Introduction to 3D NAND Flash Memories

Rino Micheloni, Luca Crippa, and Cristian Zambelli

Abstract Nowadays, NAND Flash memories are in everybody's hands, as they are the storage media used inside smartphones and tablets. At the same time NAND Flash memories, in the form of Solid State Drives (SSDs), have enabled a new generation of computers without Hard Disk Drives (HDDs), and they are also one of the key components of the modern cloud infrastructures. To win all these new applications, NAND Flash had to continuously decrease its cost per bit. Shrinking lithography has been the solution for many generations of planar NANDs, but this approach ran out of steam in the sub-20nm range due to a plethora of parasitic effects within the memory array. As such, both the industry and the academia have worked towards a different approach for many years, resulting in a tri-dimensional (3D) architecture, whose first product reached the market in 2016. In this Chapter we present the basics of 3D NAND Flash memories and the related integration challenges. There are two main variants of Flash technologies used inside 3D arrays, namely, Floating Gate (FG) and Charge Trap (CT), which are both described in this Chapter with the aid of several bird's-eye views. Finally, 3D scaling trends are discussed.

Nowadays, Solid State Drives consume an enormous amount of NAND Flash memories [1] causing a restless pressure on increasing the number of stored bits per mm^2.

This chapter is an authorized partial reprint of Micheloni R., Aritome S., Crippa L. (2018) 3D NAND Flash Memories. In: Micheloni R., Marelli A., Eshghi K. (eds) Inside Solid State Drives (SSDs). Springer Series in Advanced Microelectronics, vol 37. Springer, Singapore. https://doi.org/10.1007/978-981-13-0599-3_5.

R. Micheloni (✉) · C. Zambelli
Dipartimento di Ingegneria, Università degli Studi di Ferrara, Via G. Saragat 1, 44122 Ferrara, Italy
e-mail: rino.micheloni@ieee.org

C. Zambelli
e-mail: cristian.zambelli@unife.it

L. Crippa
Manzoni 66, 20874 Busnago (MB), Italy
e-mail: luca.crippa@ieee.org

R. Micheloni and C. Zambelli (eds.), *Machine Learning and Non-volatile Memories*,
https://doi.org/10.1007/978-3-031-03841-9_5

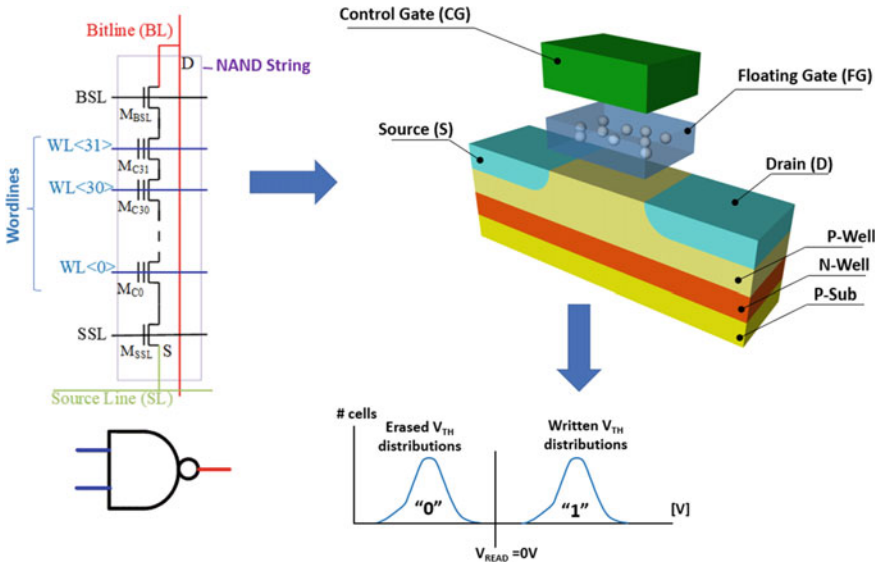


Fig. 1 NAND string

Planar memory cells (Fig. 1) have been scaled for decades by improving process technology, circuit design, programming algorithms [2], and lithography.

Unfortunately, when approaching a minimum feature size of 1x-nm, more challenges pop up: doping concentration in the channel region becomes difficult to control [3], RTN [4] and electron injection statistics [5] widen threshold distributions, thus causing a significant hit to both endurance and retention. Furthermore, by reducing the distance between memory cells, the intra-wordline electric field becomes higher, pushing the bit error rate to an even higher level.

3D arrays can definitely be considered as a breakthrough for fueling a further increase of the bit density. Identifying the right way for going 3D was not so easy though.

Historically, Flash memory manufacturers have leveraged lithography to shrink the 2-dimensional (2D) memory cell [6].

However, with 3D architectures, the "simple" reduction of the minimum feature size is running out of steam [7]: a higher number of stacked cells is the only hope for dramatically reducing the real estate of a stored bit.

3D arrays can leverage either *Floating Gate* (FG) or *Charge Trapping* (CT) technologies [8]. As a matter of fact, the vast majority of 3D architectures published to date are built with CT cells, mainly because of the simpler fabrication process. Nevertheless, Floating Gate is still around and there are commercial products who managed to integrate FG into a 3D array.

1 3D Charge Trap NAND Flash Memories

3D arrays can be efficiently built by vertically rotating the planar NAND Flash string as displayed in Fig. 2. The solution of choice is a conduction channel completely surrounded by the gate [9]: indeed, the curvature effect helps increasing the electric field E_t across the tunnel oxide, and reduces the electric field E_b across the blocking oxide [10, 11], and this has a positive impact on oxide reliability and overall power consumption.

Vertical channel arrays have been historically driven by architectures known as BiCS, which stands for *Bit Cost Scalable* [12, 13] and P-BiCS, acronym for *Pipe-Shaped BiCS* [14–16], which are both leveraging CT cells [17]. Let's get started with BiCS, which is sketched in Figs. 3 and 4 [13]. There is a stack of *Control Gates* (CGs), the lowest being the one of the *Source Line Selector* (SLS). The whole vertical stack is punched through and the resulting holes are filled with poly-silicon; each filled hole (a.k.a. pillar) forms a series of memory cells vertically connected in a NAND fashion. *Bit Line Selectors* (BLS's) and *Bitlines* (BLs) are formed at the top of the structure [18].

The poly-silicon body of memory cells is not doped or lightly doped [10, 11]; indeed, considering the bad aspect ratio of the vertical polysilicon plug, p-n junctions cannot be easily realized by either diffusion or implantation in a trench structure. As usual, a select transistor (BLS) is used to connect each NAND string to a bitline;

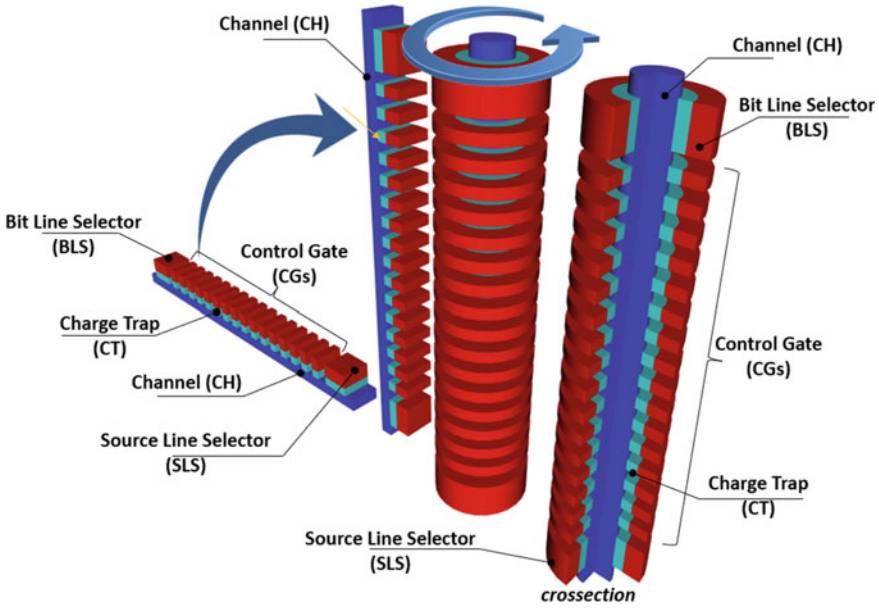


Fig. 2 The NAND flash string goes vertical

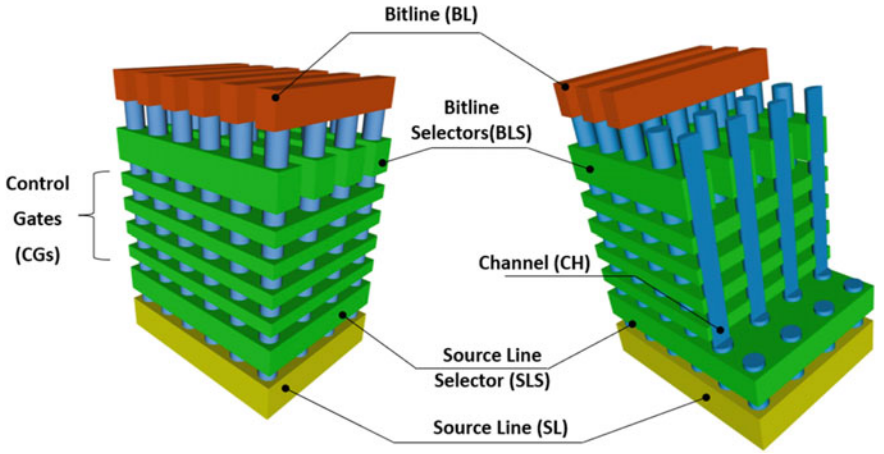
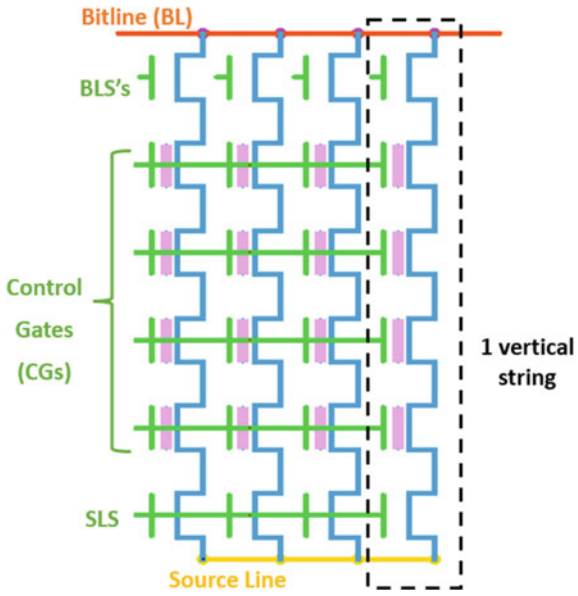


Fig. 3 BiCS architecture. Adapted with permission from [19]. ©2017 IEEE

Fig. 4 Equivalent circuit of a BiCS array



there is also another select transistor (SLS), which connects the other side of the string to the common source diffusion.

It is important to highlight that the number of critical and expensive lithography steps does not depend on the number of control gate plates because the whole 3D stack is drilled at one [20, 21].

As sketched in Fig. 5, vertical transistor have polysilicon body and this fact turned out to be one of the critical cornerstone of the 3D foundation. From a manufacturing

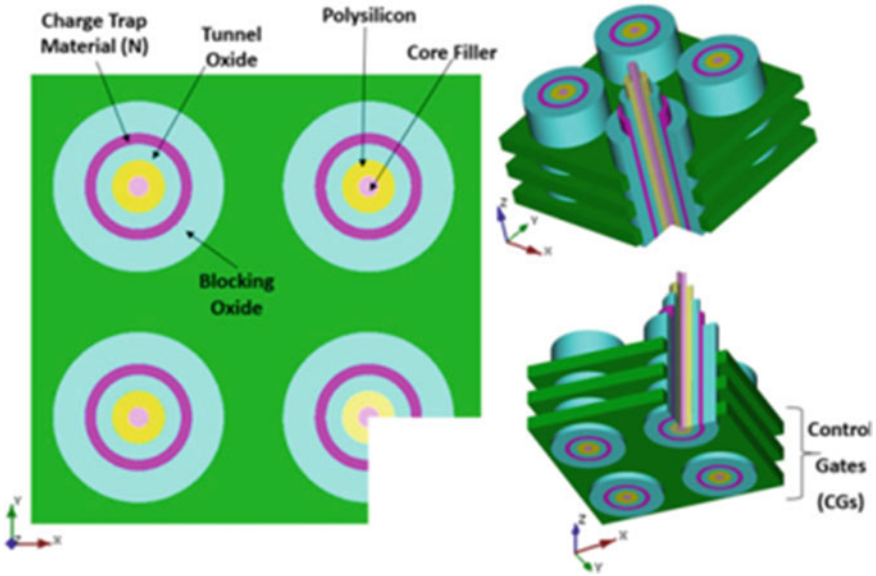


Fig. 5 BiCS memory cells

perspective, the density of the traps at the grain boundary is very difficult to control, with such a vertical shape: the bad thing is that this poor control induces significant fluctuations of the characteristics of vertical transistors.

The recipe for fixing the trap density fluctuation problem is to manufacture a polysilicon body much thinner than the depletion width. In other words, by shrinking the polysilicon volume, the total number of traps goes down (Fig. 6). This particular structure is usually referred to as *Macaroni Body* [13]. A *filler layer* (i.e. a dielectric film) is used in the central part of the macaroni structure, essentially because it makes the manufacturing process easier.

The fabrication sequence of the BiCS array [22] starts from building the layers for control gates and selectors. Then, BLS stripes are defined. After forming pillars, bitlines are laid out by using a metal layer.

Control gate edges are extended to form a ladder to connect to the fan-out region, as sketched in Fig. 7 [12, 13, 22, 23]. Actually, there are 2 ladders: one of the 2 can't be used because it is masked by the metals biasing the bitline selectors.

Over time BiCS became P-BiCS, mainly to improve the Source Line resistance [14, 15]. In a nutshell, two vertical NAND strings are shorted together at the bottom of the 3D structure: in this way, they form a single NAND string and the 2 edges are connected to the bitline and to the Source Line, respectively (Fig. 8). Thanks to its U-shape, P-BiCS has few advantages over BiCS:

- retention is better because manufacturing creates less damages in the tunnel oxide;
- being at the top, the Source Line can be connected to a metal mesh, thus lowering its parasitic resistance;

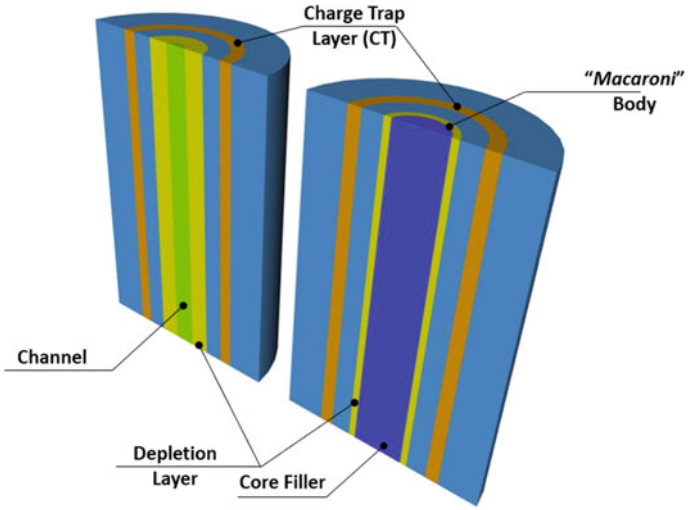


Fig. 6 A vertical transistor (right) modified with *Macaroni* body (left)

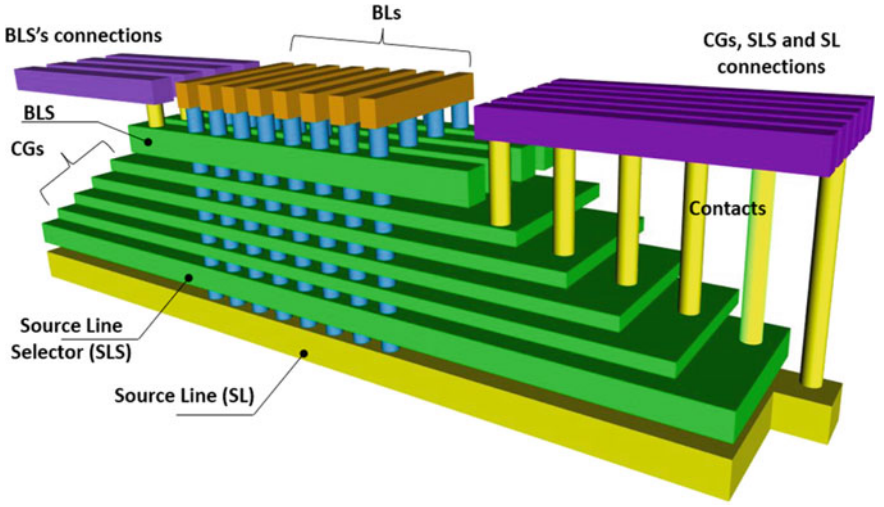


Fig. 7 Fan-out of the BiCS array. Adapted with permission from [19]. ©2017 IEEE

- Source Line and bitline selectors are at the same height of the stack and, therefore, they can be equally optimized and controlled, thus obtaining a better string functionality.

One of the biggest drawbacks of P-BiCS is the fact that at the same height of the stack there are two different control gates which, of course, can't be biased together; therefore, the two layers can't be simply shorted together. As a result, compared to

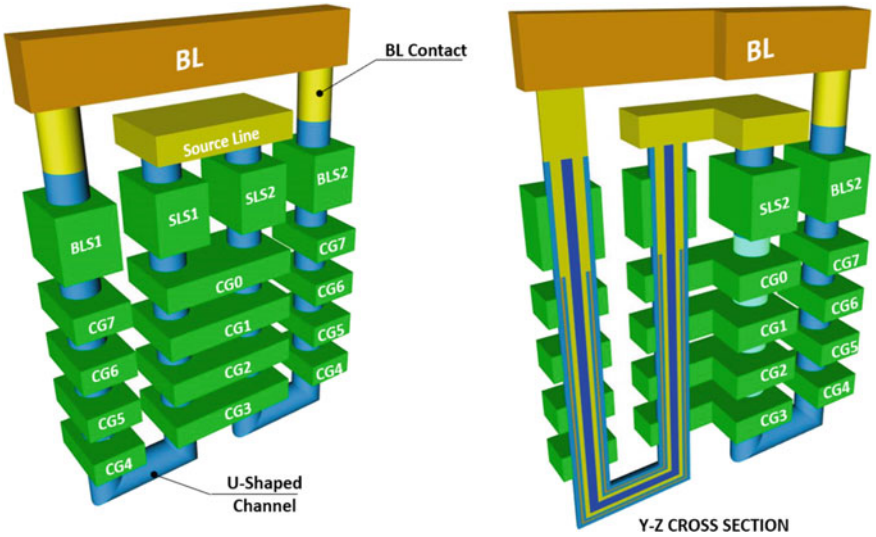


Fig. 8 P-BICS NAND strings

BiCS, a totally different and more complex fan-out is required [16], as displayed in Fig. 9: basically, a fork-shaped gate is adopted, such that each branch acts on two NAND pages.

A major advantage is the easier connection of the source line [14] through the "Top Level Source Line" of Fig. 10. This additional metal mesh guarantees a much better noise immunity for circuits.

Besides BiCS and P-BiCS, many other approaches were tried, including VRAT (*Vertical Recess Array Transistor*) [24], Z-VRAT (*Zigzag* VRAT) [24], and VSAT (*Vertical Stacked Array Transistor*) [25], and 3D-VG (*Vertical Gate*) NAND [26] which is a unique architecture where the channel runs along the horizontal direction.

TCAT (*Terabit Cell Array Transistor*) was disclosed in 2009 [27] and it was the foundation for V-NAND (Fig. 11), which is the first 3D memory device who reached the market. Except for SL + regions which are n + diffusions, the equivalent circuit of TCAT is the same of BiCS (Fig. 4). All SL + lines are connected together to form the common Source Line. There are 2 metal layers for decoding wordlines and NAND strings, respectively.

TCAT is based on *gate-replacement* [27], whereas BiCS is *gate-first*. Gate-replacement begins with the deposition of multiple oxide/nitride layers. After the stack formation, nitride is removed through an etching process. Afterwards, tungsten metal gates are deposited and, finally, gates are separated by using another etching step. Metal gates translate into a lower wordline parasitic resistance, resulting in faster programming and reading operations.

The bulk erase operation is another significant difference compared to BiCS. Because NAND strings are close to n + areas, during erasing, holes can come straight

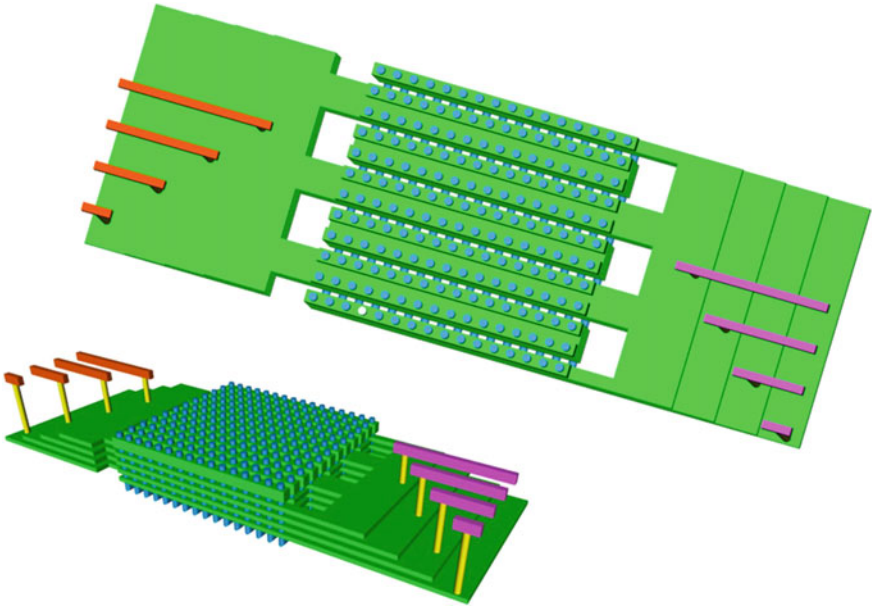


Fig. 9 Fork-shaped fan-out. Adapted with permission from [19]. ©2017 IEEE

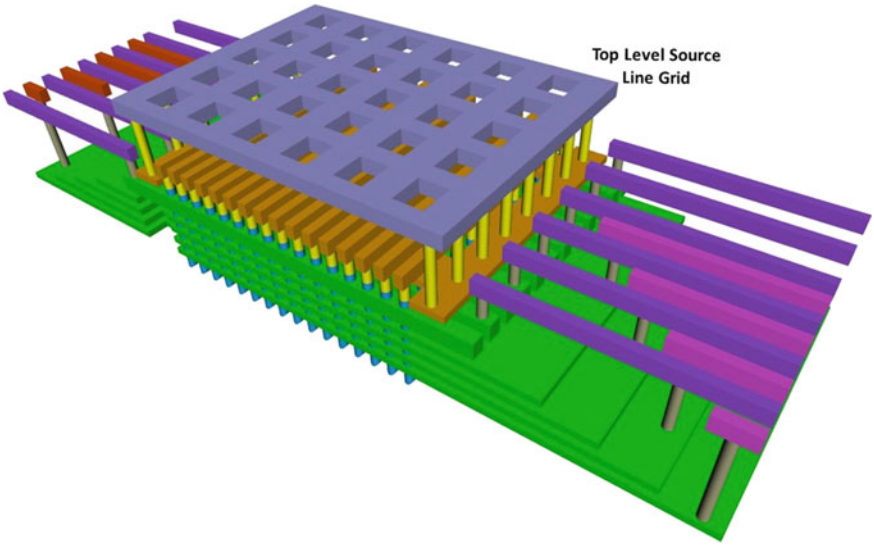


Fig. 10 P-BiCS: Source line metal mesh. Adapted with permission from [19]. ©2017 IEEE

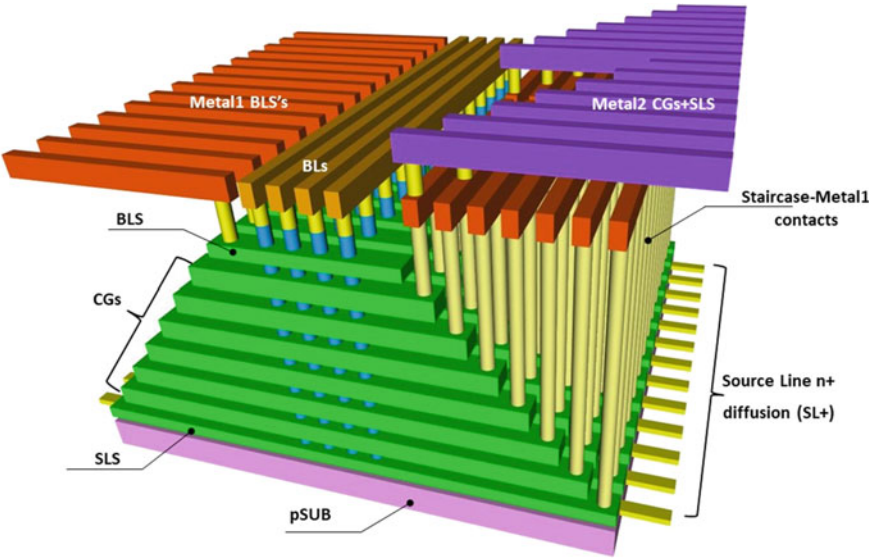


Fig. 11 TCAT NAND flash array

from the substrate, thus avoiding the GIDL (*Gate Induced Drain leakage*) on the source side, which is a well-known problem for BiCS.

BiCS and TCAT are compared in Fig. 12 [28]. Being TCAT based on a gate-last process, the charge trap layer is biconcave, and thanks to this particular shape it is much harder for charges to spread out. On the contrary, BiCS is characterized by a

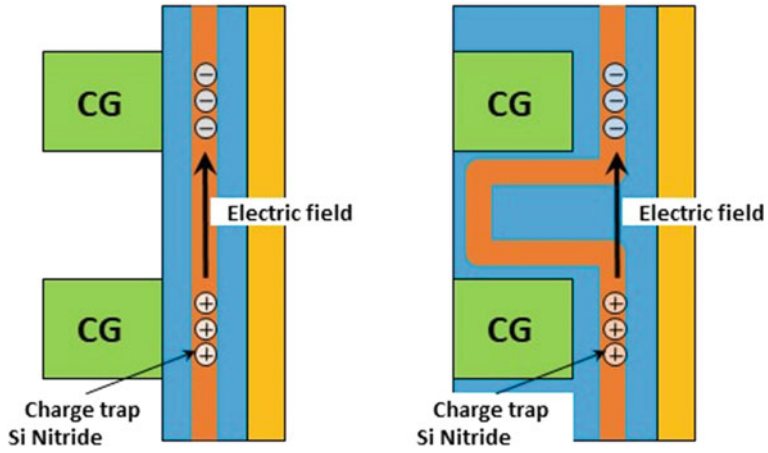


Fig. 12 BiCS versus TCAT

charge trapping layer going through all gate plates, thus acting as a charge spreading path: of course, the main consequence of this layout is a degradation of data retention.

TCAT evolved into another architecture called V-NAND [29]. The first generation had 24 wordline layers, plus additional dummy wordline layers (dummy CG) [30–32].

Why dummy layers? Mainly because of the floating body of the memory cells with vertical channel. In fact, during the programming operations, hot carriers are generated by the high lateral electric field located at the edge of the NAND string. Therefore, these hot carriers keep the voltage on the channel low during the programming operation of the first wordline (i.e. Program Disturb) [33, 34]. Dummy wordlines before the first WL are an effective and simple solution to this problem [35, 36].

A 128 Gb TLC (3 bit/cell) device manufactured by using V-NAND Gen2 was published in 2015 [37, 38]. Gen2 had 32 memory layers instead of the previous 24 and introduced the concept of Single-Sequence Programming. Conventional (mainly 2D) TLC programming techniques go through the programming sequence multiple times. To be more specific, each wordline is programmed 3 times, such that V_{TH} distributions can be progressively tightened. Because of the smaller cell-to-cell interference (compared to FG), CT cells exhibit an intrinsic narrower native V_{TH} distribution. As a result, V-NAND Gen2 could write 3 pages of logic data in a single programming sequence. There are 2 benefits to this approach: reduced power consumption and faster programming.

V-NAND Gen3 appeared in 2016 [39], in the form of a 48 layer TLC device. With such a high number of gate layers, the very high aspect ratio of the pillar becomes a serious challenge for the etching technology. To mitigate this problem, the easiest solution is to shrink the thickness of gate layers. The downside of this approach is that the parasitic RC of the wordline gets higher, thus slowing access operations to the memory array. Moreover, channel's size fluctuations become critical. Indeed, pillars are holes drilled in the gate layer and they represent a barrier for charges flowing along the wordline: in essence, a distribution of the holes diameters generates a distribution of the parasitic resistances of gate layers. In addition, pillars, once manufactured, have the conic shape sketched in Fig. 13. The overall result is that the same voltage applied to different gate layers translates into a waveform per layer. An adaptive program pulse scheme can fix the problem. In a nutshell, the program pulse duration has to be tailored to the characteristics of the wordline layer. As the number of layers increases, the pillar becomes longer with a negative impact on the aspect ratio of the pillar. To compensate for that, V-NAND Gen4 [40], which is built on a stack of 64 layers, had to shrink both the layer thickness and the intra-layer distance (spacing). The downside is an increased wordline parasitic capacitance which adversely affects cell's reliability and timings. Improved circuits and programming algorithms can be used to tackle this problem [40].

As discussed, both BiCS [41] and V-NAND use CT cells, but Floating Gate still exists, as explained in the next Section.

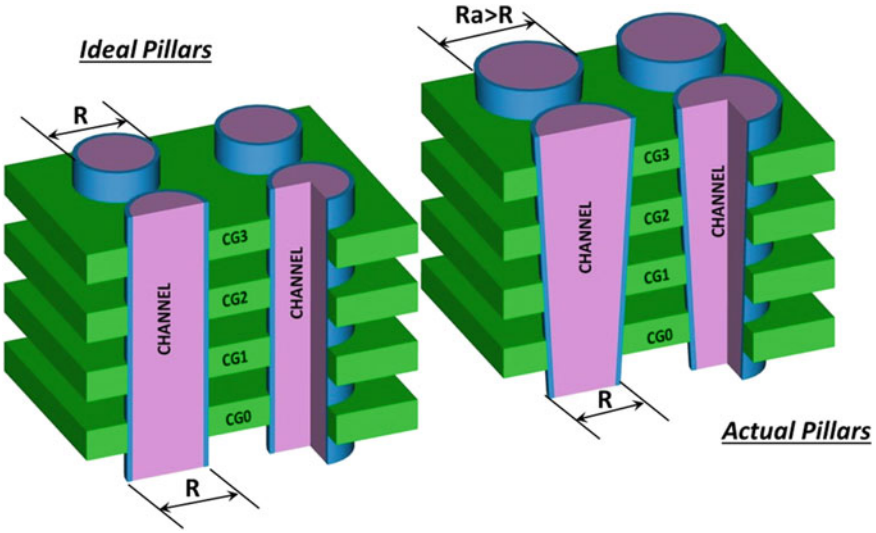


Fig. 13 Ideal versus actual shape of pillars

2 3D Floating Gate NAND Flash Memories

2D NAND Flash memories use FG cells which have been improved and optimized for decades. Of course, there have been many attempts to reuse this know-how in 3D.

The first 3D attempt is known as *3D Conventional FG* (C-FG) or S-SGT (*Stacked-Surrounding Gate Transistor*) [42–44], and it is sketched in Fig. 14.

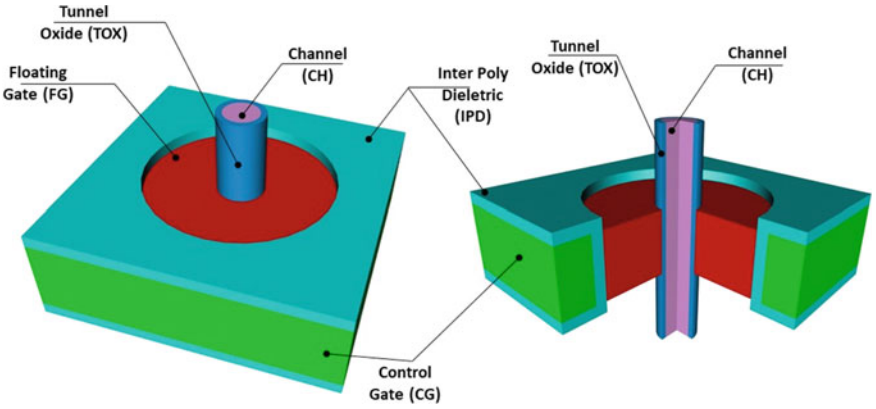


Fig. 14 3D C-FG cell. Adapted with permission from [19]. ©2017 IEEE

A C-FG NAND string is shown in Fig. 15, including select transistors. Please note that both string selectors are manufactured as standard transistors, i.e. they haven't any floating gate. Figure 16 shows a C-FG array, including the fan-out region. While all wordlines at the same height of the stack are connected, BLS lines can't, because they need to be page selective per each CG layer. On the contrary, SLS transistors can be shorted together, thus saving both power and silicon area.

Because we are talking about FG cells, FG coupling between neighboring cells is the main hurdle for vertical scaling. With enhancement-mode operations, the high resistance of source/drain (S/D) regions should also be carefully considered. In fact, these regions need high-doping and this is not very easy to accomplish when the conduction channel is made of polysilicon. The solution to this problem is to electrically invert the S/D layer by using higher voltages during read. This simple solution is hardly manageable by C-FG cells because of the thin FG.

The *Extended Sidewall Control Gate* (ESCG) structure, Fig. 17 [45], is another FG option and it was developed to contain the interference effect. Moreover, by applying a positive voltage to the ESCG structure, density of electrons on the surface of the pillar can be much higher than C-FG (even one order of magnitude): a highly inverted electrical source/drain can significantly lower the S/D resistance.

In addition, the ESCG shielding structure reduces the FG–FG coupling capacitance: the ESCG region is biased as CG, and the CG coupling capacitance (C_{CG}) is significantly increased because of the increased overlap area between CG and FG. A higher CG coupling ratio is one of the key ingredients for achieving effective NAND Flash operations [46].

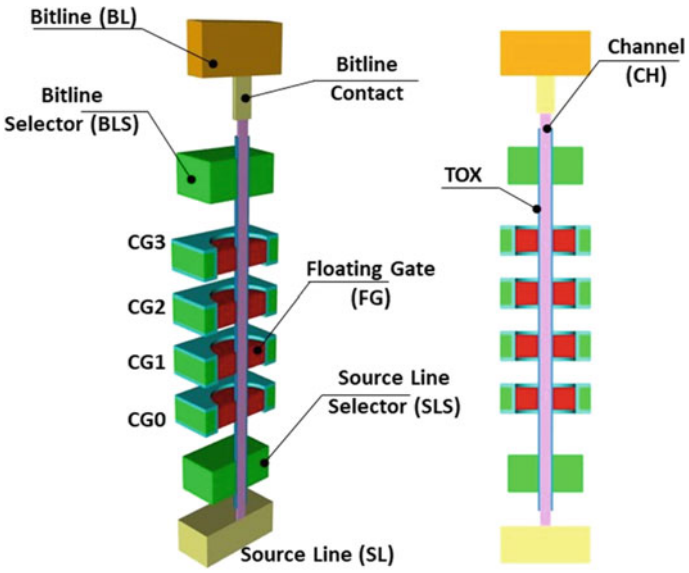


Fig. 15 C-FG NAND flash string. Adapted with permission from [19]. ©2017 IEEE

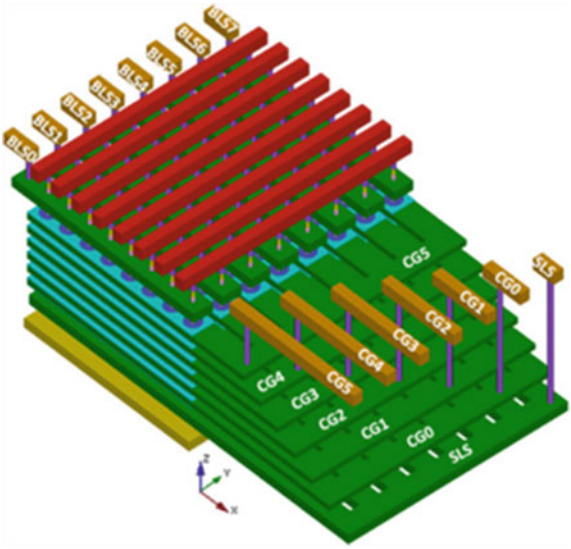


Fig. 16 C-FG NAND flash array with fan-out

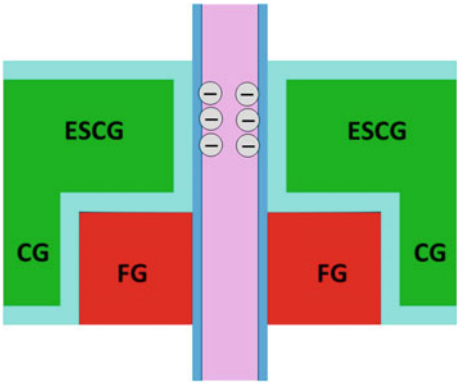


Fig. 17 ESCG NAND flash cell

Another FG cell is DC-SF (*Dual Control-Gate with Surrounding Floating Gate,* Fig. 18) [47]. This time FG is controlled by two CGs. The impact on the FG/CG coupling ratio is remarkable, thanks to the enlargement of the FG/CG overlap area. Another positive aspect is the reduction of the voltages required for programming and erasing. DC-SF eliminates the FG-FG interference because the CG between two adjacent FGs plays the role of an electrostatic shield [48].

FG is fully isolated by IPD (*Inter Poly Dielectric*) and capacitive coupled to upper and lower control gates, CGU and CGL, respectively. The tunnel oxide is located between the channel CH and FG, while IPD is on the sidewall of the CG. In this way, free charges cannot tunnel to the control gates.

BiCS and DC-SF NAND strings are sketched in Fig. 19. In BiCS the nitride layer, going across all gates, makes the cell prone to data retention issues [49]. On

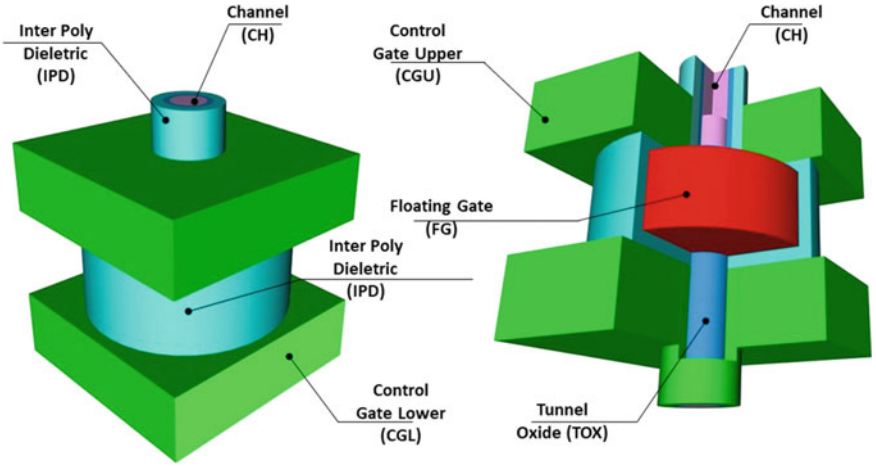


Fig. 18 DC-SF NAND flash cell. Adapted with permission from [19]. ©2017 IEEE

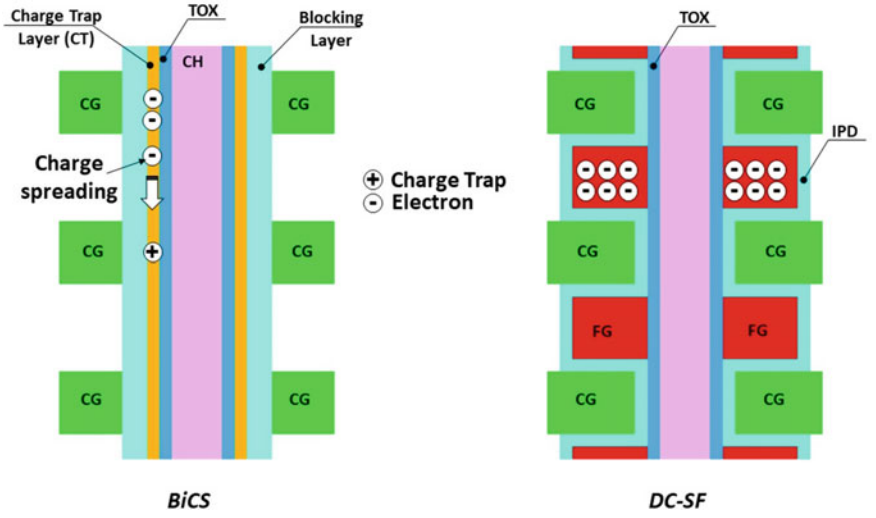


Fig. 19 BiCS versus DC-SF

the contrary, the surrounding FG is totally isolated: it is much easier for DC-SF to retain electrons [50, 51]. Of course, the downside of DC-SF is the fact there are two gate layers instead of one, coupled with much more complex biasing schemes [52, 53].

The *Separated Sidewall Control Gate* (S-SCG) Flash cell [54] displayed in Fig. 20 is another 3D FG option developed around the sidewall concept.

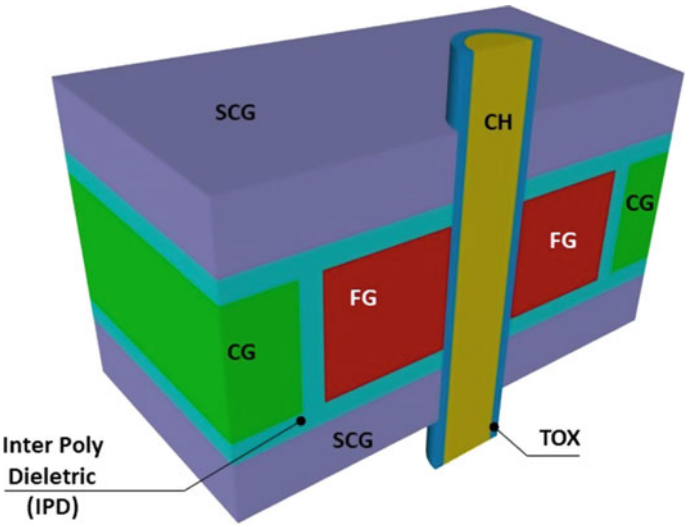


Fig. 20 S-SCG NAND flash cell. Adapted with permission from [19]. ©2017 IEEE

One of major drawbacks of this cell is the "direct" disturb to the neighboring passing cells, caused by the high SCG/FG coupling capacitance. We define it as "direct" because the sidewall CG is shared between adjacent cells: as a matter of fact, biasing SCG means biasing both FGs.

To minimize the decoding complexity, all SCGs belonging to one block adopt a common SCG scheme; besides their electrostatic shield functionality, sidewall gates can help all memory operations [55]. For instance, the common SCG is biased at 1 V during read operations, thus electrically inverting the channel (same as ESCG). Compared to ESCG, the electrical inversion happens simultaneously on source and drain, exactly because of the sidewall gates. Same thing happens during programming: the common SCG is biased at a medium voltage to improve the channel boosting efficiency.

Besides the direct disturb, another problem of Sidewall Gates is the limitation of vertical scaling to around 30 nm; indeed, the thicknesses of SCG and IPD can't be scaled too much, otherwise they would breakdown when voltages are applied.

Let's now take a look at examples of 3D FG NAND memory arrays of hundreds of Gb. The first 3D FG device was published in 2015 [56], in the form of a 384 Gb TLC NAND based on C-FG. This memory device was built with a stack of 32 (+ dummy) memory layers.

A 768 Gb 3D FG NAND became public in the following year [57]. What is unique in this case is the fact that the area underneath the array was used for circuitry. More details about this approach are provided in Sect. 3.

3 Key Challenges for 3D Flash Development

In this Section we cover some of the key challenges that technologists and designers are facing to push 3D memories even further.

3.1 Number of Layers

To reduce the bit size, the number of stacked cells needs to go up, but this causes a bunch of problems hard to solve [6].

Pillar's *Aspect Ratio* (AR) is definitely the first challenge to overcome; in a stack of 32 cells AR can already be as high as 30. In this context, hole etching and gate patterning are extremely difficult, but of paramount importance.

A possible solution to this problem is to divide the stacking process in more steps to reduce the corresponding AR. For example, a NAND string made of 128 cells can be divided in 2 groups of 64 cells each, as shown in Fig. 21. The downside of this solution is the cost of the stacking process (in this example, 4 times higher than the cost of the plain solution).

Second problem is the small cell current [58]. With 2D sensing schemes, a 200 nA/cell saturation current is considered the right value because it gives a reasonable sensing margin. Unfortunately, already with a stack of 24 layers, the cell current is just

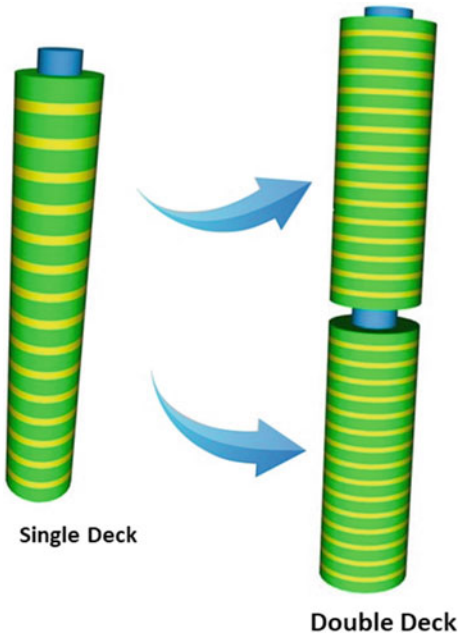


Fig. 21 Multi-stacked or multi-deck process [6]

~20% of FG cell. And it becomes lower and lower as the number of cells in the vertical stack increases. There are a couple of possible paths to solve this problem: sensing schemes with higher sensitivity, and the introduction of new materials enabling a higher cell mobility in the poly-Si channel (i.e. a higher current) [59–62].

All the above-mentioned problems can be fixed if entire NAND strings could be stacked one on top of each other. In this case, either bitlines or source lines are fabricated between NAND strings. This special architecture can simultaneously reduce the aspect ratio and increase the sensing current at same time.

3.2 Peripheral Circuits Under Memory Arrays

In the first 3D generations [63, 64], peripheral circuits (charge pumps, logic, etc.) and core circuits (like Page Buffers and Row decoders) are located outside the memory matrix, like in a conventional 2D chip floorplan, as sketched in Fig. 22. However, 3D memory cells are vertically stacked: in other words, memory transistors are not formed on the Si substrate; on the contrary, they are built around a deposited poly-Si (vertical pillar). Therefore, 3D architectures allow placing some circuits directly on the Si substrate under the memory array. Of course, this solution offers a significant reduction of the chip size.

Figure 23 shows a layout of a Flash memory with *Circuits Under the Array* (CuA) [65, 66].

This big area saving doesn't come for free. The most important challenge is manufacturing low resistance metal layers under the array: this is absolutely critical for a reliable circuit functionality. Usually, metal layers used in 2D NAND flash memories are made of Cu. However, when circuits are under the array, the high temperature

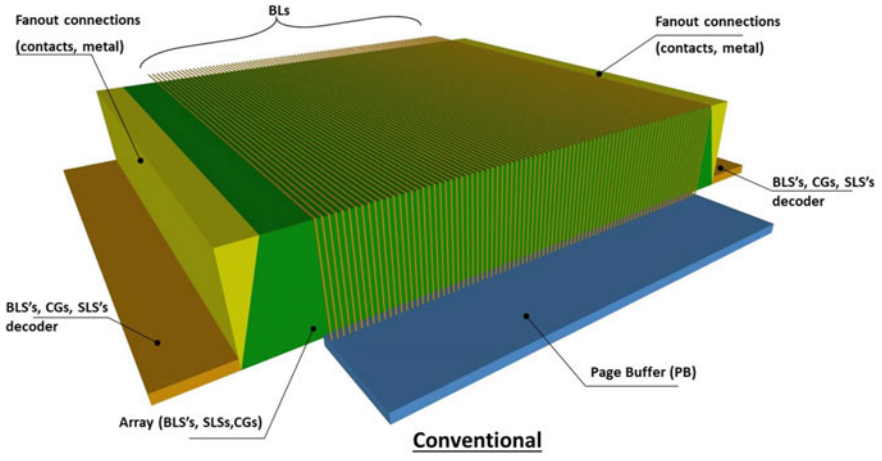


Fig. 22 Conventional 3D NAND flash memory layout

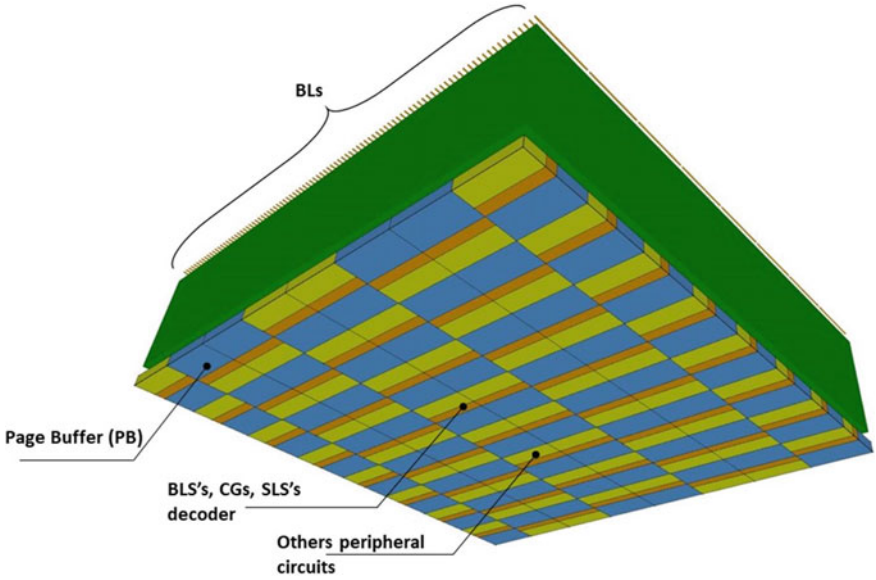


Fig. 23 3D NAND flash memory layout with Circuits Under Array

processes (i.e. >800 °C) that 3D requires can seriously degrade the resistance of metal layers. Therefore, circuits under the array require 3D "low" temperature fabrication processes.

4 Future Trend for 3D NAND Flash

Figure 24 shows cell's size scaling trend, based on published die photographs. 2D became flat below 20 nm, while 3D cell showed a significant reduction going from 24 to 64 layers. This 3D scaling speed will continue by increasing the height of the memory stack, and exploiting technological innovations like Multi-stacked and Stacked NAND string [67].

3D NAND arrays based on CT vertical channel were selected for volume production because the fabrication process is simpler than other 3D architectures. Volume production of 3D NAND Flash started in late 2013 with a 24 layer MLC (2 bit/cell) V-NAND [63, 68]. Year after year, the number of stacked cells grew up, as shown in [7, 64, 69], thus reducing the cost per bit and fueling an even more pronounced diffusion of Solid State Drives.

In this chapter we have presented many architectural options for building a 3D NAND array, including some of the latest and greatest layout options, but the 3D evolution is just at the beginning. In fact, two fundamentally different technologies, Floating and Charge Trap, are fighting each other, trying to prove that they can win

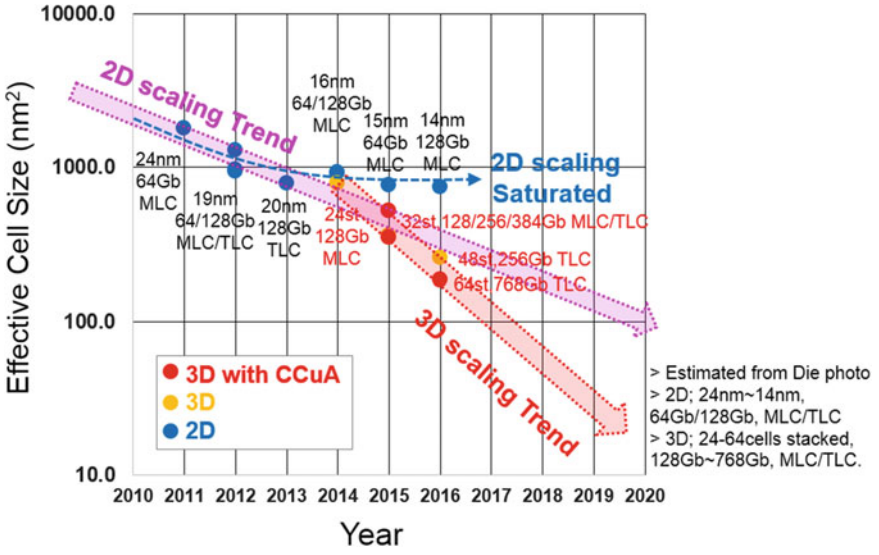


Fig. 24 Effective cell size trend. Reproduced with permission from [19]. ©2017 IEEE

in the long run, i.e. when scaling will be pushed to the limit. Flash manufactures are already shooting for 200 vertical layers with multi-level capabilities, including 4 bit/cell and 5 bit/cell. No doubt that we'll see a lot of innovations in the near future: engineers and scientists are called to give their best effort to make this vertical evolution happen.

References

1. F. Masuoka, M. Momodomi, Y. Iwata, R. Shirota, New ultra high density EPROM and flash EEPROM with NAND structure cell, electron devices meeting. International **33**, 552–555 (1987)
2. R. Micheloni, L. Crippa, A. Marelli, Inside NAND Flash Memories, Chap. 6 (Springer, 2010)
3. T. Mizuno et al., Experimental study of threshold voltage fluctuation due to statistical variation of channel dopant number in MOSFET's. IEEE Trans. Electron Devices **41**(11), 2216–2221 (1994)
4. H. Kurata et al., The impact of random telegraph signals on the scaling of multilevel flash memories, in *Symposium on VLSI Technology* (2006)
5. C.M. Compagnoni et al., Ultimate accuracy for the NAND flash program algorithm due to the electron injection statistics. IEEE Trans. Electron Devices **55**(10), 2695–2702 (2008)
6. S. Aritome, NAND Flash Memory Technologies. IEEE Press Series on Microelectronics System, (Wiley-IEEE Press, 2015)
7. S. Aritome, 3D Flash Memories, International Memory Workshop 2011 (IMW 2011), short course
8. R. Micheloni, L. Crippa, A. Marelli, Inside NAND Flash Memories, Chap. 5 (Springer, 2010)
9. http://www.samsung.com/us/business/oem-solutions/pdfs/VNAND_technology_WP.pdf. Samsung V-NAND technology, White Paper, 2014

10. R. Micheloni, L. Crippa, Chapter 3, Multi-bit NAND flash memories for ultra high density storage devices, in *Advances in Non-volatile Memory and Storage Technology*, ed. by Y. Nishi (Woodhead Publishing, Sawston, 2014)
11. R. Micheloni et al., Chapter 7, High-capacity NAND flash memories: XLC storage and single-die 3D, in *Memory Mass Storage*, ed. by G. Campardo et al. (Springer, 2011)
12. H. Tanaka et al., Bit cost scalable technology with punch and plug process for ultra high density flash memory, in *VLSI Symposium Technical Digest* (2007), pp. 14–15
13. Y. Fukuzumi et al., Optimal integration and characteristics of vertical array devices for ultra-high density, bit-cost scalable flash memory, in *IEDM Technical Digest* (2007), pp. 449–452
14. M. Ishiduki et al., Optimal device structure for pipe-shaped BiCS flash memory for ultra high density storage device with excellent performance and reliability, in *IEDM Technical Digest* (2009), pp. 625–628
15. T. Maeda et al., Multi-stacked 1G cell/layer pipe-shaped BiCS flash memory, in *Digest Symposium on VLSI Circuits* (2009), pp. 22–23
16. R. Katsumata et al., Pipe-shaped BiCS flash memory with 16 stacked layers and multi-level-cell operation for ultra high density storage devices, in *Symposium on VLSI Technology* (2009), pp. 136–137
17. H.-T. Lue, S.-Y. Wang, E.-K. Lai, K.-Y. Hsieh, R. Liu, C.Y. LuA, BESONOS (Bandgap Engineered SONOS) NAND for Post-Floating Gate Era Flash Memory, in *Symposium on VLSI Technology* (2007)
18. H. Aochi, BiCS flash as a future 3-D non-volatile memory technology for ultra high density storage devices, in *Proceedings of International Memory Workshop* (2009), pp. 1–2
19. R. Micheloni, S. Aritome, L. Crippa, Array architectures for 3-D NAND flash memories. Proc. IEEE **105**(9), 1634–1649 (2017). https://doi.org/10.1109/JPROC.2017.2697000
20. Y. Yanagihara et al., Control gate length, spacing and stacked layers number design for 3D-Stackable NAND flash memory2, in *IEEE IMW* (2012), pp. 84–87
21. K. Takeuchi, Scaling challenges of NAND flash memory and hybrid memory system with storage class memory and NAND flash memory, in *IEEE Custom Integrated Circuits Conference (CICC)* (2013), pp. 1–6
22. A. Nitayama et al., Bit Cost Scalable (BiCS) flash technology for future ultra high density storage devices, in *International Symposium on VLSI Technology Systems and Applications (VLSI TSA)*, (2010), pp. 130–131
23. Y. Komori et al., Disturbless flash memory due to high boost efficiency on BiCS structure and optimal memory film stack for ultra high density storage device, in *IEDM Technical Digest* (2008), pp. 851–854
24. J. Kim et al., Novel 3-D structure for ultra high density flash memory with VRAT (vertical-recess-array-transistor) and PIPE (planarized integration on the same plane), in *IEEE Symposium on VLSI Technology* (2008)
25. J. Kim et al., Novel vertical-stacked-array-transistor (VSAT) for ultra-high-density and cost-effective NAND flash memory devices and SSD (solid state drive), in *IEEE Symposium on VLSI Technology* (2009)
26. H.T. Lue, T.H. Hsu et al., A highly scalable 8-layer 3D vertical-gate (VG) TFT NAND flash using junction-free buried channel BE-SONOS device, in *VLSI Symposia on Technology* (2010)
27. J. Jang et al., Vertical cell array using TCAT (terabit cell array transistor) technology for ultra high density NAND flash memory, in *IEEE Symposium on VLSI Technology* (2009)
28. W. Cho et al., Highly reliable vertical NAND technology with biconcave shaped storage layer and leakage controllable offset structure, in *Symposium on VLSI Technology (VLSIT)* (2010), pp. 173–174
29. J. Elliott, E.S. Jung, Ushering in the 3D memory era with V-NAND, in *Proceedings of Flash Memory Summit*, (Santa Clara, CA, 2013), www.flashmemorysummit.com
30. K.-T. Park, Three-dimensional 128 Gb MLC vertical NAND flash memory with 24-WL stacked layers and 50 MB/s high-speed programming, in *IEEE ISSCC, Digest Technical Papers* (2014), pp. 334–335

31. K.-T. Park, Three-dimensional 128 Gb MLC vertical NAND flash memory with 24-WL stacked layers and 50 MB/s high-speed programming. IEEE J. Solid-State Circ. **50**(1), (2015)
32. K.T. Park, A world's first product of three-dimensional vertical NAND flash memory and beyond, in *NVMTS* (27–29 October 2014)
33. K.-S. Shim et al., Inherent issues and challenges of program disturbance of 3D NAND flash cell, in *Memory Workshop (IMW), 2012 4th IEEE International* (20–23 May 2012), pp. 1, 4
34. H.S. Yoo et al., Modeling and optimization of the chip level program disturbance of 3D NAND Flash memory, in *Memory Workshop (IMW), 2013 5th IEEE International* (26–29 May 2013), pp. 147, 150
35. E. Choi et al., Device considerations for high density and highly reliable 3D NAND flash cell in near future, in *IEEE International Electron Devices Meeting* (2012), pp. 211–214
36. K. Shim et al., Inherent issues and challenges of program disturbance of 3D NAND flash cell, in *IEEE International Memory Workshop* (2012), pp. 95–98
37. J.-W. Im, 128 Gb 3b/cell V-NAND flash memory with 1 Gb/s I/O rate, in *IEEE International Solid-State Circuits Conference* (2015), pp. 130–131
38. J.-W. Im, 128 Gb 3b/cell V-NAND flash memory with 1 Gb/s I/O rate. J. Solid-State Circuits. **51**(1), (2016)
39. D. Kang et al., 256 Gb 3b/Cell V-NAND flash memory with 48 stacked WL layers, in *IEEE International Solid-State Circuits Conference (ISSCC), Digest Technical Papers* (2016), pp. 130–131
40. C. Kim et al. A 512 Gb 3b/cell 64-stacked WL 3D V-NAND flash memory, in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2017 IEEE International* (2017) pp. 202–203
41. R. Yamashita et al., A 512 Gb 3b/cell Flash Memory on 64-Word-Line-Layer BiCS Technology, in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2017 IEEE International* (2017), pp. 196–197
42. T. Endoh et al., Novel ultra high density flash memory with a stacked-surrounding gate transistor (S-SGT) structured cell. IEDM Tech. Dig. 33–36 (2001)
43. T. Endoh et al., Novel ultra high density flash memory with a stacked-surrounding gate transistor (S-SGT) structured cell.2. IEEE Trans. Electron Devices **50**(4), 945–951 (2003)
44. T. Endoh et al., Floating channel type SGT flash memory, in *The 1999 Joint International Meeting*, Hawaii, vol. 99–2, Abstract No. 1323, 17–22 Oct 1999
45. M.S. Seo et al., The 3-dimensional vertical FG nand flash memory cell arrays with the novel electrical S/D technique using the extended sidewall control gate (ESCG), in *Proceedings of IEEE International Memory Workshop* (2010), pp. 1–4
46. M.S. Seo et al., 3-D Vertical FG NAND flash memory with a novel electrical S/D technique using the extended sidewall control gate. IEEE Trans. Electron Devices **58**(9), (2011)
47. S. Whang et al., Novel 3-dimensional dual control gate with surrounding floating-gate (DC-SF) NAND flash cell for 1 Tb file storage application, in *Proceedings of International Electron Devices Meeting (IEDM)* (2010), pp. 668–671
48. Y. Noh et al., A new metal control gate last process (MCGL process) for high performance DC-SF (dual control gate with surrounding floating gate), in *3D NAND flash memory in Symposium on VLSI Technology* (2012), pp. 19–20
49. C.-P. Chen et al., Study of fast initial charge loss and its impact on the programmed states Vt distribution of charge-trapping NAND Flash, in *Electron Devices Meeting (IEDM), 2010 IEEE International* (6–8 Dec 2010), pp. 5.6.1, 5.6.4
50. R. Micheloni, L. Crippa, Multi-bit NAND flash memories for ultra high density storage devices (chapter 3), in *Advances in Non-volatile Memory and Storage Technology*, ed, by Y. Nishi (Woodhead Publishing, 2014)
51. R. Micheloni et al., High-capacity NAND flash memories: XLC storage and single-die 3D (chapter 7) in *Memory Mass Storage*, eds. by G. Campardo et al. (Springer, 2011)
52. H. Yoo et al., New read scheme of variable Vpass-read for dual control gate with surrounding floating gate (DC-SF) NAND flash cell, in *Proceedings of 3rd IEEE International Memory Workshop* (2011), pp. 1–4

53. S. Aritome et al., Advanced DC-SF cell technology for 3-D NAND flash. IEEE Trans. Electron Devices **60**(4), 1327–1333 (2013)
54. M.S. Seo et al., A novel 3-D vertical FG nand flash memory cell arrays using the separated sidewall control gate (S-SCG) for highly reliable MLC operation, in *Proceedings of 3rd IEEE International Memory Workshop (IMW)* (2011), pp. 1–4
55. M.S. Seo et al., Novel concept of the three-dimensional vertical FG nand flash memory using the separated-sidewall control gate. IEEE Trans. Electron Devices **59**(8), 2078–2084 (2012)
56. K. Parat, C. Dennison, A floating gate based 3D NAND technology with CMOS under array, in *Conference on International Electron Devices Meeting (IEDM)* (San Francisco, USA, 2015)
57. T. Tanaka et al., A 768 Gb 3 b/cell 3D-floating-gate NAND flash memory, in *2016 IEEE International Solid-State Circuits Conference (ISSCC), Digest of Technical Papers* (San Francisco, USA, 2016), pp. 142–143
58. E.-S. Choi, S.-K. Park, Device considerations for high density and highly reliable 3D NAND flash cell in near future, in *Electron Devices Meeting (IEDM), 2012 IEEE International*, vol. no., pp. 9.4.1–9.4.4, 10–13 Dec 2012
59. Subirats et al., Impact of discrete trapping in high pressure deuterium annealed and doped poly-Si channel 3D NAND macaroni, in *IEEE International Reliability Physics Symposium (IRPS)* (2017)
60. L. Breuil, Improvement of poly-Si channel vertical charge trapping NAND devices characteristics by high pressure D2/H2 annealing, in *IEEE 8th International Memory Workshop (IMW)* (2016)
61. E. Capogreco et al. MOVPE In1-xGaxAs high mobility channel for 3-D NAND memory, in *IEEE International Electron Devices Meeting (IEDM)* (2015)
62. J. G. Lisoni et al., Laser Thermal Anneal of polysilicon channel to boost 3D memory performance, in *Symposium on VLSI Technology (VLSI-Technology)*, Digest of Technical Papers (2014)
63. K.-T. Park et al., Three-Dimensional 128 Gb MLC Vertical nand Flash Memory With 24-WL Stacked Layers and 50 MB/s High-Speed Programming. IEEE J. Solid-State Circuits **50**(1), 204–213 (2015)
64. J.-W. Im et al., A 128Gb 3b/cell V-NAND Flash Memory with 1Gb/s I/O rate, in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2015 IEEE International* (2015), pp. 23–25
65. S. Aritome, NAND flash memory revolution, in *2016 IEEE 8th International Memory Workshop (IMW)* (Paris, 2016), pp. 1–4
66. T. Tanaka et al., 7.7 A 768Gb 3b/cell 3D-floating-gate NAND flash memory, in *2016 IEEE International Solid-State Circuits Conference (ISSCC)* (San Francisco, CA, 2016), pp. 142–144
67. R. Micheloni (ed.), *3D Flash Memories* (Springer, 2016)
68. K.-T. Park et al., 19.5 Three-dimensional 128Gb MLC vertical NAND Flash-memory with 24-WL stacked layers and 50MB/s high-speed programming, in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2014 IEEE International* (9–13 Feb 2014) pp. 334–335
69. S. Aritome, Joint Rump session in VLSI Symposium 2012, Scaling challenges beyond 1Xnm DRAM and NAND Flash