on networks on chip (NoCS), 2013. pp. 1–8.

44. Kodi A, Morris R. Design of a scalable nanophotonic interconnect for future multicores. In: The 5th ACM/IEEE symposium on architectures for networking and communications systems, ACM, 2009. pp. 113–22
45. Morris RW, Kodi AK. Power-efficient and high-performance multi-level hybrid nanophotonic interconnect for multicores. In: 4th ACM/IEEE international symposium on networks-on-chip (NoCS), 2010. pp. 207–14.
46. Bahirat S, Pasricha S. Uc-photon: a novel hybrid photonic network-on-chip for multiple use-case applications. In: 2010 11th International symposium on quality electronic design (ISQED), IEEE, 2010. pp. 721–9.
47. Xue J, Garg A, Ciftcioglu B, Hu J, Wang S, Savidis I, Jain M, Berman R, Liu P, Huang M, Wu H, Friedman E, Wicks G, Moore D. An intra-chip free-space optical interconnect. In: Proceedings of the 37th annual international symposium on Computer architecture, ISCA '10, (New York, NY, USA), ACM, 2010. pp. 94–105.
48. Gratz P, Keckler SW. Realistic workload characterization and analysis for networks-on-chip design. In: The 4th workshop on chip multiprocessor memory systems and interconnects (CMP-MSI), 2010.
49. Tan MRT, Rosenberg P, Mathai S, Straznicky J, Kiyama L, Yeo JS, Mclaren M, Mack W, Mendoza P, Kuo HP. Photonic interconnects for computer applications. In: Communications and photonics conference and exhibition (ACP), 2009 Asia, 2009. pp. 1–2.
50. Prabhu S, Grot B, Gratz P, Hu J. Ocin tsim-DVFS aware simulator for NoCs. In: Proc. SAW. vol. 1, 2010.
51. Bienia C, Kumar S, Singh JP, Li K. The PARSEC benchmark suite: characterization and architectural implications. In: The 17th international conference on parallel architectures and compilation techniques (PACT), 2008.
52. Hestness J, Keckler S. Netrace: dependency-tracking traces for efficient network-on-chip experimentation. Tech. Rep., Technical Report TR-10-11, The University of Texas at Austin, Department of Computer Science, http://www.cs.utexas.edu/~netrace, 2010.
53. Kim H, Ghoshal P, Grot B, Gratz PV, Jimenez DA. Reducing network-on-chip energy consumption through spatial locality speculation. In: 5th ACM/IEEE international symposium on networks-on-chip (NoCS), 2011. pp. 233–40.

Chapter 8
Design Automation for On-Chip Nanophotonic Integration

Christopher Condrat, Priyank Kalla, and Steve Blair

Abstract Recent breakthroughs in silicon photonics technology are enabling the integration of optical devices into silicon-based semiconductor processes. Significant developments in silicon photonic manufacturing and integration are enabling investigations into applications beyond that of traditional telecom: sensing, filtering, signal processing, quantum technology—and even optical computing. In effect, we are now seeing a convergence of communications and computation, where the traditional roles and boundaries of optics and microelectronics are becoming blurred. As the applications for opto-electronic integrated circuits (OEICs) are developed, and manufacturing capabilities expand, design support is necessary to fully exploit the potential of this technology. Photonic design automation represents an opportunity to take OEIC design to a larger scale, facilitating design-space exploration, and laying the foundation for current and future optical applications—thus fully realizing the potential of this technology.

This chapter describes our work on design automation for integrated optic system design. Using a building-block model for optical devices, we provide an EDA-inspired design flow and methodologies for optical design automation. Underlying these flows and methodologies are new supporting techniques in behavioral and physical synthesis. We also provide modeling for optical devices, and determine optimization and constraint parameters that guide the automation techniques. Starting from a logic design model, we describe how conventional logic synthesis and physical design techniques (placement, global and detail routing) can be applied in a top-down fashion to engineer a fully automated design flow for integrated optical systems.

This research was funded in part by a sub-contract to AFOSR grant FA9550-09-1-0661.

C. Condrat
Calypto Design Systems, Wilsonville, OR, USA
e-mail: chris@g6net.com

P. Kalla (✉) • S. Blair
Department of Electrical and Computer Engineering, University of Utah, Salt Lake City, UT, USA
e-mail: kalla@ece.utah.edu; blair@ece.utah.edu

R.O. Topaloglu (ed.), *More than Moore Technologies for Next Generation Computer Design*, DOI 10.1007/978-1-4939-2163-8_8

8.1 Introduction

Advancements in integrated optics are expanding the role of optical devices in system design. Opto-electronic integrated circuits (OEICs) [1], merging optics and control electronics on a monolithic substrate, are now a reality and enable optical integration in a diverse set of applications, such as sensing, signal processing, communications, and also computing [2–9]. The driving forces behind optics technology come from different, but inter-related areas. One area is *optical interconnects.* As semiconductors feature sizes have scaled downward, metal interconnects are now the dominant cause of delay and power usage in system design. In addition, the trend towards greater parallelism at the system level [10] has prioritized the role of communications in computing. Optics are therefore being pushed as an inter- and intra-chip *interconnect technology* to provide high-speed, long-haul, *low-power* communications [11–16].

A second driving force behind optical technology is that of manufacturing. Silicon is the mainstay of the semiconductor industry. The ease of manufacturing for semiconductors in well-characterized silicon-based processes, and steady improvements in performance and density at each process node makes CMOS-based technology the dominant computing manufacturing technology. For the same reasons, attempts were also made to develop silicon-based integrated optics—*silicon photonics*. Silicon's high refractive index and transparency to telecom wavelengths make it a suitable material for integrated optical waveguides, but silicon's use had traditionally been limited to *passive* optical devices such array waveguide gratings (AWGs) [17]. In the active domain, silicon's indirect band gap limiting silicon-based lasers, inability to detect light at telecom wavelengths, and slow modulation due to weak or absent electro-optic effects [18, 19] stymied nanophotonic device development. III-V semiconductor compounds such as gallium arsenide (GaAs) and indium phosphide (InP), or materials such as lithium niobate ($LiNbO_3$) would become the materials of choice for active photonic devices.

This changed in 2005, when Intel Corporation announced the first all-silicon optical modulator operating beyond the 1 Ghz threshold [20]. Fast optical modulation would be a significant breakthrough in silicon photonics, enabling viable optical networks to be fabricated in all-silicon processes. This development ushered in a number of subsequent breakthroughs in silicon photonic device development, including faster modulators [21, 22], hybrid lasers [23], and other device technologies [24] including hybrid silicon-germanium processes [25] for on-chip detectors. This promise of monolithic integration of OEICs in silicon-based processes opens the door to a great number of opportunities in system design. Already a number of architectures have been proposed for connecting systems via optical interconnect networks [14, 26], including as separate layers in 3D ICs. Investigations have also been made into optical digital signal processing [27], sensing, and even computing frameworks that can leverage optics in ways that would have been cost-prohibitive. In essence, silicon-based integrated optics are enabling the *convergence of computation and communication.*

As optical devices are integrated on larger scales, the need for design automation becomes apparent to handle greater levels of complexity in design. Scalability requires abstractions, which in turn enables and requires the use of optimization algorithms, design methodologies and tool-flows. What is now required is an Electronic Design Automation (EDA) type tool flow replicated and adapted to the optical domain. Such a *Photonic Design Automation (PDA)* represents an opportunity to take OEIC design to a larger scale, facilitating design-space exploration, and laying the foundation for current and future optical applications—thus fully realizing the potential of this technology. In this chapter, we describe the research and development efforts towards PDA by our research group: proposing a design automation flow with abstractions, optimization algorithms, tool-flows, and methodologies —enabling the synthesis of OEICs through automated means. We demonstrate our approach on *optical computing* applications, even though our approach is quite generic and not restricted to optical computing.

Contributions: In our work, we consider optical switching devices such as Mach-Zehnder Interferometers (MZI) and ring resonators connected by waveguides, splitters, couplers, detectors, etc., to form optical logic computing systems. We describe how Boolean logic synthesis techniques can be adapted to such a technology to design optical logic circuits. Subsequent to logic design, we describe a physical design methodology for placement and routing of optical circuits. We analyze technology-specific cost metrics during each stage of the synthesis process, and design algorithmic techniques that optimize for them. The chapter provides an overview of our contributions, and interested readers are referred to [2, 28–32] for more details.

Organization: The chapter is organized as follows: The following subsection depicts the photonic design flow proposed in this work. Section 8.2 describes the overall view of an integrated optical system. Section 8.3 covers the background and switching device models employed in this work. Section 8.4 describes the proposed photonic logic synthesis model. Section 8.5 describes the physical design automation methodology and cost models. Section 8.6 covers the global routing approach, whereas the detail routing model is elaborated in Section 8.7. Section 8.8 concludes the chapter.

8.1.1 The Photonic Design Automation Flow

Our design flow, depicted in Fig. 8.1, draws inspirations from EDA design flows and methodologies, and it consists of behavioral synthesis and physical synthesis. Ancillary to this flow is *technology modeling,* where the groundwork is laid for design automation in terms of building-block models and optimization metrics used throughout the design flow. System integration also plays an important role by introducing external constraints and effects on the optical system such as: area-limitations, packaging, and thermal interactions between on-chip heat sources and optical devices.

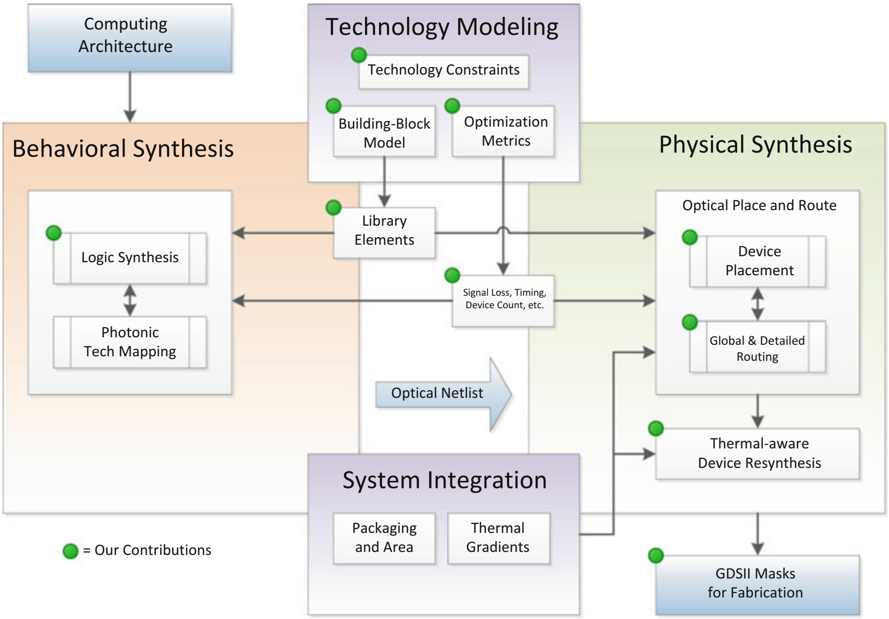


Fig. 8.1 The proposed design flow

8.2 Integrated Optic Systems

Figure 8.2 depicts a high-level view of an integrated optics system. We describe the components of this system and their operations; the details of the individual devices can be found in [33]. At the optical inputs of a system are lasers that provide light at the wavelengths the system is designed for, around $1,550\,\text{nm}$ for SOI systems. For silicon-based processes, this light is usually coupled into the system from outside using fiber couplers or grating couplers. To inject data into the system, modulation devices such as Mach Zehnder interferometers (MZIs), are used to vary the intensity of the input light. The light is then routed throughout the substrate using waveguides and optical switching devices with electrical switching inputs or in some cases employing all-optical switching.

The routing network also includes passive devices such as waveguide splitters, waveguide crossings, and passive multiplexing devices such as array waveguide gratings. Splitters divide the input among two outputs, with each output receiving half the input power, minus losses. Crossings are necessary for waveguides to cross each other on the single-layer planar substrate with minimal losses; crossings will feature into our physical design work in subsequent sections. Devices such as array waveguide gratings enable (de)multiplexing of various wavelengths, and have been a useful application for 1st-generation silicon photonics.

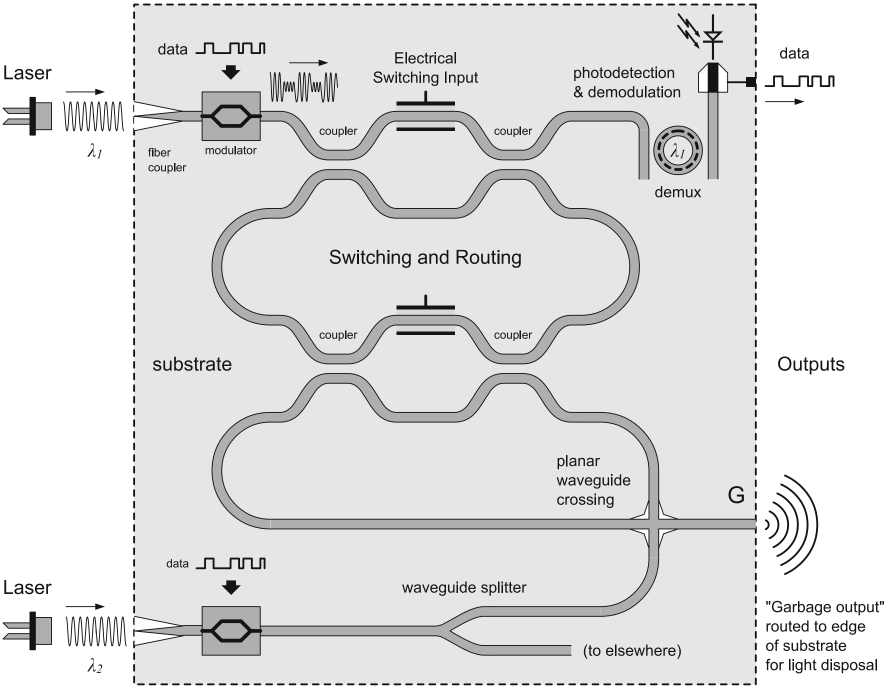


Fig. 8.2 High-level view of an integrated optic system

At the outputs of the system are demultiplexers for multi-wavelength systems, photodetectors and garbage outputs. Waveguides can support ranges of wavelengths, and therefore multiple channels of data may be present on a waveguide that need to be demultiplexed at the output. After demultiplexing, a photodetector (receiver) is required to translate optical signals into electrical signals, to read the transmitted data. Such photodetectors utilize materials such as germanium [34], which are incorporated into modern silicon photonics processes [35]. Finally, some routing networks need to dispose of unused light. To prevent interference and noise, the light from these “garbage outputs” must either be routed to the edge of the substrate for disposal, or absorbed by a material such as germanium, placed near the exit-point of the waveguide.

8.3 Device Models for Synthesis

One of the goals of this work is to develop synthesis techniques that utilize conventional integrated optics devices that can be fabricated with current technology, while also being applicable to future design processes. We describe the basic operation of the integrated optic devices we utilize.

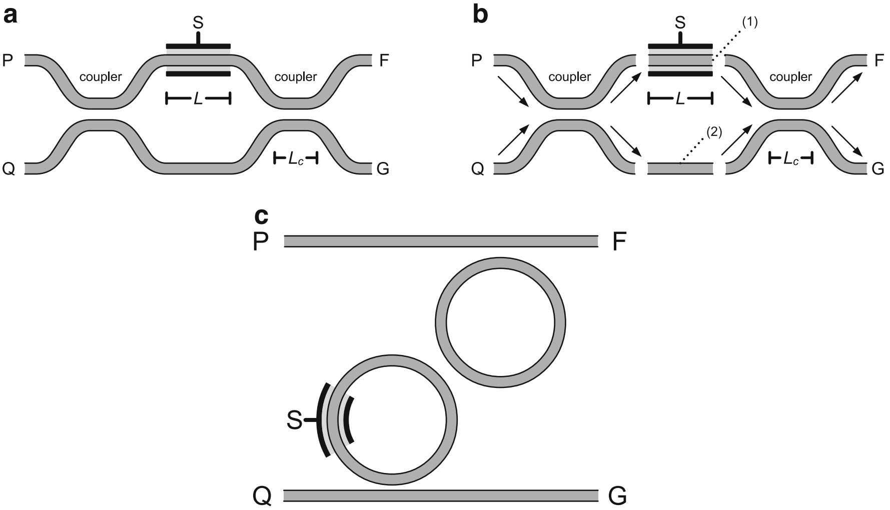


Fig. 8.3 Mach-Zehnder interferometer routing devices. (**a**) Mach-Zehnder interferometer (MZI); (**b**) MZI in parts; (**c**) Ring-resonator modulator

Routing light using waveguides is performed through the use of coupling and controlled interference. Consider the Mach-Zehnder Interferometer (MZI) depicted in Fig. 8.3a. The paths connected between P and F and Q and G are waveguides. Under certain conditions, when waveguides are brought in close proximity to each other, energy transfers between one waveguide to the other, and vice-versa. The couplers in this device are 3dB couplers, dividing and/or combining the signal from both inputs equally between the two outputs. The actual routing is controlled by input S, described by the following equations:

$$\phi_1 = \frac{\omega}{c} \cdot n \cdot L \qquad \phi_2 = \frac{\omega}{c} \cdot (n + \Delta n) \cdot L \tag{8.1}$$

$$\Delta\phi = |\phi_2 - \phi_1| = \pi = \frac{\omega}{c} \cdot \Delta n \cdot L \tag{8.2}$$

where ω is the angular frequency of the light (dependent on wavelength), ϕ_1 and ϕ_2 represent the phase of the light in the two center waveguides, and n is the index of refraction for the waveguide.

The input S is used to change the refractive index of Fig. 8.3b(1) by Δn via heating, carrier injection, or other means. This causes a path-length difference, and therefore a phase difference, between the signals in Fig. 8.3b(1) and b(2), causing constructive or destructive interference at the second coupler. A phase difference of 0 or π [36] will route each input *completely* to one output or the other, and the device acts as the controlled crossbar depicted in Fig. 8.4a. Similarly, other designs [16,37], as depicted in Fig. 8.3c, can be used to reduce the amount of phase-shift needed and the size of the overall device. Changing the refractive index can be accomplished by

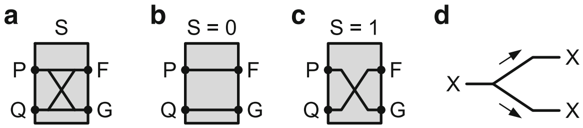


Fig. 8.4 Crossbar switch, and different routing configurations. (**a**) Gate; (**b**) Bar; (**c**) Cross; (**d**) Splitter

using a microheater or more advanced methods such as the MOS-capacitors used in Intel's high-speed modulator [20]. Modulation is also possible using devices such as ring resonators. The operation of such devices will be covered in later chapters. In our work, we can utilize either an MZI or ring resonators as an electrically controlled optical crossbar switch to design digital optical logic.

The operation of the MZI allows us to model it as a crossbar *gate* that routes light signal completely between two paths depending on the state of **S**, and depict it symbolically in Fig. 8.4a, with its two states Fig. 8.4b and c (bar and cross respectively). The waveguides are sourced by light (logical "1") or darkness ("0"), and the output of a function is read using optical receivers at the end. In our model, the switching input **S** is an *electrical* signal; it is an outside signal that controls the cross/bar configuration and cannot be switched by optical inputs. Connections to **p** and **q**, and **f** and **g** are waveguides, and for simplicity, light is assumed to move from the **p** and **q** side to **f** and **g**. In our model, *an optical signal cannot directly switch a crossbar's* $\boldsymbol{S}$ *input*[1]. More formally:

$$\begin{aligned}(S = 0) &\Rightarrow (P = F) \wedge (Q = G)\\ (S = 1) &\Rightarrow (Q = F) \wedge (P = G)\end{aligned} \tag{8.3}$$

These constraints affect how functions may be composed, and imply that the inputs to a crossbar are the primary inputs for that network. Waveguide connections between crossbar gates are depicted symbolically as black "wires." All designs created using the above model can be physically realized, including allowing waveguides to cross each other without interference.

In addition to MZIs, we also utilize *optical splitters,* depicted symbolically in Fig. 8.4d. A splitter divides the light from one waveguide into two output waveguides, each of which contain the original signal, but at half the power (a 3 dB loss). In our model, splitters are a significant signal degradation mechanism for a given topology; losses due to waveguide bends, waveguide crossings and insertion losses for MZI devices are the other mechanisms.

[1] Switching a crossbar gate with an optical signal requires an opto-electrical interface comprising an optical receiver unit feeding switching hardware. This can be expensive and slow, and is currently beyond the scope of the synthesis technique applied to this device model.

8.4 Optical Boolean Logic

Static-CMOS benefits from two important properties: metals and semi-conductors conduct when physically connected, and logic is restorative in nature. These two properties grant static-CMOS a great level of flexibility for implementing and optimizing logic functions, especially as it allows fanout for multi-level logic implementation. Unfortunately, this flexibility does not extend to optical circuits.

Consider the two networks in Fig. 8.5 implementing functions $f_1 = a + b$ and $f_2 = c \cdot (a + b)$. The first network implements f_2 by using the output of f_1 to drive the switching input of a gate. This is an unworkable design under our model, because an optical signal f_1 cannot switch the electrical input of another gate. A more optimal solution is found in the second design Fig. 8.5b, which uses f_1 as an *optical* input to another gate. This design benefits from using fewer gates, but more importantly, the sub-function is kept entirely in the optical domain. In such a way sub-functions *can* be shared, but with limitations.

8.4.1 Waveguide Splitters

The device which enables signal sharing using waveguides is the *waveguide splitter*. A waveguide splitter shares the signal of the input waveguide between two output waveguides, dividing the input power between two outputs, generally with a 50:50 ratio (3 dB loss). As the outputs of the splitter have only half the power of the original signal, there are limitations on how many may be used, which can serve as a cost-metric in the design of an optical logic network. Furthermore, as an optical signal, the sub-function may still only be switched and routed further using primary inputs to the network.

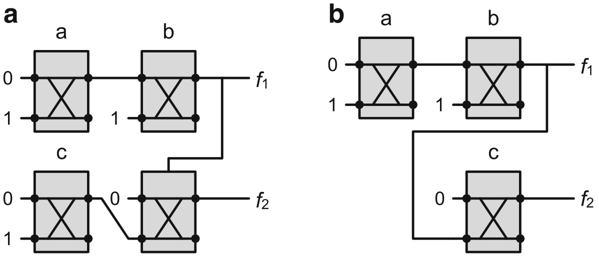


Fig. 8.5 Two configurations for $f_1 = a + b$ and $f_2 = c \cdot (a + b)$. (**a**) Incompatible design. (**b**) Compatible design

8.4.2 Garbage Outputs

A “garbage output” is a waveguide output that is not connected to a receiver (a function output), i.e. it is left unused. These unconnected outputs cause problems because the signals, and the light/energy it carries, may interfere with the operation of the network if not properly “disposed.” This is demonstrated in Fig. 8.6, which is the visual output of a Finite Difference Time Domain (FDTD) simulation [38] of an MZI device. The FDTD simulation technique models wave propagation through a (discrete) wave medium; Fig. 8.6 depicts the MZI device routing light from the top-left input to the lower-right output. The lower-right output of the device is left unconnected. Light arriving at this unconnected output can do a number of things, including dispersing into the substrate as noise and heat (as shown in the figure as ripples in the substrate) and/or reflecting back into the device, interfering with other signals.

These unconnected, or “garbage” outputs are problematic, and must be properly routed to the edges of the substrate where they can be dispersed away from the logic devices. The additional waveguides needed for this can cause congestion and complicate the overall physical routing of a network. *Every crossbar gate output that is left unconnected is a garbage output.* For example, the network shown in Fig. 8.5b would require three garbage outputs to be routed to the edges of the substrate, leading to a far-less compact design. Minimizing gate count, in general, reduces the number of garbage outputs, and is an important part of any synthesis procedure.

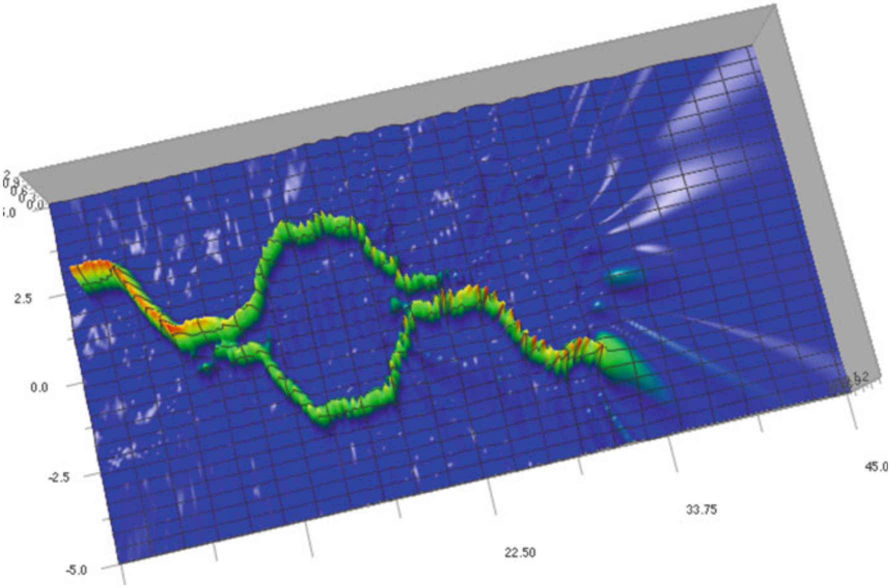


Fig. 8.6 Dispersion of light into the substrate from a garbage output

With these constraints in mind, we now explore two basic design styles/methods for creating optical crossbar logic networks: BDD-based design and Virtual Gate design. We show how these design styles operate, and highlight their abilities, as well as limitations. These limitations motivate more advanced approaches using Boolean decomposition as a means to derive designs that may be more optimal and beyond the ability of the other approaches to optimize for. All these described methods lend themselves to automation, and provide a comparison of these approaches near the end of the chapter, using metrics which are described in the coming sections.

8.4.3 BDD Based Design

The 2 × 2 crossbar can be modeled as two multiplexers with complemented inputs. As multiplexers, each crossbar gate effectively implements Shanon's expansion in one variable:

$$f = \bar{x} f_{\bar{x}} + x f_x \tag{8.4}$$

$$output_f = \bar{s}p + sq \tag{8.5}$$

$$output_g = sp + \bar{s}q$$

We can therefore utilize logic structures that employ Shanon's expansion, namely (Reduced Order) Binary Decision Diagrams (BDDs) [39] for direct implementation using crossbar gates.

Consider the ROBDD in Fig. 8.7a, which implements two functions: $f_1 = ab+c$ and $f_2 = \bar{a}b + c$, using variable order $a \prec b \prec c$. A dashed line indicates the negative cofactor, and a solid line the positive cofactor, which are connected to the p and q ports of a gate respectively. This is reflected in Fig. 8.7b. A crossbar network can therefore be technology-mapped from the BDD. The BDD's variable-switched function form directly maps to crossbar gate networks, and does not violate our crossbar model. In addition, the properties of the resulting network are also directly related to the properties of the BDD structure, including the effects of variable ordering on the canonical structure of an ROBDD.

8.4.3.1 Salient Features

A BDD-based crossbar network will, in general, have a number of garbage outputs equal to the number of nodes present in the BDD. The physical aspects of crossbar gates also mean that networks cannot take advantage of ROBDD extensions such as complemented edges as the signal in a waveguide cannot be "inverted" without extra hardware; complemented functions will need to be derived as separate BDD

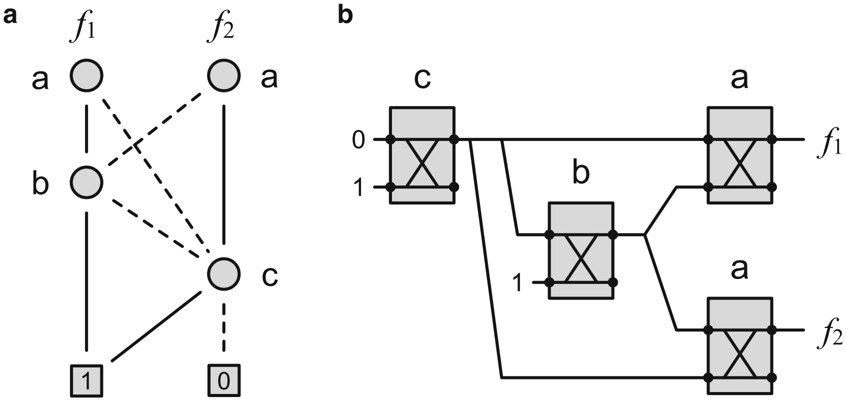


Fig. 8.7 BDD-based design for $f_1 = ab + c$, $f_2 = \bar{a}b + c$. (**a**) BDD Graph; (**b**) Resulting BDD-based design

function. Common subexpression extraction is possible in the form of shared functions is possible through the use of splitters; however, the effects of the signal degradation must be accounted for.

BDD-crossbar networks are relatively path-delay balanced, as they have a feed-forward design topology. The longest path is computed as:

$$l_{max} = h \cdot l_0 \tag{8.6}$$

where h is the height of the BDD graph.

Where BDD-based design suffers is in the number of garbage outputs produced by the approach. Each gate has the potential to produce a garbage output that must be accounted for through routing or a light absorbing structure. The canonical structure of ROBDDs can also lead to networks of extremely large gate counts for a given function. Though BDD-based design is attractive for its predictable signal delay, the number of garbage outputs and unpredictability of logic composition in terms of gate counts leads us to abandon this logic composition method for crossbar gate logic. We therefore investigate a composition methodology using "virtual gates."

8.4.4 Virtual Gates Based Design

Consider the device networks depicted in Fig. 8.8. We denote these logic composition functions "Virtual Gates" (VGs). A *virtual gate* (VG) is—functionally and conceptually—a crossbar gate that is switched by a *function,* not necessarily a primary input. The gate is "virtual" in the sense that it is a black box for a function composed of "real" gates—those driven by primary inputs—as well as other virtual gates. A novel form of nesting can be used to compose VG function

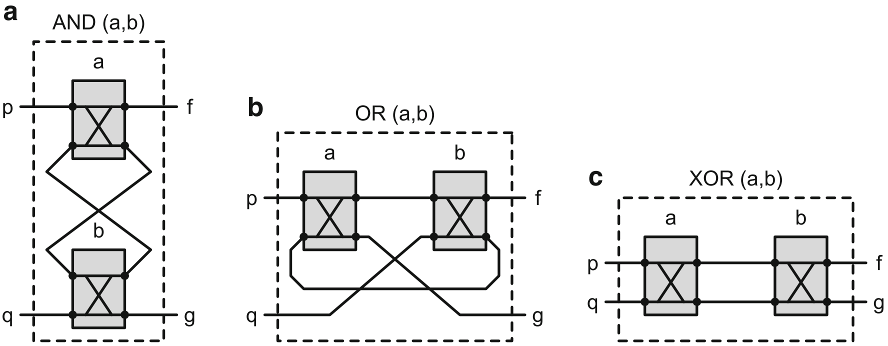


Fig. 8.8 Virtual gate functions for 2-input Boolean operators. (**a**) AND; (**b**) OR; (**c**) XOR

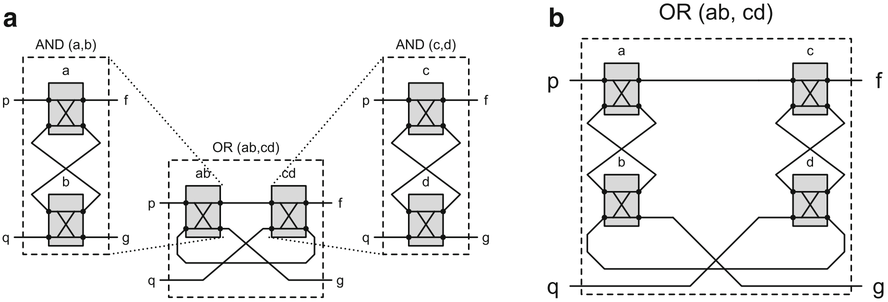


Fig. 8.9 Composing functions with virtual gates. (**a**) Virtual gates implementing $f = ab + cd$; (**b**) Resulting network

implementations, where Boolean operators are implemented by replacing child gates with other gates, a real or virtual.

A given VG implementation comprises two input waveguide ports **p** and **q** connected by waveguides and crossbar gates to two output ports **f** and **g**. The nesting operation comprises the Boolean operator forms depicted in Fig. 8.8, and is illustrated in Fig. 8.9a where two AND virtual gates are nested within an OR virtual gate, creating the final function $ab+cd$. Evaluation of a VG, given a primary input assignment, involves assigning **p** and **q** inputs logical 0 and 1 respectively, and applying *cross* or *bar* configurations to gates as defined in Fig. 8.4. The output of the function is detected at **f**, with $\mathbf{g} = \neg\mathbf{f}$.

The process of composition is illustrated in Fig. 8.9a, where a function $f = ab + cd$ is implemented by replacing (or *nesting*) the gates of an OR function with VGs implementing $a \cdot b$ and $c \cdot d$. The result is depicted in Fig. 8.9b.

While it may seem strange to see feedback loops in device designs, the physical devices can indeed implement self-feedback. As an experiment, the model for the AND gate depicted in Fig. 8.8a was simulated in a 2D FDTD simulator

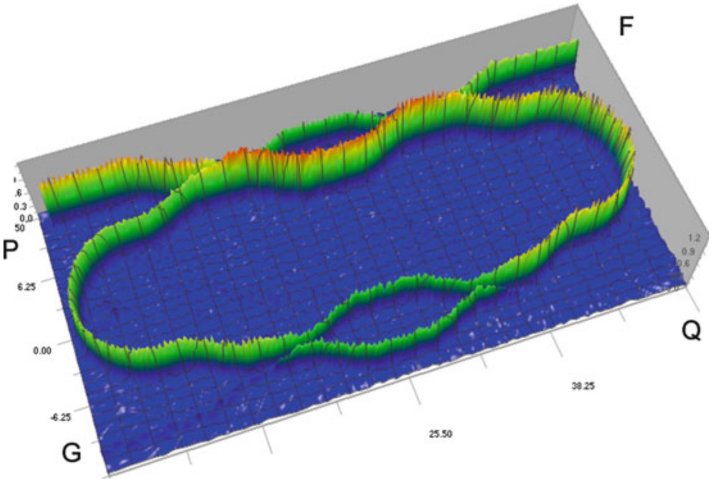


Fig. 8.10 FDTD simulation of an AND virtual gate

OptiFDTD®by Optiwave Software; the visual output[2] of which can be seen for $a = 1, b = 0$ in Fig. 8.10. The signal from the top-left crosses in the top gate, but passes through in the bottom gate, returning to the top gate where it crosses again to appear in the top-right output.

8.4.4.1 Salient Features

Networks composed of virtual gates have exactly two optical inputs p and q and two outputs f and g, as the entire network is, in itself, a virtual gate; in addition, for a given function, a maximum of one garbage output is created. The existence of a complete logic enables virtual gates to implement any logic function using crossbar gates *comprising only primary inputs.* This includes factored functions, and any other single-output representation using Boolean operators. Control signals (S) are connected via the primary inputs of the function. The f port implements the function, and $g = \neg f$. Furthermore, the total number of *real* gates is the number of primary literals in the original logic expression the network is derived from.

[2] Note that there are differences from the virtual gate diagram: the bottom two ports are swapped because the waveguides are not crossed in the center, and that the "light" source is positioned at the p input rather than at the q input.

Virtual gates also suffer from very unbalanced signal paths, depending on the state of the switches, with the potential for a signal to traverse every waveguide present in a VG network. The maximum signal path l_{max} is roughly computed as:

$$l_{max} = 2 \cdot p \cdot l_0 \tag{8.7}$$

where p is the number of operators in the virtual gate, and l_0 is a "unit length" of waveguide. This is based on the fact that all virtual gate operators connect two gates (virtual or real) by two waveguides, and a signal could possibly traverse all paths to reach the destination. For example, the network in Fig. 8.9a would have a $2 \cdot 3 \cdot l_0 = 6l_0$ long maximum signal path, which is close to the longest possible signal path from p to f with variable assignment $\{a, b, c, d\} = \{1, 0, 1, 0\}$ at $5l_0$. The value l_{max} is a reasonable rough estimate; it can be further refined by estimating routing distances for operators and physical network topology.

8.4.4.2 Expression Sharing

The major limitation of designing with virtual gates is that the nesting of gates *prevents the extraction/sharing of arbitrary common sub-expressions (CSE).* For example, in Fig. 8.11 one cannot simply share the ab term from $f = ab + bc$ for use with another gate; assignments such as $abcd = \{1, 1, 1, 1\}$ will cause all crossbar gates to assume a cross-configuration, isolating the top input of the h-gate from the optical inputs of the network. In effect, any operator employing feedback for its inputs can produce an undefined state. Only the XOR operator does not exhibit this behavior as it has no feedback, but XOR-based CSE is not well studied in contemporary logic synthesis. To address this issue particularly for optical logic synthesis, we investigated a XOR-based functional decomposition technique for CSE, and implemented it within our virtual-gate paradigm. Interested readers may refer to our publication [2] for more details.

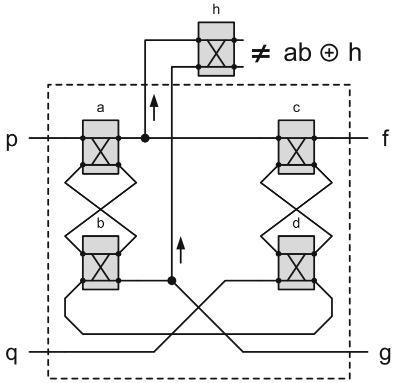


Fig. 8.11 Internal functions of virtual gates cannot be shared

8.5 Physical Synthesis Methodology for Integrated Optics

Subsequent to high-level and logic design, the need for automated design space exploration and optimization also begins to appear for *physical synthesis of integrated electro-optical systems*. For this reason, the Electronic Design Automation (EDA) community is investigating how automatic design space exploration techniques can be adapted to the photonics domain [40–44]. Such circuits are complex in their device interconnections, often featuring high device counts and large amounts of feedback loops. These designs comprise a set of pre-designed optical devices—modulators, switches, splitters—placed on a planar substrate, connected together via waveguides. For example, in our previous work [2], our multi-level logic synthesis methodology for implementing logic demonstrates how optical designs can scale beyond the ability of custom design. The physical synthesis of such applications now has to be addressed. For this purpose, we describe the design constraints, layout models and methodologies for integrated optics automation.

8.5.1 Design Constraints

At the physical automation level, we identify signal power and substrate area as our core guiding metrics.

8.5.1.1 Signal Power

Signal power is the primary guiding metric in our methodology. All devices, including bulk waveguides, have insertion losses, measured in decibels (dB). Our assumption is that these losses are pre-characterized through device-analysis (FDTD, etc.) for the following type devices:

- **Pre-designed devices [device-specific]** (e.g. modulator devices, switches, splitters, etc.). Losses are characterized from inputs to outputs. For example, waveguide splitters have their signal power from the input effectively halved at each output (a 3dB loss).
- **Waveguide crossings [0.1–0.2 dB / crossing]** Per-crossing losses are on the order of 0.1–0.2dB per crossing [45–47], affecting both crossing waveguides.
- **Waveguide bends [0.001–0.3 dB / bend]** Losses dependent on inherent waveguide properties (materials, geometry, etc.), radius of curvature of the bend, and surface roughness due to fabrication [48–50].
- **Bulk waveguides [0.01–2 dB / cm]** As these losses are extremely low (dB per *centimeter*, e.g. 0.03dB/cm [51]), we consider bulk waveguides essentially *lossless*.

Losses due to the presence of pre-designed devices are effectively fixed. Therefore, the design automation problem concerns itself with designing within the permitted losses *between* such devices–the routing fabric. We identify three main routing loss mechanisms in descending importance: 1) *waveguide crossings,* which induce a relatively large fixed loss per crossing; 2) waveguide bends, especially bends close to the minimum radius of curvature; 3) bulk waveguides, which generally have low losses; however surface roughness can induce losses over larger distances for smaller waveguides.

8.5.1.2 SOI Waveguides

Si-photonic waveguides, with their large refractive index differentials, provide strong mode confinement, and therefore bends can be much sharper, saving area. While waveguide bends can be effectively lossless given a large enough radius of curvature, accepting small per-bend losses can be advantageous in reducing the area occupied by a bend [49]. The choice of minimum routing grid size can therefore affect the weighting of metrics used to guide the routing, whether losses due to bends, waveguide crossings [45], or area.

8.5.1.3 Area

Many optical devices, such as those used for switching, are designed such that their input and output ports appear on only opposing sides. This feed forward device design often extends to the device networks as a whole, resulting in overall networks that are very wide. Wide substrates may not be desirable when integrating optics into designs, and a more suitable aspect ratio may need to be enforced. The side-effect of this is that devices must be rearranged on the substrate in a manner that can affect inter-device locality as well as increase waveguide routing complexity. This becomes an important part of the placement phase of our methodology.

8.5.2 Methodology

We propose the following methodology for the overall physical design problem for integrated optics. As depicted in Fig. 8.12c, pre-designed optical devices are represented as rectangular blocks (a) that are arranged (placed) in fixed-width columns (b). Such a placement gives rise to *vertical routing channels* (c), which are routing regions that separate the placed devices. Waveguides are routed between devices at "ports" (d) that face the channels. For ports in different columns, these waveguides may pass through *horizontal routing channels*, as depicted in (e). While the substrate is planar, waveguides may also cross each other perpendicularly (f) without sharing signals.

Overall, the physical design methodology requires that the problem be solved in three steps:

- Placement of optical switching devices into columns, i.e. a grid-based layout.
- Global routing of waveguides that connect these devices. Global routing solution will determine the overall routing topology of all the nets.
- Detailed routing of all the nets, which manifests itself as a well-defined channel routing problem.

While this methodology is analogous to that employed in the VLSI domain, the design and optimization constraints imposed by the optical technology are different. Any CAD solution to this problem will have to incorporate such technology specific constraint models and design rules.

8.5.3 Device Placement

Pre-designed optical devices are placed into columns. Consider the layout of devices in Fig. 8.12a. While devices maintain ports on only their left and right sides, connections may be made to any other device in the network by routing through vertical columns and between columns. In such a manner, connectivity is preserved, but the overall network has a smaller aspect ratio. Placement techniques, such as those used for row placement and chip floorplanning [52], can therefore be employed for placing devices within an optical substrate. The placement of devices into columns enables us to utilize routing techniques designed for such placement strategies. In our applications we use the Capo placer [52] to arrange devices in rows given a specific aspect ratio. Connected devices are localized as much as possible, reducing congestion. Such a placement simplifies the subsequent signal-loss constrained global and local routing methods for integrated optics.

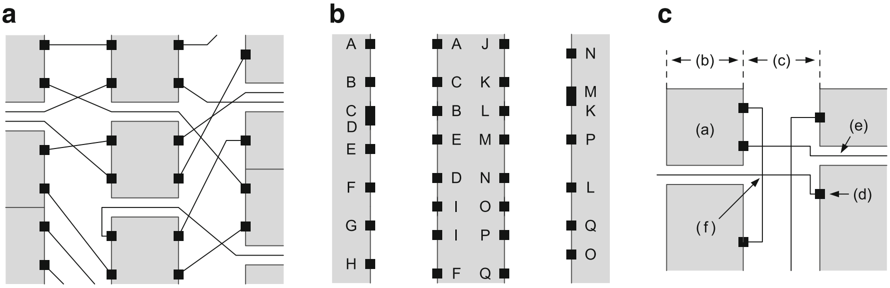


Fig. 8.12 Stages of the Physical Design Methodology. (**a**) Columns of optical devices, and global routes; (**b**) Resulting channels for detailed routing; (**c**) Ports, routes and channels

8.5.4 Global and Detail Routing

Global routing determines the high level topology a signal may take through the channels from source to destination. The chosen routes induce bends and crossings with other nets. The optimization goal of the global router is to minimize losses due to waveguide crossings and waveguide bends. In addition, global routing also takes into account overall net lengths and routing congestion.

Given a device placement, a graph is derived from the vertical and horizontal channels separating the device blocks. Nodes are placed at locations where ports are located, and where horizontal and vertical routing regions meet. Any device placement topology may also be used; however, we assume a channel-based placement is used. In a channel-based placement, such as depicted in Fig. 8.13a, nodes and edges are first derived for the vertical channels from the location of device ports and horizontal channels. These channels are then connected to other channels via horizontal inter-channel edges, such as depicted in Fig. 8.13b.

The presence of inter-route loss can affect signal quality more than route length, forcing longer routes to be exercised. Consider the nets in example Fig. 8.13c, where a net q can utilize one of two distinct routes (1) and (2). Route (1), though shorter than route (2), must cross the chosen route for p; to avoid the crossing, route (2) could be utilized. Route (2), however, crosses over the chosen route for r. Should route r have less stringent loss constraints than p, route (2) may be chosen over (1), despite a longer overall path. Ultimately, the final route choice is derived from a combination of all loss factors.

The global router provides a set of vertical routing channels with net/port connectivity, such as depicted in Fig. 8.12c. At this stage, detail routing is performed, determining the actual placement of horizontal and vertical connections

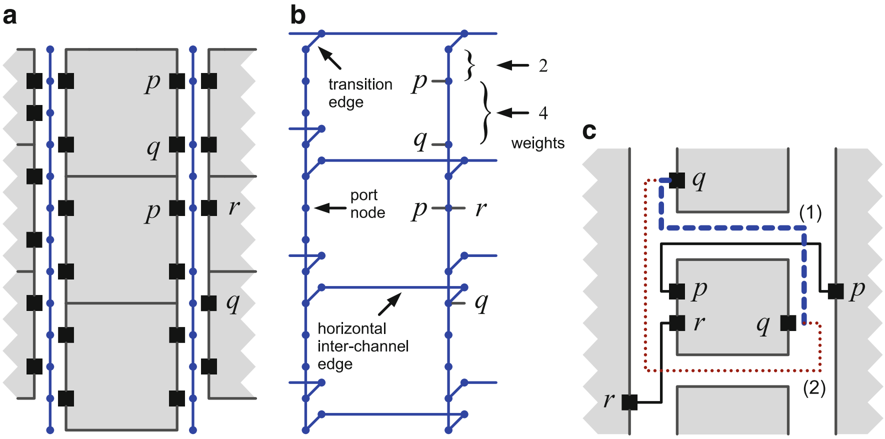


Fig. 8.13 Construction of routing graph from channel layout. (**a**) Vertical nodes and edges from layout; (**b**) Complete routing graph; (**c**) Different route choices inducing different crossings

within the vertical channel. Consider the routing channels depicted in Fig. 8.12b. The channel routing area is a grid between the pins on either side of the channel, where waveguides are routed between pairs of pins. Traditional VLSI channel routing seeks to minimize the area of a fully routed channel. In our channel routing techniques, we optimize for crossings and bends, with channel height a subsequent metric. The details of our channel routing approaches are found in Sect. 8.7.

8.5.5 Routing Grid Realization

The result of routing algorithms must be transformed into the physical waveguide layout. This entails converting the routing grids into waveguide bends satisfying the material bend constraints, which are generally defined in terms of minimum radius of curvature and coupling distance.

A rectilinear routing grid is realized as waveguides by converting all 90° grid transitions to 90° waveguide bends. This requires that such bends complete within a quarter of the routing grid. This is illustrated in Fig. 8.14b where a horseshoe-shaped bend utilizes two 90° waveguide bends, each taking place within a quadrant of the routing grid. This mapping represents the smallest grid that can be suitably used for complete routing grid flexibility.

The physical routing can also exploit the spacing between curves at the corners of grids. These "knock-knee" style bends, as depicted in Fig. 8.14c, enable additional track sharing—potentially reducing the overall number of tracks needed for a routing. For example, in the solution depicted in Fig. 8.15b, the knock-knee bends between signals C-E, F-G, D-I, and G-J allow each respective pair to occupy the same track, with the net effect of reducing the total number of tracks to four (4). Routing techniques enabling knock-knee track sharing must account for shared rectilinear grid locations, e.g. Fig. 8.15a, during channel construction.

The waveguide's minimum radius of curvature r has an important role in determining the routing grid's minimum size. In some cases, r may be chosen for

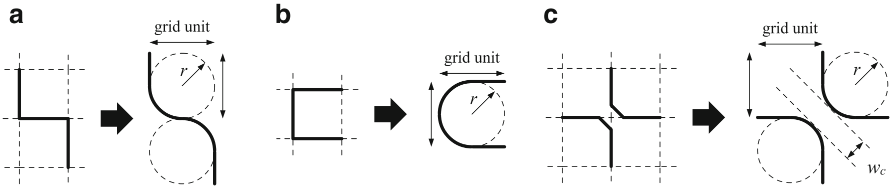


Fig. 8.14 Conversion of grid units to waveguide curves. (**a**) S-shaped grid to bends; (**b**) Horseshoe-shaped grid to waveguide bends; (**c**) Knock-knee grid with 90° bends, radius of curvature r, and minimum coupling distance w_c

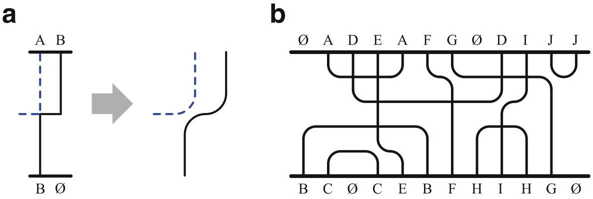


Fig. 8.15 Knock-knee model for grid spacing. (**a**) Shared grid corners enable knock-knees; (**b**) Channel routing incorporating knock-knee bends (Four tracks, eight crossings)

area reduction, at the expense of per-bend losses [49]. For example, to enable knock-knee routing patterns, the distance w_c in Fig. 8.14c must be sufficient to prevent significant coupling between waveguides.

8.6 Global Routing for Integrated Optics

Global routing provides the high-level overall placement of routes throughout the device network, while detailed routing determines the localized routing necessary to complete routing. In the VLSI domain, global routers are mostly concerned with 1) wire-length, 2) congestion, and 3) overflow—all of which are interrelated. For integrated optics, some of these aspects, such as wire-length, are deemphasized and in their place the router must also account for signal loss in terms of crossings and bends. In addition, optical routing must be performed on a *single routing layer.*

Despite the advanced state of VLSI global routers [53–56], their applicability is limited within the optical routing domain. VLSI routing is inherently multi-layer, and VLSI global routers are designed to take advantage of multiple layers in order to produce routing solutions. As such, global routers are also not designed to minimize crossings. Minimization techniques for metrics such as vias, though applicable to bends, cannot be applied to waveguide crossings, as a single via can facilitate multiple crossings due to multiple-layers. We therefore investigate global routing specifically for integrated optics.

8.6.1 Routing Using Mixed Integer Linear Programming

We conceptually frame global routing through mixed integer linear programming (MILP). Each net i has a set of n_i candidate routes. Only one route k may be chosen for a given net, and that route has a cost associated with it α^i_k. The cost α^i_k is formulated in terms of signal loss: *static* losses induced by bends and length, and *inter-route* losses caused by crossings between of different routes of different nets. More formally

$$\alpha_k^i = \alpha_{k,static}^i \cdot x_k^i + \sum_j^{j \neq i} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} \alpha_{k,l}^{i,j} \cdot x_{k,l}^{i,j} \tag{8.8}$$

$$\alpha_{total}^i = \sum_{k=1}^{n_i} \alpha_k^i \qquad \alpha_{total}^i < \alpha_{max}^i \tag{8.9}$$

where $x_k^i = 1$ if net i uses route k, otherwise 0, and $x_{k,l}^{i,j} = 1$ if respective nets i and j use routes k and l, respectively, otherwise 0. A loss-coefficient $\alpha_{k,l}^{i,j}$ is the inter-route loss associated with those two routes. α_{total}^i constrains the maximum losses acceptable for the given net. Equations (8.8) and (8.9) provide the basic structure for optimization. What remains is to determine the coefficient weights $\alpha_{k,static}^i$ and $\alpha_{k,l}^{i,j}$.

8.6.2 Route Analysis

Routes are defined on a graph G comprising a set of grid-edges E derived from layout of the device placement. For example, the routing regions between devices in Fig. 8.16 produce a set of edges connecting between port-endpoints.

Static route costs $\alpha_{k,static}^i$ are derived from the set of edges E_k^i that a route traverses, comprising a sum of edge-cost weights. Waveguide bends also have a cost associated with them, as they can be a significant loss mechanism. In order to penalize their use, we modify the graph by adding weighted *via-edges* connecting between vertical and horizontal edges, as depicted in Fig. 8.16. Though straight-waveguide losses at these scales are negligibly small, longer routes have a greater potential for intersecting other routes, potentially causing more crossings. Edges are therefore weighted according to their length in the substrate to favor locality. For simplicity, a basic approach for choosing the set of candidate routes for a given net is to choose the set of routes with the least static route-costs. Other metrics, such as potential edge-capacity utilization, and diversity of routes may also provide better route candidates.

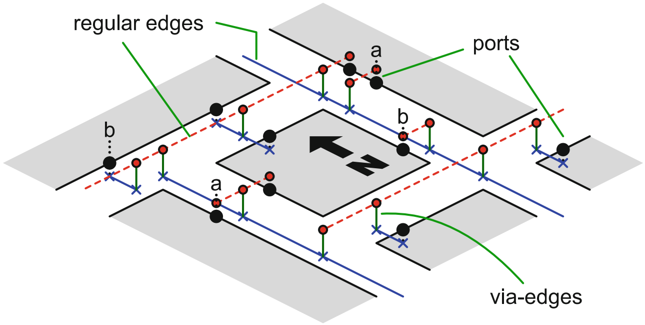


Fig. 8.16 Routing graph derived from device placement

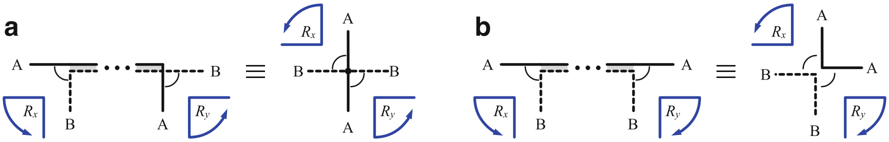


Fig. 8.17 Functionally equivalent configurations of path-endpoints and their rotations. (**a**) Same direction ($A \rightarrow B$) induce crossings; (**b**) Opposite directions require no crossings

Given the sets of candidate routes, inter-route losses $\alpha_{k,l}^{i,j}$ are determined by pairwise analyzing candidate routes for different nets to determine whether routes cross. The given pair of routes may have multiple shared sets of edges (paths) where a crossing may occur. A crossing, if it is required, will occur only once for a given shared path; we can treat the shared path edges as a single node that retains the crossing of the original.

Consider the two nets depicted in Fig. 8.17, where route A and B share edges in the middle. At the endpoints of the shared edges, the two routes diverge; we denote these as diverging endpoint edges (DEEs). For a given endpoint, *rotation* is the direction a route's DEE must rotate towards DEE of the other route, pivoted on their shared node, on the arc that does not contain the shared route edges. For example, in the left path endpoint of Fig. 8.17a, the DEE of A rotates counter-clockwise towards the DEE of B. Likewise, on the right side, DEE of A again rotates counter-clockwise towards the DEE of B. *A crossing is only required if-and-only-if the rotation of both endpoints is the same, otherwise no crossing is required.*

8.6.2.1 Minimization Function

With the per-route and per-net equations in place, and their coefficient weights determined, the final minimization function is a sum of all net costs. The minimization function is implemented as

$$minimize : \sum_{i=1}^{m} W_i \cdot \alpha_{total}^{i} \tag{8.10}$$

where m is the total number of nets and each W_i is a per-net weight to prioritize certain nets over others during optimization. Though not detailed here, congestion can be accounted for by the number of routes that utilize given edges in the routing graph.

The presented global router is relatively basic as compared to contemporary VLSI routers; however, it accounts for many aspects specific to single-layer integrated optic routing. For VLSI routers, crossing minimization at a global level is not incorporated. Therefore, instead of utilizing VLSI-centric global routers, we developed our own global router for integrated optics. Subsequent to the global routing, the final detail routing problem is formulated and solved as a channel routing problem.

8.7 Channel Routing for Integrated Optics

In column-based optical device placement, the detailed routing problem manifests itself as a *channel routing* problem, where (Silicon) optical waveguides are fabricated on a planar substrate and are connected to devices at the ends of the channel. Planar routes require waveguides to bend (curve) and cross each other—causing loss of signal power. Channel routing techniques are therefore needed that minimize waveguide crossings and bends. We present a channel router based on crossing-aware, graph-constraint track-assignment. The router minimizes signal loss as a function of waveguide crossings and bends within the channel, while also reducing area.

8.7.1 Optimization Objective

The primary optimization objective in our routing formulation is *signal loss* minimization. Within the channel, this is achieved by: 1) minimization of the total number of waveguide crossings; and 2) minimization of the number of waveguide bends. Minimization of the number of tracks (channel height) is the subsequent secondary objective.

We optimize for the *total signal loss within the channel* due to optical feedback within the system. A signal may be routed such that it enters a given channel multiple times and may cross multiple other nets. Therefore, instead of minimizing losses on a per-net basis, we minimize for *total* losses within a channel (Fig. 8.18).

8.7.2 Left-Edge-Style Channel Routing

Traditional left-edge-style channel routers [57–59] represent the channel routing problem using horizontal and vertical constraint graphs (HCG, VCG). An alternate representation of the HCG is the zone representation, which is derived from the HCG, where every zone is defined by a maximal clique. The number of signals in the largest zone is the lower bound on the number of tracks needed for routing. These graphs encode constraints on how tracks may be assigned to nets in the channel. Consider the channel routing problem depicted in Fig. 8.19a. The resulting zone representation is depicted in Fig. 8.19b. Likewise, the VCG for the problem is represented in Fig. 8.20a.

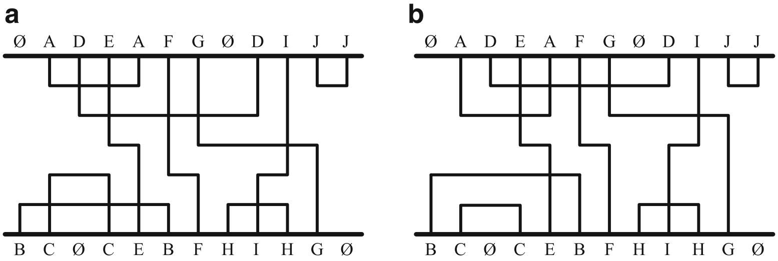


Fig. 8.18 Channel routing solutions under differing constraints. (**a**) Track-optimized (five tracks, ten crossings); (**b**) Crossing-constrained (five tracks, eight crossings)

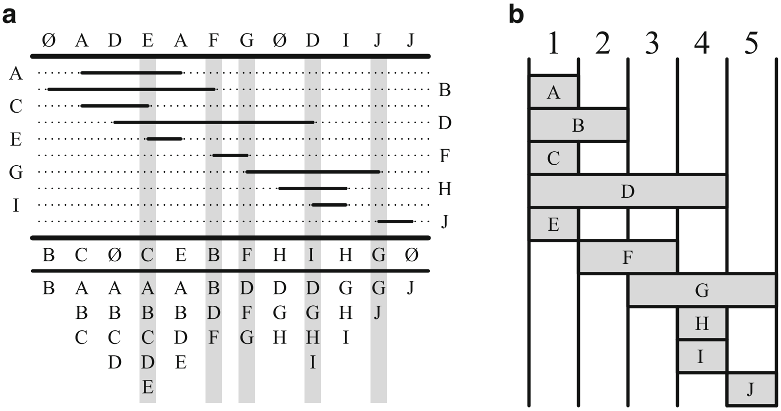
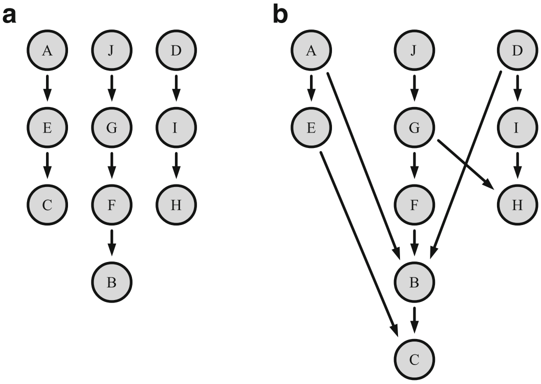


Fig. 8.19 Horizontal constraints and zone representation. (**a**) Five maximal subsets of signals; (**b**) Resulting five zones

Fig. 8.20 Crossing-constraints modifications to the VCG. (**a**) Original VCG; (**b**) With crossing constraints



A net may be assigned to a track should it have no descendants on the VCG, and have no overlapping zone conflicts with previously assigned nets on a given track. Nets are removed from the VCG as they are assigned to tracks. When a track cannot contain more nets, a new track is created and the process is repeated until no more nets are left for assignment.

Multiple nets can be candidates for assignment to a given track, each with different horizontal overlaps. Therefore heuristics are used to choose which nets are assigned first. One of the simplest is a greedy heuristic used in *constrained left-edge channel routing* [57], where the left-most available nets in channel are assigned first to tracks. This can lead to sub-optimal track-utilization; more sophisticated heuristics analyze the graph structure for better results, such as [59], which attempts to reduce the longest path in the VCG for better track utilization. We refer to the class of track assignment algorithms above generically as "left-edge-style" channel routing. The approach we describe below can be incorporated into any such techniques.

8.7.2.1 Crossing-Constrained Track Assignment

Figure 8.18a depicts the output of a (VLSI) left-edge 2-layer channel router, and Fig. 8.18b, a channel routing constrained for crossing-minimization. Both solutions are minimal in terms of tracks; however, the total number of crossings in Fig. 8.18a is 10, compared to 8 in Fig. 8.18a. The discrepancy in the number of crossings is attributed to the two crossing points caused by nets B and C. By forcing C to appear below B, two crossings are avoided. However, transforming from Fig. 8.18a to b is not as simple as moving net C below B, not if track height is to be kept minimal. Crossing minimization must therefore be encoded into the routing process itself as constraints.

We constrain the channel routing problem to favor crossing minimization. The VCG is modified such that avoidable crossings impose vertical constraints on the net ordering. Only nets that share zones have the possibility of crossing, and pairwise analysis takes place after the zones are derived.

A crossing constraint is only encoded into the VCG if a crossing can be avoided. For example, the pair of nets in Fig. 8.21c would not normally be constrained in the VCG; however, a net crossing can be avoided if B is assigned a track above A. Therefore, an edge connecting B to A is added to the VCG. Conversely, the two nets in Fig. 8.21b cannot avoid crossing, and therefore no constraint is added.

We introduce the concept of *pin-rotation* to detect avoidable crossings. If we were to map the pins of nets on a unit circle, a crossing is unavoidable if rotating from one pin to the next is not possible without first passing through the pin of another net. Consider the nets depicted in Fig. 8.21a. Collapsing the shared horizontal region, and considering the areas Fig. 8.21a(1) and a(2) shows how pins of a given side rotate with respect to each other (clockwise/counter-clockwise) around an axis fixed at the center. In the case of Fig. 8.21a(1), the rotation of the left pin of A to the left pin of B is *counterclockwise,* and likewise the pins on the right-side also rotate in the same *counterclockwise* direction. If the pins on both left and right terminals rotate in the same direction a crossing is unavoidable.

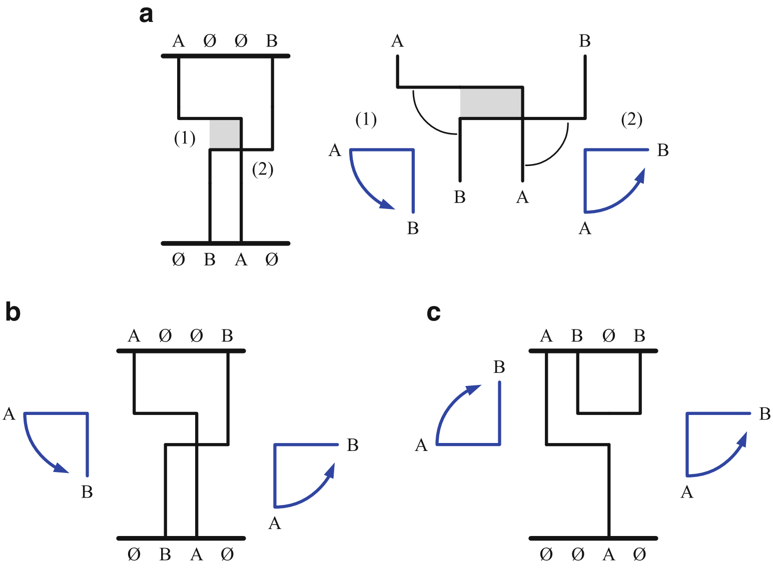


Fig. 8.21 Crossing detection via rotation from A to B. (**a**) Rotation direction with respect to pin locations; (**b**) Same rotation direction ⇒ Unavoidable crossing; (**c**) Opposite rotation directions ⇒ Avoidable crossing

More formally:

$$X^{left}_{A,B,CW} = \begin{cases} X^{left}_{B,top} & \textbf{if } C^{left}_{A} < C^{left}_{B} \\ \neg X^{left}_{A,top} & \text{otherwise} \end{cases} \tag{8.11}$$

$$X^{right}_{A,B,CW} = \begin{cases} X^{right}_{A,top} & \textbf{if } C^{right}_{A} < C^{right}_{B} \\ \neg X^{right}_{B,top} & \text{otherwise} \end{cases} \tag{8.12}$$

$$X_{\text{avoidable}}(A, B) = \left(X^{left}_{A,B,CW} \neq X^{right}_{A,B,CW} \right) \tag{8.13}$$

where $C^{left/right}_{N}$ is the integer-valued column-position of a pin of net N on a given side (left, right); the Boolean variable $X^{left/right}_{N,top}$, using the same notation, denotes whether that pin resides on the top side of the channel. Equations (8.11) and (8.12) utilize the horizontal relationships of pins and their channel-sides (top/bottom) to determine the *clockwise rotation* (CW) of a given pair of *left* or *right* pins for nets A and B, rotating from A to B. A crossing is avoidable only if left and right rotations are *not* the same, the result of (8.13).

For example, in Fig. 8.21a, consider the left side of the shared span Fig. 8.21a(1):

- The variables C_A^{left} and C_B^{left} are the column positions of the respective left-terminals of nets A and B. In the example, $C_A^{left} = 1$, $C_B^{left} = 2$.
- $C_A^{left} < C_B^{left}$ implies $X_{A,B,CW}^{left} = X_{B,top}^{left}$ from (8.11).
- The left pin of net B is *not* on the top side of the channel ($X_{B,top}^{left} =$ **false**). Therefore, the left side of the pair of nets is *not* rotating clockwise from A to B, i.e. $X_{A,B,CW}^{left} = X_{B,top}^{left} =$ **false**.
- On the right side of the shared span Fig. 8.21a(2), $C_A^{right} < C_B^{right}$. This condition implies that $X_{A,B,CW}^{right} = X_{A,top}^{right} =$ **false**. The right side is therefore also *not* rotating clockwise from A to B.

Having the same direction of rotation ($X_{A,B,CW}^{left} = X_{A,B,CW}^{right} =$ **false**) implies that a crossing is *unavoidable,* as determined by (8.13); this is reflected in the figure.

Applying crossing constraints to the problem depicted in Fig. 8.19a results in the VCG depicted Fig. 8.20b. As compared to the original VCG Fig. 8.20a, the crossing-constrained VCG is more heavily constrained, ensuring that unnecessary crossings do not occur, such as the double-crossing of nets B and C in Fig. 8.18a.

8.7.2.2 Knock-Knee Track Sharing

Though the modified VCG is effective in preventing waveguide crossings, the additional constraints can affect overall track height, and may produce a worse solution in terms of number of tracks. However, we observe that the bend geometry of optical waveguides can be exploited to further reduce channel height. This is discussed below.

Consider the two nets in Fig. 8.22a. The endpoints of the two nets occupy the same column and therefore net A should be placed above B in the VCG. However, given the same track, the two nets would intersect at a corner of each horizontal

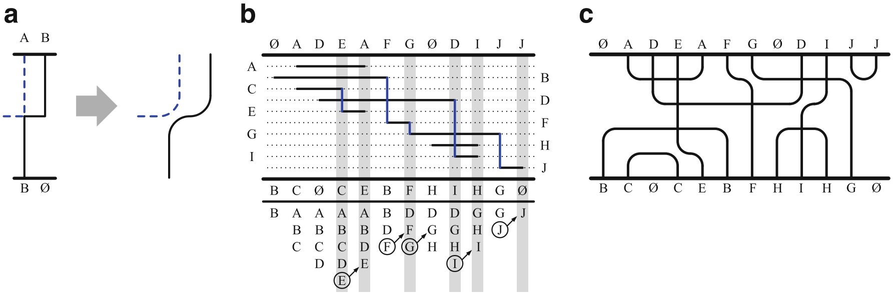


Fig. 8.22 VCGs for Fig. 8.19a and knock-knee extension. (**a**) Knock-knee implementation; (**b**) Knock-knee-constrained zone representation; (**c**) 4-track routing solution utilizing knock-knees

span—a *knock-knee*. In VLSI, this situation is untenable, and different tracks would need to be assigned to each net. However, for waveguides, the minimum grid spacing for a channel can permit knock-knees in the routing grid. This is depicted in Fig. 8.22a, where a track is shared between the two nets without overlap.

A knock-knee occurs where one net ends and another begins, e.g. nets C and E in Fig. 8.22c. During zone construction, at columns where knock-knees appear, the net that is beginning its horizontal span is only added to the *subsequent* column set, rather than the current column set under consideration. For example, in Fig. 8.22c knock-knee signals E, F, G, I, and J are removed from the marked columns and only appear in the subsequent columns.

The effect of this column change on the resulting zones is demonstrated in Fig. 8.22b, where there are six (6) zones rather than the five (5) from the previous zone analysis Fig. 8.19. Despite containing an additional zone, the largest column set now contains *one fewer* net than the original, resulting in the 4-track solution depicted in Fig. 8.22c.

Overall, the effect incorporating knock-knees into a routing solution is that two knock-knee nets can now occupy separate zones, and therefore can be placed on the same track. Additional zones may be created; however, those zones are equal in size or *smaller* in terms of nets—*potentially reducing the lower bound on the number of tracks required for routing*.

8.7.2.3 Cycles Induced by Crossing Constraints

Crossing constraints can induce cycles in the VCG. Consider the three nets depicted in Fig. 8.23a. Without crossing constraints, nets A and B would be unconstrained, and no cycle would occur; however, due to the constraint edge between B and A such a cycle occurs. Cyclic constraints cannot be routed without additional tracks and require "doglegging" to complete routing [58]. In order to avoid crossings, the routes for A and B are converted into doglegging routes as depicted in Fig. 8.23b, utilizing the same columns as the original. Unfortunately, this results in an additional

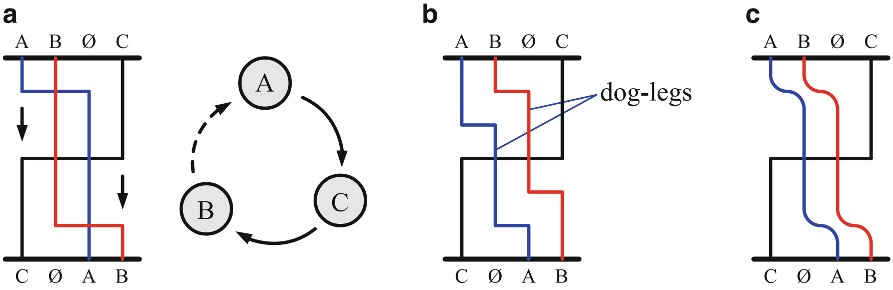


Fig. 8.23 Cycles induced by crossing constraints. (**a**) Vertical cyclic constraints; (**b**) Dog-legging avoids crossings; (**c**) Knock-knees avoid additional tracks

two (2) tracks being added to the routing solution should spare tracks not be available adjacent to the cycle. However, in the presence of knock-knees, both the crossings, and the additional tracks can be avoided, as depicted in Fig. 8.23c. The experimental results show that knock-knees can have a marked difference in track utilization especially in the presence of cyclic constraints induced by crossings.

In our work, we have designed two detail routers based on the above channel routing and crossing-aware signal loss models. Our routers provide an effective means to automate optical waveguide routing and track assignment, with signal loss as the main metric. Interested readers can refer to our publication [60] for details on our algorithms and experimental results.

8.8 Conclusion

This chapter has described design automation for integrated optics. We have demonstrated photonics design automation through a building-block methodology with optics technology-specific constraints and objectives. Our design flow is broken into behavioral and physical synthesis stages. In the behavioral synthesis phase, we describe multi-level logic design and synthesis techniques for optical digital logic. Mach-Zehnder Interferometer (MZI) and ring resonators devices are employed as switching devices—connected with waveguides—to construct optical logic systems. A virtual gate based design methodology is introduced that enables device-minimal logic synthesis for such a technology.

Post logic synthesis, the logic network needs to be placed and the interconnection-network needs to be routed. For this purpose, we introduce a row/column based placement methodology that exploits the regularity of the MZI-based optical logic network. Post placement, a global and detail routing framework is presented that minimizes signal loss as the primary optimization constraint. Signal loss models are incorporated to account for insertion losses, waveguide crossing and bends incurred due to routing on a planar substrate. This work essentially demonstrates the feasibility of *silicon photonic design automation*, though significant research is needed to make silicon-photonics design technology widely applicable and scalable.

References

1. Soref R. The past, present, and future of silicon photonics. IEEE J Sel Top Quantum Electron. 2006;12:1678–87
2. Condrat C, Kalla P, Blair S. Logic Synthesis for Integrated Optics. In: Proceedings of the 21st Edition of the Great Lakes Symposium on Great Lakes Symposium on VLSI, GLSVLSI ’11, New York:ACM; 2011. pp. 13–18.
3. Condrat C, Kalla P, Blair S. Exploring Design and Synthesis for Optical Digital Logic. International Workshop on Logic Synthesis, 2010.

4. Caulfield HJ, Vikram CS, Zavalin A. Optical logic redux. Optik. 2006;117:199–209
5. Politi A, Matthews J, O'Brien J. "Shor's Quantum Factoring Algorithm on a Photonic Chip. Science. 2009;325:1221
6. Hardy J, Shamir J. Optics Inspired Logic Architecture. Opt Express. 2007;15:150–65
7. Caulfield et al. HJ. Generalized optical logic elements GOLEs. Opt Commun. 2007;271: 365–76
8. Ganapati P. Germanium laser breakthrough brings optical computing closer. Wired Mag. 2010
9. Blair S, Wagner K. Collision-based computing. Chapter gated logic with optical solitons. London: Springer, 2002. p. 355–80.
10. Shan A. Heterogeneous Processing: a Strategy for Augmenting Moore's Law (http://www.linuxjournal.com/article/8368). Linux J. 2006; 142
11. Dokania Rk, Apsel AB. Analysis of challenges for on-chip optical interconnects. In: GLSVLSI, GLSVLSI. New York: ACM; 2009. pp. 275–80.
12. Batten C, Joshi A, Stojanovic V, Asanovic K, Designing chip-level nanophotonic interconnection networks. IEEE J Emerging Sel Top Circuits Syst. 2012;2:137–53
13. Cianchetti M, Kerekes J, Albonesi D, Phastlane: A rapid transit optical routing network. In: Proceedings of the 36th annual International Symposium on Computer Architecture, ISCA '09, New York:ACM; 2009. p. 441–450
14. Beausoleil et al. R. A nanophotonic interconnect for high-performance many-core computation. Symposium on High-Performance Interconnects, 2008. p. 182–189
15. Chan J, Hendry G, Bergman K, Carloni L. Physical-layer modeling and system-level design of chip-scale photonic interconnection networks. IEEE Trans Comput-Aided Design Integr Circuits Syst. 2011;30(10):1507–1520.
16. Emelett SJ, Soref R. Design and simulation of silicon microring optical routing switches. J Lightwave Technol. 2005;23:1800
17. Pearson et al. MR. Arrayed waveguide grating demultiplexers in silicon-on-insulator. In: Proceedings of SPIE, vol. 3953, 2000. p. 11–18
18. Boyd RW. Nonlinear optics, third edition. 3rd ed. Academic Press: New York; 2008.
19. Dinu M, Quochi F, Garcia H. Third-order nonlinearities in silicon at telecom wavelengths. Appl Phys Lett. 2003;82:2954–56
20. Liao L et al. High speed silicon Mach-Zehnder modulator. Opt Express. 2005;13:3129–35
21. Green W et al. Ultra-compact, low RF power, 10 Gb/s silicon Mach-Zehnder modulator. Opt Express. 2007;15:17106–113
22. Liao L, Liu A, Basak J, Nguyen H, Paniccia M, Rubin D, Chetrit Y, Cohen R, Izhaky N. Gbit/s silicon optical modulator for highspeed applications. Electron Lett. 2007;43(22):1196–1197
23. Park H, Fang A, Kodama S, Bowers J. Hybrid silicon evanescent laser fabricated with a silicon waveguide and III-V offset quantum wells. Opt Express. 2005;13:9460–9464
24. Lipson M. Compact electro-optic modulators on a silicon chip. IEEE J Sel Top Quantum Electron. 2006;12:1520–1526
25. Gunn C, Masini GLI. Closing in on photonics large-scale integration. Photon Spectra. 2007
26. Miller DAB. Optical interconnects to electronic chips. Appl Opt. 2010;49:F59–F70
27. Madsen C, Zhao J. Optical filter design and analysis: a signal processing approach. NewYork: Wiley, 1999.
28. Condrat C, Kalla P, Blair S. A methodology for physical design automation for integrated optics. In: Proceedings of IEEE International Midwest Symposium on Circuits and Systems, 2012.
29. Condrat C, Kalla P, Blair S. Channel routing for integrated optics. In: Proceedings of ACM/IEEE System-Level Interconnect Prediction Workshop, 2013.
30. Condrat C, Kalla P, Blair S, Crossing-Aware Channel Routing for Photonic Waveguides. In: Proceedings of IEEE International Midwest Symposium on Circuits and Systems, 2013.
31. Condrat C, Kalla P, Blair S. Thermal-aware Synthesis of Integrated Photonic Ring Resonators. In: To appear in Proceeding of the International Conference on CAD (ICCAD), Nov. 2014.
32. Condrat C. Design Automation for Integrated Optics, PhD thesis, University of Utah, 2014.
33. Pollock C, Lipson M, Integrated photonics. Dordrecht:Kluwer Academic Publishers; 2003.

34. Koester SJ et al. Ge-on-SOI-detector/si-cmos-amplifier receivers for high-performance optical-communication applications. J Lightwave Technol. 2007;25:46–57
35. OpSIS: Optoelectronic System Integration in Silicon. http://www.opsisfoundry.org.
36. Okamoto K. Fundamentals of optical waveguides. London: Academic Press; 2000.
37. Emelett S, Soref R. Analysis of dual-microring-resonator cross-connect switches and modulators. Opt Express. 2005;13:7840–53
38. Shlager KL, Schneider JB. A Selective survey of the finite-difference time-domain literature. Advances in Computational electrodynamics: the finite-difference time-domain method. Boston:Artech House Inc; vol. 37, 1995. p. 39–56.
39. Bryant RE. Graph based algorithms for boolean function manipulation. IEEE Trans Comput. 1986;C-35:677–91
40. Ding D, Pan D. Oil: a nano-photonics optical interconnect library for a new photonic network architecture. In: System-level interconnect prediction workshop (SLIP), 2009.
41. Ding D, Zhang Y, Huang H, Chen RT, Pan DZ. O-Router: an optical routing framework for low power on-chip silicon nano-photonic integration. In: Design Automation Conference 2009. p. 264–69.
42. Orcutt J, Ram R. Photonic device layout within the foundry cmos design environment. IEEE Photonics Technol Lett. 2010.
43. Ding D, Yu B, Pan D. "GLOW: a global router for low-power thermal-reliable interconnect synthesis using photonic wavelength multiplexing. In: 2012 17th Asia and South Pacific Design Automation Conference (ASP-DAC), 30 Jan–02 Feb 2012, p. 621–26.
44. Zheng Y, Lisherness P, Gao M, Bovington J, Cheng K, Wang H, Yang S. Power-Efficient Calibration and Reconfiguration for Optical Network-on-Chip. J Opt Commun Networking. 2012;4:955–66
45. Bogaerts W, Dumon P, Thourhout DV, Baets R. Low-loss, low-cross-talk crossings for silicon-on-insulator nanophotonic waveguides. Opt Lett. 2007;32:2801–03
46. Sanchis P, Villalba P, Cuesta F, Håkansson A, Griol A, Galán JV, Brimont A, J. Martí J. Highly efficient crossing structure for silicon-on-insulator waveguides. Opt. Lett. 2009;34:2760–62.
47. Xu F, Poon AW, Silicon cross-connect filters using microring resonator coupled multimode-interference-based waveguide crossings. Opt. Express. 2008;16:8649–57.
48. Cardenas J, Poitras CB, Robinson JT, Preston K, Chen L, Lipson M. Low loss etchless silicon photonic waveguides. Opt. Express. 2009;17:4752–57.
49. Vlasov Y, McNab S. Losses in single-mode silicon-on-insulator strip waveguides and bends. Opt. Express. 2004;12:1622–31.
50. Qian Y, Kim S, Song J, Nordin GP, Jiang J. Compact and low loss silicon-on-insulator rib waveguide 90° bend. Opt. Express. 2006;14:6020–28.
51. Li G, Yao J, Thacker H, Mekis A, Zheng X, Shubin I, Luo Y, Lee J, Raj K, Cunningham JE, Krishnamoorthy AV. Ultralow-loss, high-density SOI optical waveguide routing for macrochip interconnects. Opt. Express. 2012;20:12035–39.
52. Roy J, Papa D, Adya S, Chan H, Ng A, Lu J, Markov I. Capo: robust and scalable open-source min-cut floorplacer. In: Proceedings of the 2005 international symposium on Physical design, ISPD '05. New York: ACM; 2005. p. 224–6.
53. Larry M, Carl E. PathFinder: A Negotiation-based Performance-driven Router for FPGAs. In: Proceedings of the 1995 ACM Third International Symposium on Field-programmable Gate Arrays, FPGA '95. New York: ACM; 1995. p. 111–7.

54. Pan M, Chu C. FastRoute 2.0: A High-quality and Efficient Global Router. In: Design Automation Conference, 2007. ASP-DAC '07. Asia and South Pacific, 2007. p. 250–5.
55. Cho M, Lu K, Yuan K, Pan DZ. BoxRouter 2.0: Architecture and Implementation of a Hybrid and Robust Global Router, Ť ICCAD. In: In Proceeding of ICCAD 2007, 2007. pp. 503–8.
56. Chang Y, Lee Y, Wang T. NTHU-Route 2.0: A fast and stable global router. In: IEEE/ACM International Conference on Computer-Aided Design, 2008. ICCAD 2008. 2008. p. 338–43.
57. Hashimoto A, Stevens J. Wire routing by optimizing channel assignment within large apertures. In: Proceedings of the 8th Design Automation Workshop, DAC '71. New York: ACM; 1971. p. 155–69.
58. Deutsch D. A dogleg channel router. In: Proceedings of the 13th Design Automation Conference, DAC '76. New York:ACM; 1976. p. 425–33.
59. Yoshimura T, Kuh ES. Efficient algorithms for channel routing. IEEE Trans Comput Aided Des Integr Circuits Syst 1982;1:25–35.
60. Condrat C, Kalla P, Blair S. Crossing-aware channel routing for integrated optics. IEEE Trans CAD, special section on optical interconnects 2014;33,6:814–25.