

# Chapter 10

## High-Speed 3D Memories Enabling the AI Future



Zvi Or-Bach

### 10.1 Stacked Capacitor DRAM

For the last two decades, Stacked Capacitor DRAM has been the technology of choice for high speed ( $<100$  ns), high endurance ( $>10^{12}$ ), and low cost ( $< \$0.5/\text{Gb}$ ) memory. Thus far, no alternative technology has been positioned to challenge DRAM. Figure 10.1 was presented by John Hennessy in multiple events during 2018 stating: “For many years we were achieving increases of about 50 percent a year that is going up slightly faster than Moore’s law. Then we began a period of slowdown and if you look at what’s happened in the last seven years, this technology we were used to seeing increased the number of megabits per chip more than doubling every two years but is now going up at about 10% a year and it’s going to take about seven years to double now.” Capacitor based DRAM technology needs a minimum size capacitor to keep enough charge so that the refresh rate would be kept, while scaling with reduced size make it harder to keep the charge leakage under control. It is now clear that capacitor-based DRAM scaling has leveled off.

During the last decade, it was observed that the need for DRAM in computing systems is limited, while the need for storage has kept growing. Accordingly, industry analysts were expecting the NAND market will become far larger than the DRAM market by now. But the re-birth of AI technology has reversed that trend in the recent years and DRAM demand has seen a rapid growth resulting in dramatic price increases for DRAM devices (Fig. 10.2).

The diminishing effectiveness of conventional scaling, at a time of accelerating AI-driven use, presents a tough challenge for the industry. However, at nearly the same time, the NAND industry was facing a scaling challenge. But then the industry was able to change course and adopt 3D scaling (Fig. 10.3).

---

Z. Or-Bach (✉)

MonolithIC 3D Inc., 3555 Woodford Drive, San Jose, CA 95124, USA  
e-mail: [Zvi@MonolithIC3D.com](mailto:Zvi@MonolithIC3D.com)

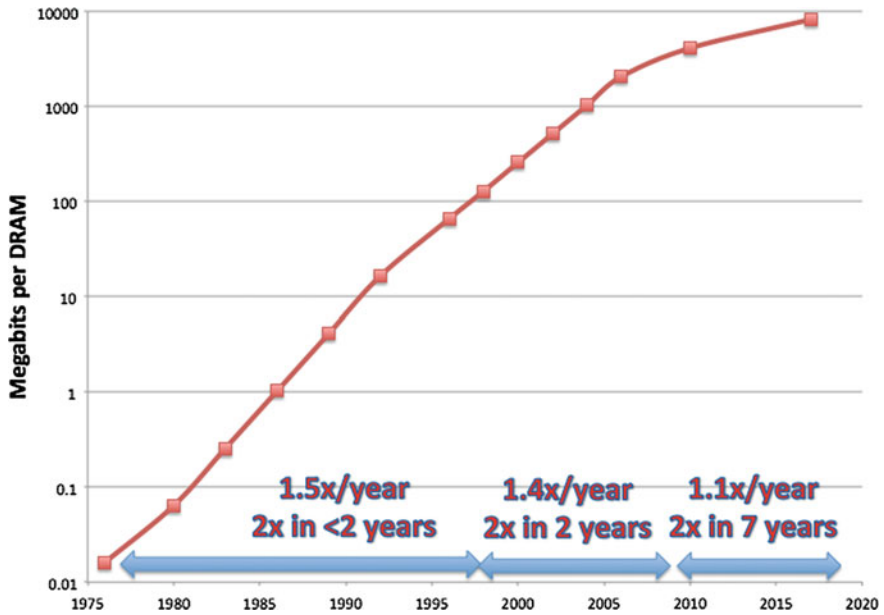


Fig. 10.1 Moore’s law for DRAM—J. Hennessy 2018

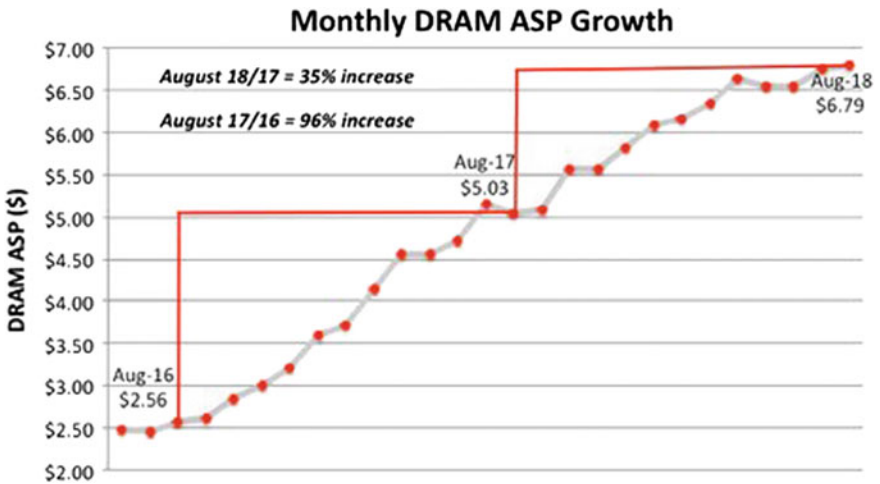


Fig. 10.2 Recent years DRAM device price appreciation

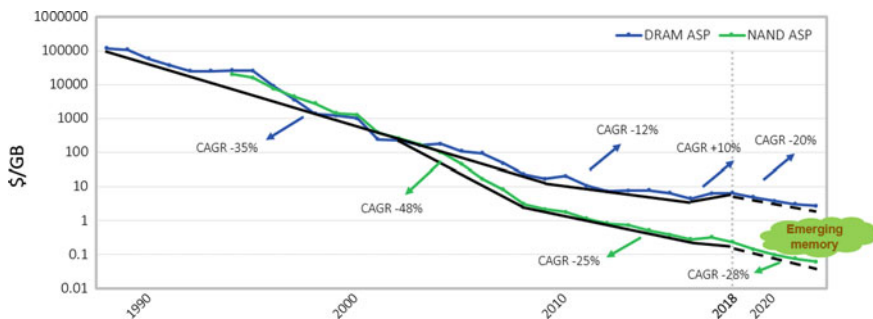


Fig. 10.3 DRAM versus flash ASP (average selling price). Source IDC

Early 3D NAND products used 24 layers in the 3D stack, and then the industry released 32, 64, 72, and recently 96 layers to production. This 3D-stack driven roadmap suggests continuing the 3D scaling towards a few hundred layers, thus keeping the scaling of NAND memory products to increase the memory capacity with the corresponding reduction in cost per bit.

Capacitor based DRAM would not allow such 3D scaling and no alternative has been proposed so far to do so for DRAM.

## 10.2 Alternative Memory Technologies

Over the past decades, a significant R&D effort has been devoted to developing alternative memory technologies. The leading alternative technologies are based on: Phase Change Materials (PCM), Resistive Memory (R-RAM) or Magnetic Memory (M-RAM). These alternative memory technologies have many variations and derivatives with other name branding as well. None of these alternative memory technologies seems to challenge the mainstream technologies—DRAM and NAND. And none of these technologies has been considered as a potential alternative to DRAM.

### 10.2.1 PCM—3D XPoint

Intel and Micron collaborated in releasing to the market a product named Optane™ as a Storage Class Memory (SCM) to bridge the growing gap between DRAM and 3D NAND. Also, it is considered a 3D memory as it is designed as a cross-point architecture and would not fit the low-cost 3D Scaling in which many memory layers are processed together following the same lithography step. 3D XPoint is not considered as a potential DRAM alternative due to access speed and endurance limitations.

### ***10.2.2 R-RAM***

R-RAM has been a very popular memory candidate with many alternative material and configurations. So far it seems that most of the effort is to position R-RAM as an attractive alternative for embedded non-volatile memory. R-RAM is not been proposed as a DRAM alternative mostly due to endurance limitations.

### ***10.2.3 M-RAM***

M-RAM has recently made good progress and is now being offered as a qualified non-volatile embedded memory by multiple vendors including TSMC, Samsung, and Intel. M-RAM has not been proposed as a DRAM alternative mostly due to the much larger memory cell size and concerns with the challenge of scaling to smaller technology nodes.

### ***10.2.4 F-RAM***

Ferro-Electric memory (F-RAM) is an established high-speed non-volatile specialty memory technology currently offered by Fujitsu and Cypress. It was considered a low-density memory due to the prohibitive thickness of the special Ferro-Electric materials. Recently, it was discovered that doped hafnium oxide ( $\text{HfO}_2$ ) exhibits ferro-electric properties and could enable a high-density F-RAM [1, 2]. The technology was proposed to support capacitor-based DRAM or single transistor memory cells. So far, the endurance of single transistor FRAM memory cell has been at about  $10^6$  cycle, which is too low to be considered as a DRAM alternative.

## **10.3 Charge-Trap DRAM**

Charge-Trap is the dominating technology for 3D NAND which is considered a slow Non-Volatile memory technology. In CT 3D NAND, the charge is trapped in a nitride layer of about 5–8 nm thick. A high-quality tunneling oxide is placed as a barrier between the trapping layer and the channel to keep the charge trapped for about 10 years. In a landmark paper by Wann and Hu in 1995 [3] it was presented that thinning the tunneling oxide to about 1 nm provides a memory with performance attractive for DRAM applications from all aspects (endurance, access time, size). So, by accepting the concept of refreshing the CT NAND memory cells, a Charge-Trap memory of thin tunneling oxide will give up a 10-year retention for few seconds of retention, but in return possess fast write times in the tens of ns and an endurance

higher than  $10^{12}$  cycles. A similar concept was published by IBM [4]. This work was confirmed and improved on by work such as Fujitsu's [5] and covered in patents filed by Macronix [6] and Micron [7]. It seems that the thin tunneling concept was proposed at a time floating gate rather than charge-trap was the Non-Volatile industry's technology of choice. Moreover, at that time DRAM scaling was in-step with the rest of the industry and a thin tunneling charge-trap did not offer enough of an advantage to be pursued by the memory industry.

In the Flash market for storage applications, NAND architecture became the industry choice as it provides a significantly higher density (lower cost) than a NOR architecture. As illustrated in Fig. 10.4 a NAND architecture with only two diffusion contacts could provide access to a long NAND string, thus reducing the effective size of a memory cell to  $4F^2$  [8]. In the NOR architecture, the one diffusion contact per cell increases the cell size to  $8F^2$ , thus a higher memory cost. The NOR architecture does provide direct access to the selected cell which result in much faster read access time, consequently making it attractive for applications such as program code storage. An alternative architecture shown in Fig. 10.4 as AND architecture provides direct access with a better density than conventional NOR. This architecture often is also called NOR and could be attractive for 3D random access memory structures.

The success of the NAND industry with 3D NAND scaling could now be followed by adopting Charge-Trap for DRAM and changing the memory architecture from a NAND to a NOR (AND) architecture. Such a 3D architecture has been first proposed by Macronix [9], later by MonolithIC 3D Inc. with single crystal channel option [10], and then by Eli Harari [11] and his new company Sunrise Memory Corp (Eli Harari was the founder of SanDisk and won the National Medal of Technology and Innovation from President Barack Obama for his innovations and contributions to flash memory storage solutions). These proposals could be grouped into those with horizontal bit-line orientation and vertical bit-line orientation. In the following, the details of 3D NOR with a vertical bit-line orientation are presented. An important advantage of these structures is the similarity to the common 3D NAND 'Punch and Plug' process and accordingly the advantage in sharing the industry accumulated know-how and manufacturing infrastructure.

## 10.4 Charge-Trap 3D NOR (AND)

Just like in 3D NAND, the foundation fabric is a multi-layer fabric such as oxide layers with poly-silicon in-between. The number of poly-silicon layers is a linear relation to the number of memory cells in the 3D memory structure. And just as in 3D NAND the memory process is done for the full multi-layer fabric affecting all the levels together—hence 3D scaling (Fig. 10.5).

Figure 10.6 illustrates a side cut-view of the structure overlaying the structure transistor schematic. It represents an aggressive 3D NOR (AND) structure in which the bit-lines (B0–B4), in blue, serves as Source and Drain to cells on their right side and on their left side. These bit lines could be formed by filling the punch holes with

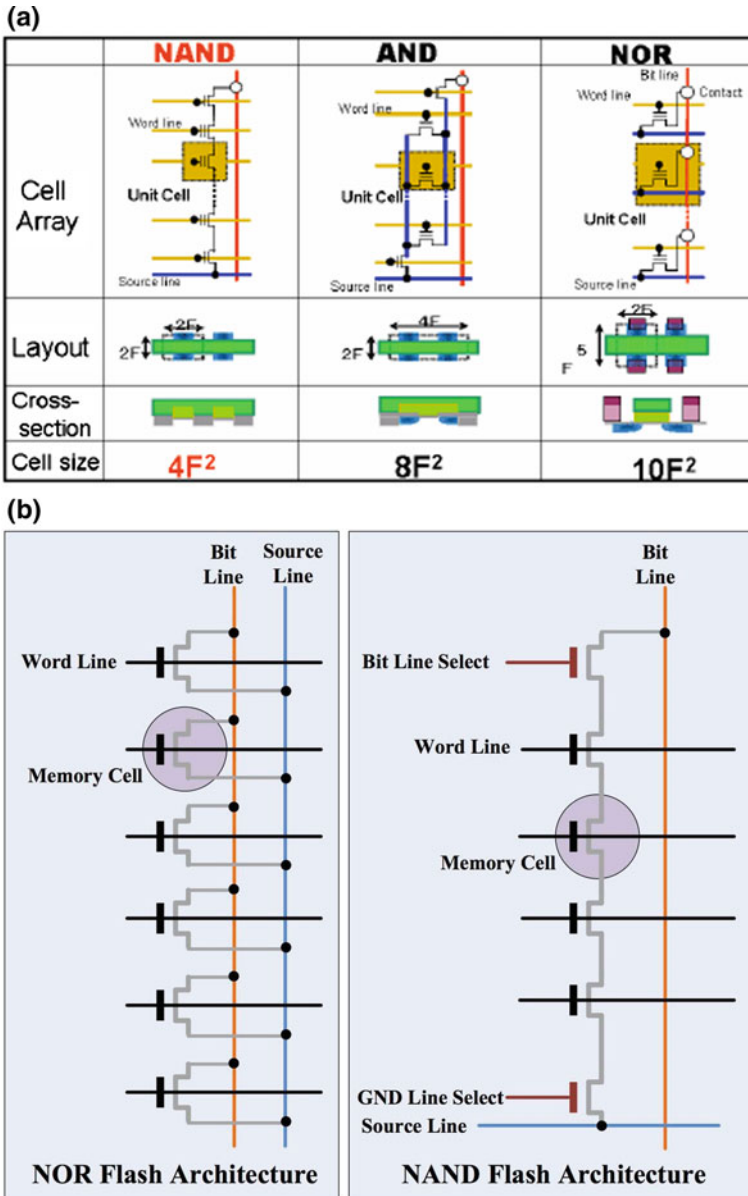
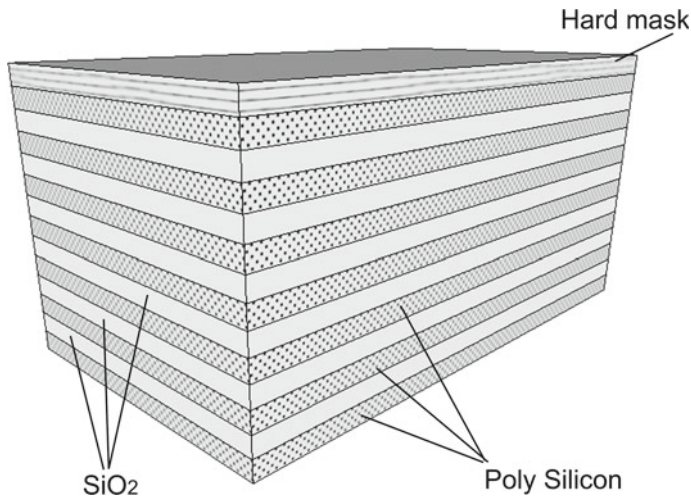
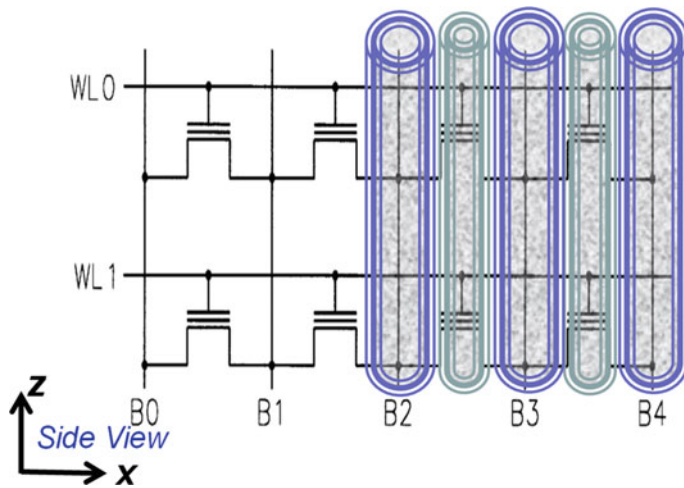


Fig. 10.4 a NOR versus NAND flash architecture. b NOR versus NAND flash architecture



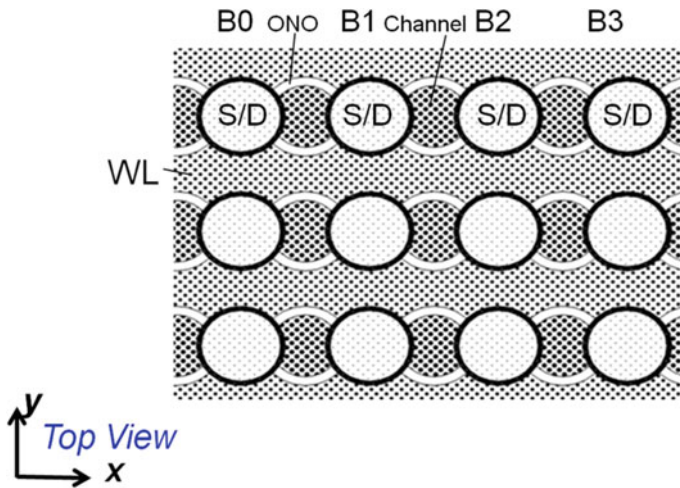
**Fig. 10.5** Multilayer fabric as foundation for 3D NAND and 3D NOR



**Fig. 10.6** Transistor schematic overlaid by punch and fill source/drain (bit-lines) in the odd holes and channel in the even holes

N+ silicon, or through a combination of N+ layers on the holes' walls and core of metal or even just metal for a Schottky-based structure. The channel holes in between are filled with un-doped polysilicon.

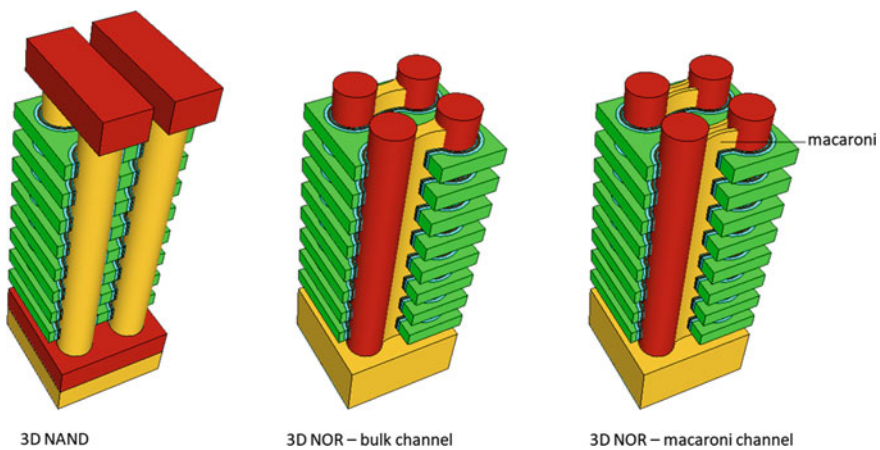
The structure looks like a 3D NAND with N+ holes punched between channel holes. Figure 10.7 illustrates a top view of the structure.



**Fig. 10.7** Top-view, source/drain (bit-lines) in the odd holes and channel in the even holes

Figure 10.8 illustrates an alternative for the 3D NOR structure in which no additional holes are ‘punched’ for the channel but rather forming the channels by use of etch and deposition through the S/D holes.

Additional details for the 3D NOR structures and alternative process flows to form them could be found in the referenced patents and applications [9–12].



**Fig. 10.8** Some alternatives for 3D memory structures



### 10.5 Schottky Barrier and Dopant Segregated Schottky Barrier (“DSSB”)

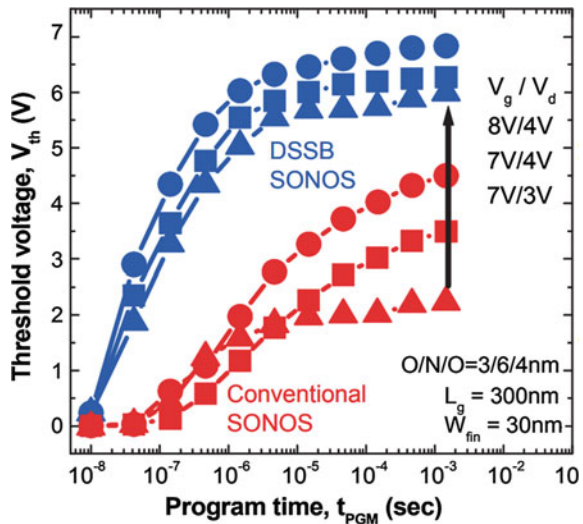
In flash devices there are a few writing mechanisms that are frequently used. One is Fowler–Nordheim (FN) tunneling commonly used in NAND flash devices and another is Hot Carrier Injection (HCI), also called Hot Electron, often used in NOR flash devices. Flash cell writing using FN tunneling is orders of magnitudes more efficient than HCI as in FN most of the current is the tunneling current while in HCI only a small fraction of the current through the channel is actually the hot carriers being driven over the quantum barrier thus to be trapped.

In a paper [13] titled “Performance breakthrough in NOR flash memory with dopant-segregated Schottky-barrier (DSSB) SONOS devices” a few orders of magnitude improvements were reported by the use of Schottky Barrier devices, as is illustrated in Fig. 10.9 (Fig. 3 of the paper [13]).

This improvement in hot carrier write time and efficiency was reported in other papers including devices without dopant segregation, and devices utilizing poly silicon channels [14–16]. Using metalized Source/Drain lines in the 3D NOR device improves the bit-line conductivity and thus enhances the device P/E efficiency and speed.

Comparing such a 3D NOR technology to Stacked Capacitor DRAM suggests many advantages such as: higher density, 3D scaling, lower power, reduced rate of refresh, non-destructive read. Yet Charge-Trap 3D NOR is expected to have much longer erase time. Proper design of a 3D NOR device could support a full segment erase scheme, which combined with proper system design and support software, could compensate for the erase time deficiency.

**Fig. 10.9** Program/erase characteristics for NOR flash memory cell (double gate), DSSB and conventional SONOS devices



## 10.6 Periphery Under Cell (“PUC”) or Over Cell (“POC”)

To further enhance the 3D NOR structure to support DRAM applications, it is useful to have the memory control circuits, often called periphery circuits, either under the memory array or on top of it. Some of the 3D NAND products in the market use periphery under cell, also called CMOS under Array (‘CUA’), currently being produced by Micron and Intel. And as discussed in Chap. 8, YMTC uses Xtacking to form the periphery over the memory array. For DRAM applications the 3D NOR structure could utilize these ideas further to break the array to hundreds or even thousands of small arrays, each with its own control circuits, to keep the memory control lines short and accordingly support very high-speed access.

## 10.7 Further Applications

3D NOR high speed memory could be an attractive architecture to many memory applications such as Storage Class Memory (SCM) and AI applications such as Neuromorphic Computing [17, 18]. The NOR (AND) architecture provides direct access to the selected cell, the 3D structure allows high density packing and reduces costs with 3D scaling. Supporting it, with periphery under the cell or over the cell, further enables high speed access and partitioning the memory into small arrays helps keep the memory access lines short. In summary, 3D integration technology is already a part of memory scaling and it is positioned to support the full range of memory applications required to keep advancing device integration to drive the AI era.

## References

1. S. Mueller et al., Incipient ferroelectricity in Al-doped HfO<sub>2</sub> thin films. *Adv. Funct. Mater.* **22**(11), 2412–2417 (2012)
2. J. Müller et al., Ferroelectric hafnium oxide: a CMOS-compatible and highly scalable approach to future ferroelectric memories, in *2013 IEEE International Electron Devices Meeting* (IEEE, 2013)
3. H.C. Wann, C. Hu, High-endurance ultra-thin tunnel oxide in MONOS device structure for dynamic memory application. *IEEE Electron Device Lett.* **16**(11), 491–493 (1995)
4. H.I. Hanafi et al., A scalable low power vertical memory, in *Proceedings of International Electron Devices Meeting* (IEEE, 1995)
5. K. Tsunoda et al., Ultra-high speed direct tunneling memory (DTM) for embedded RAM applications, in *Digest of Technical Papers. 2004 Symposium on VLSI Technology, 2004* (IEEE, 2004)
6. Patents: US 7,848,148, US 8,705,278
7. Patents: US 6,249,460, US 6,639,835, US 6,730,960
8. J. Cooke, Flash Memory 101: An Introduction to NAND Flash. *EE Times*, 20 Mar 2006
9. Patents: US 8,203,187, US 8,426,294

10. Patents: US 10,014,318 and PCT application WO 2017/053329
11. Patents: US 9,842,651, US 9,892,800, US 9,911,497
12. Patent application: WO/2018/144957
13. S.-J. Choi et al., Performance breakthrough in NOR flash memory with dopant-segregated Schottky-barrier (DSSB) SONOS devices, in *2009 Symposium on VLSI Technology* (IEEE, 2009)
14. S.-J. Choi et al., A novel TFT with a laterally engineered bandgap for of 3D logic and flash memory, in *2010 Symposium on VLSI Technology* (IEEE, 2010)
15. C.-H. Shih et al., Source-side injection Schottky barrier flash memory cells. *Semicond. Sci. Technol.* **24**(2), 025013 (2009)
16. C.-H. Shih et al., Schottky barrier silicon nanowire SONOS memory with ultralow programming and erasing voltages. *IEEE Electron Device Lett.* **32**(11), 1477–1479 (2011)
17. Y. Noh et al., Synaptic devices based on 3-D AND flash memory architecture for neuromorphic computing, in *IEEE International Memory Workshop (IMW)* (2019)
18. H.-T. Lue et al., A novel 3D AND-type NVM architecture capable of high-density, low-power in-memory sum-of-product computation for artificial intelligence application, in *2018 IEEE Symposium on VLSI Technology* (IEEE, 2018)