Chapter 4
3D Charge Trap NAND Flash Memories

Luca Crippa and Rino Micheloni

4.1 Introduction

Different criteria can be adopted to sort 3D architectures but, probably, the classification based on topological characteristics is the most effective since the choice of a specific topological organization has a direct impact on cost, electrical performances and process technology integration. The following cases can be considered:

- horizontal channel and gate (Chap. 3);
- vertical channel and horizontal gate;
- horizontal channel and vertical gate (mainly Chap. 7).

The array with horizontal channel and gate is the 3D Stacked option discussed in the previous chapter. In the 3D approach with horizontal gate and vertical channel, the planar (2D) NAND Flash string of Fig. 4.1a is rotated by 90°, as shown in Fig. 4.1b. In order to improve electrical performances, a channel fully wrapped around by gate is adopted (Fig. 4.1c and d) [1]. With this specific configuration, thanks to the curvature effect, it is possible to enhance the electric field across the tunnel oxide and to relax the electric field across the blocking oxide [2, 3], thus improving power and reliability (Chap. 2).

To simplify descriptions, in this book "vertical channel with horizontal gate" is shortened to "vertical channel".

This chapter starts off with 2 vertical channel architectures named BiCS (*Bit Cost Scalable*) and P-BiCS (*Pipe-Shaped* BiCS), respectively. BiCS was proposed for the

L. Crippa (✉) · R. Micheloni
Performance Storage BU, Microsemi Corporation, Vimercate, Italy
e-mail: luca.crippa@ieee.org

R. Micheloni
e-mail: rino.micheloni@ieee.org

R. Micheloni (ed.), *3D Flash Memories*, DOI 10.1007/978-94-017-7512-0_4

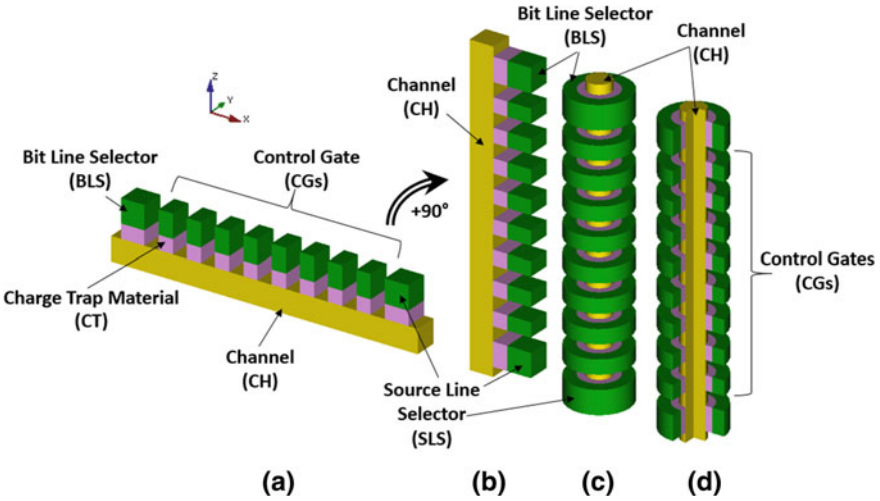

Fig. 4.1 NAND Flash string with horizontal gate and vertical channel: **a** planar, **b** planar rotated by 90°, **c** vertical channel with cylindrical shape and **d** its cross section

first time by Toshiba in 2007 [4, 5], and another version called P-BiCS was presented in 2009 [6–8] to improve retention, source selector performances and source line resistance. Both options are based on a Charge Trap memory cell (Chap. 2). BiCS can definitely be considered as a major milestone in the history of 3D Flash.

In the next 2 sections we will dig into the details of BiCS and P-BiCS, while the second part of the chapter is devoted to the evolutionary path of V-NAND, which is the first 3D architecture that reached volume production.

4.2 BiCS

Bird's-eye views of the BiCS Flash memory are sketched in Fig. 4.2, while Figs. 4.3 and 4.4 include diagrams of the equivalent circuit [5].

First of all, there is a stack of *Control Gate* (CG) plates (green rectangles). The lowest gate plate corresponds to the gate of the *Source Line Selector* (SLS) of the NAND string. Holes are punched through the entire stack and plugged with poly-silicon, thus forming a series of vertical memory cells connected in a NAND architecture. At the top of the structure we have *Bit Line Selectors* (BLS's) and *Bitlines* (BLs) [9].

Memory cells work in depletion-mode [2, 3] with the body poly-silicon being un-doped or lightly uniformly n-doped, to avoid the process complexity of forming p-n junctions within the vertical polysilicon plug (or pillar). The intersection of a control gate plate with a pillar identifies a single memory cell; each string is connected to a bitline through an upper select transistor (BLS). The bottom of the

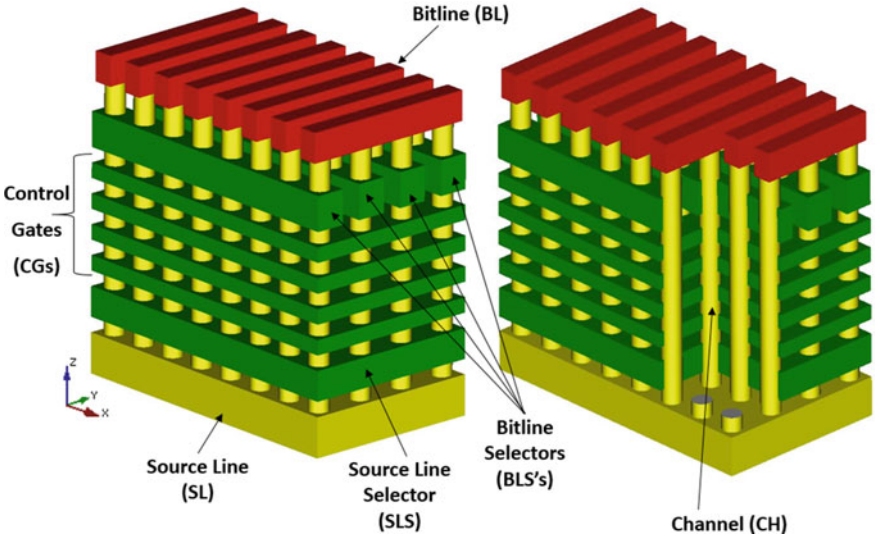

Fig. 4.2 BiCS architecture

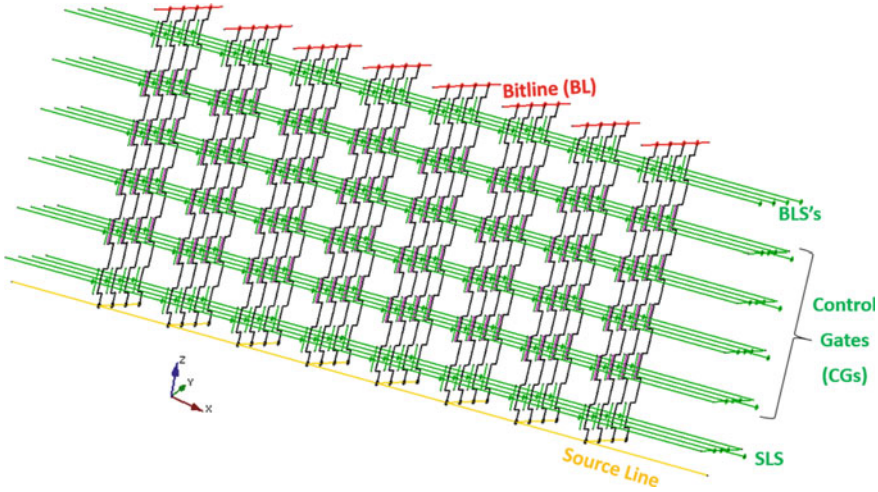

Fig. 4.3 Schematic circuit of a BiCS array

memory string is connected to the common source diffusion, which is formed on the silicon substrate. Figure 4.5 highlights a single NAND Page in the SLC case (i.e. 1 bit/cell); basically, there is one selected memory cell per bitline.

Bit density can be increased by adding more control gate plates [10, 11], while the number of the critical lithography steps remains constant because the whole stack of control gates is completely punched through with one lithography step only.

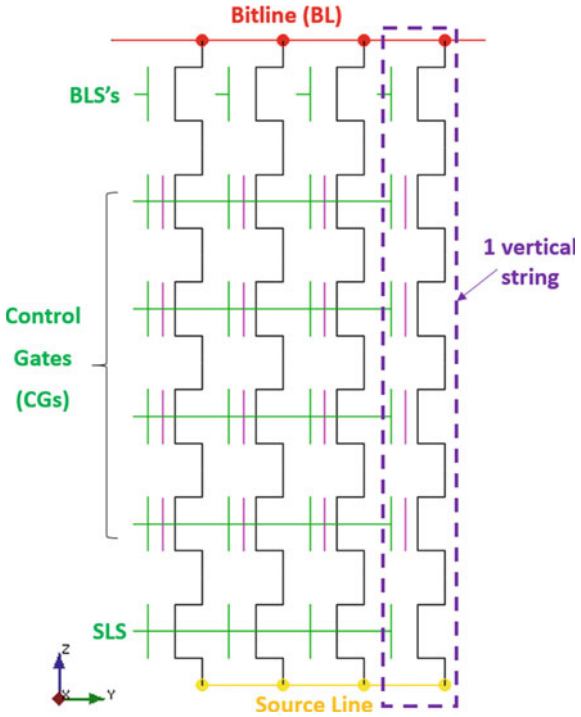

Fig. 4.4 Y-Z view of BiCS equivalent circuit schematic

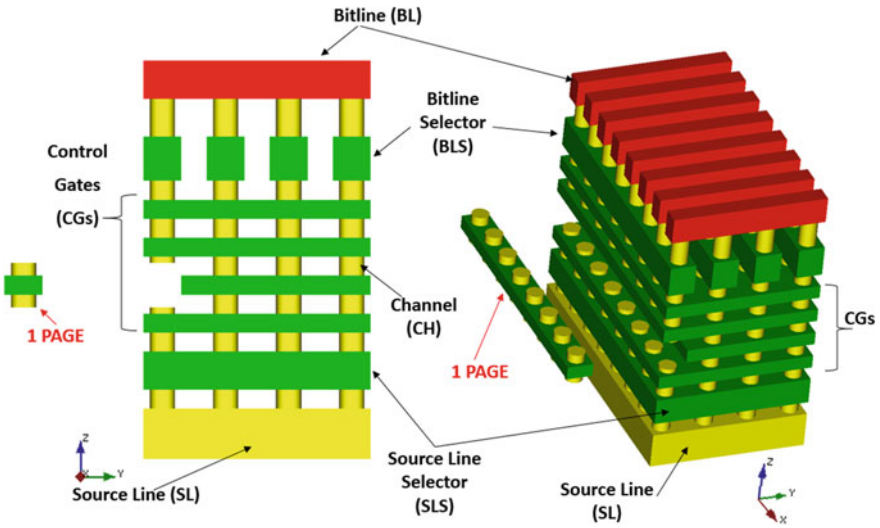

Fig. 4.5 SLC NAND flash page in the BiCS architecture

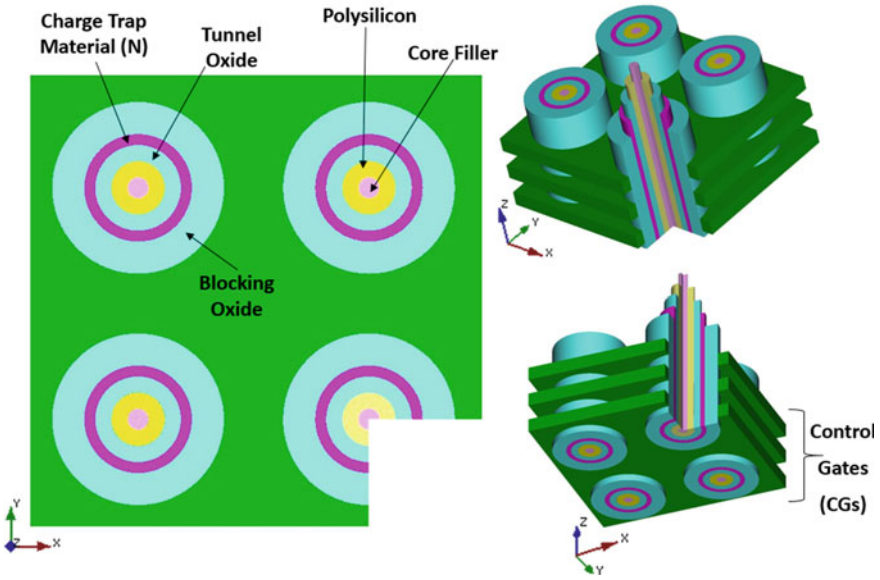

Fig. 4.6 BiCS memory cells

As discussed in Chap. 2, in the BiCS design the floating gate is replaced by a charge trap material. Multiple views of BiCS cell's structure are reported in Fig. 4.6, where tunnel oxide, field oxide, polysilicon channel and core filler are highlighted.

It is worth highlighting the fact that the body of the vertical transistor is completely made of polysilicon; in fact, the aspect ratio of the punched hole is so high that it becomes almost impossible to achieve good production yield without non-selective deposition process steps.

Unfortunately, is such a configuration, it is very difficult to control the trap density at grain boundary, and this causes a large variation of the sub-threshold characteristics of the vertical transistor. In order to reduce the trap density fluctuation, the body polysilicon needs to be much thinner than the depletion width. Basically, the concept is to reduce the volume of the polysilicon and, therefore, the total number of traps, as sketched in Fig. 4.7. Given its shape, the body of this vertical transistor is referred to as *Macaroni Body* [5]. The center of the macaroni shape is filled with dielectric film (*filler* in the following) to make the 3D process integration easier.

In Figs. 4.8 and 4.9 the basic building block of BiCS memory is split in different sections in order to provide a greater level of details for the different elements [6, 9]. Cell's structure is clearly visible in the Middle section where the reader can find all the layers seen in Fig. 4.6. In the Upper section it is worth highlighting that the charge trap material stops before the bitline selector, which is a standard transistor built with field oxide only. In other words, both the tunnel oxide and the charge trap

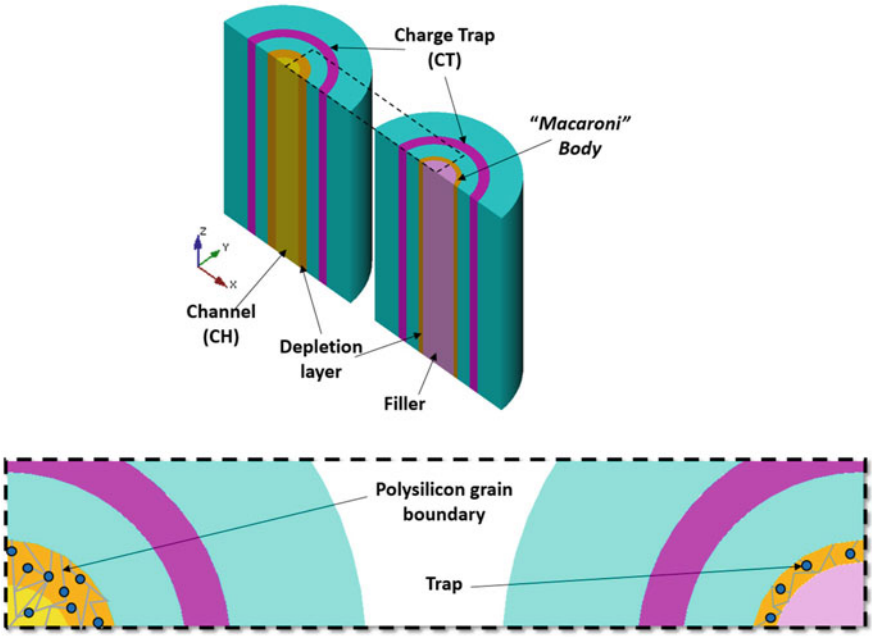

Fig. 4.7 Vertical transistor with (*right*) and without (*left*) *macaroni* body

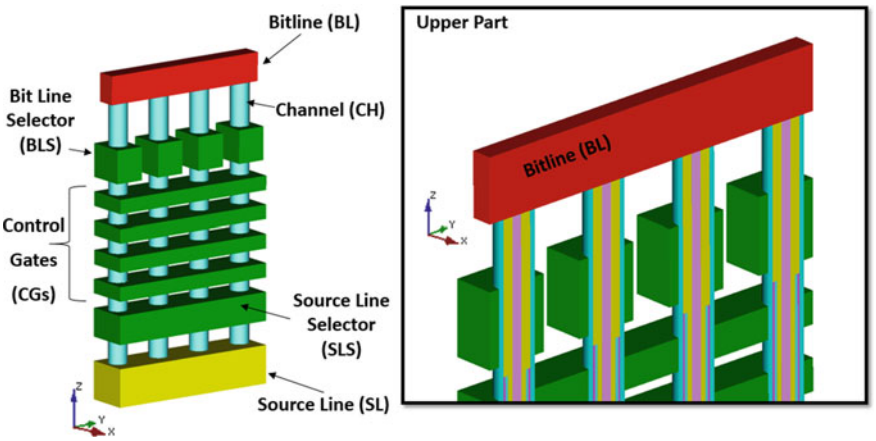

Fig. 4.8 BiCS vertical cross-section: *upper part*

material are replaced by the polysilicon. The same applies to the Source Line Selector which becomes a standard nMOS transistor, as shown in Fig. 4.9.

Figures 4.10 and 4.11 show a simplified fabrication sequence of a BiCS array [12]. In the first step all the plates (green parallelepipeds) for SLS, control gates, and bitline selectors are manufactured. Then each single stripe of BLS's is defined.

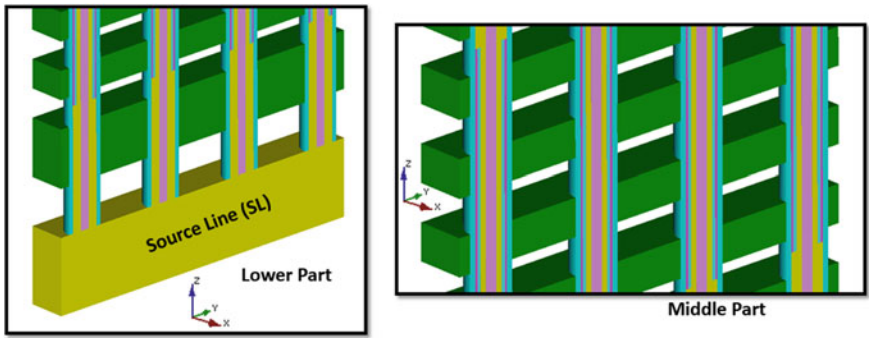

Fig. 4.9 BiCS vertical cross-section: *middle* and *lower parts*

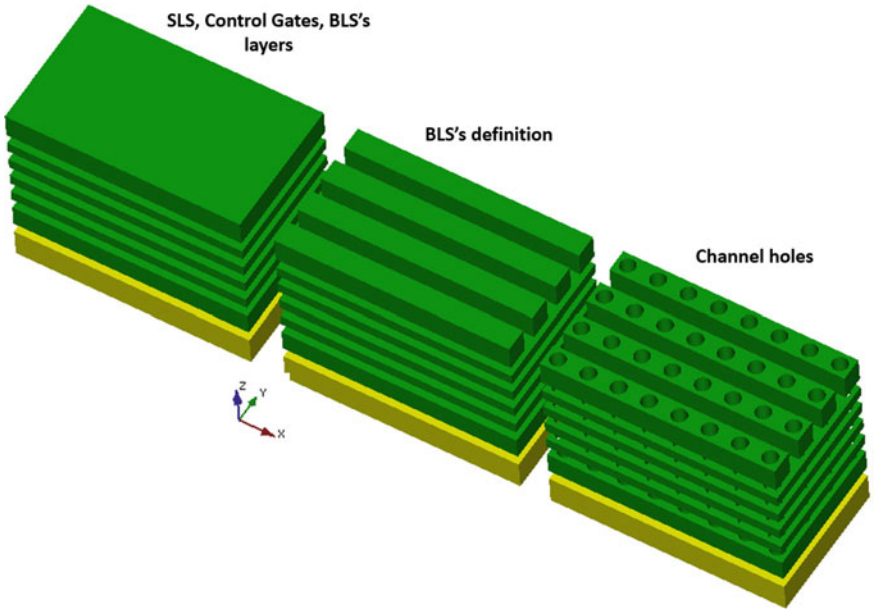

Fig. 4.10 Simplified BiCS fabrication sequence: CGs, BLS definition, and channel holes

At this point the stack is ready to be drilled: in Fig. 4.10 channel holes are clearly visible. Figure 4.11 shows the next step: each single channel hole is filled with polysilicon (yellow cylinder), thus forming a pillar. Bitlines are defined during the back-end phase. To sum up, the BiCS cell array consists of multi-stacked control gate plates and polysilicon pillars fabricated through control gate plates. Memory cells are placed at the intersections of control gate plates and polysilicon pillars.

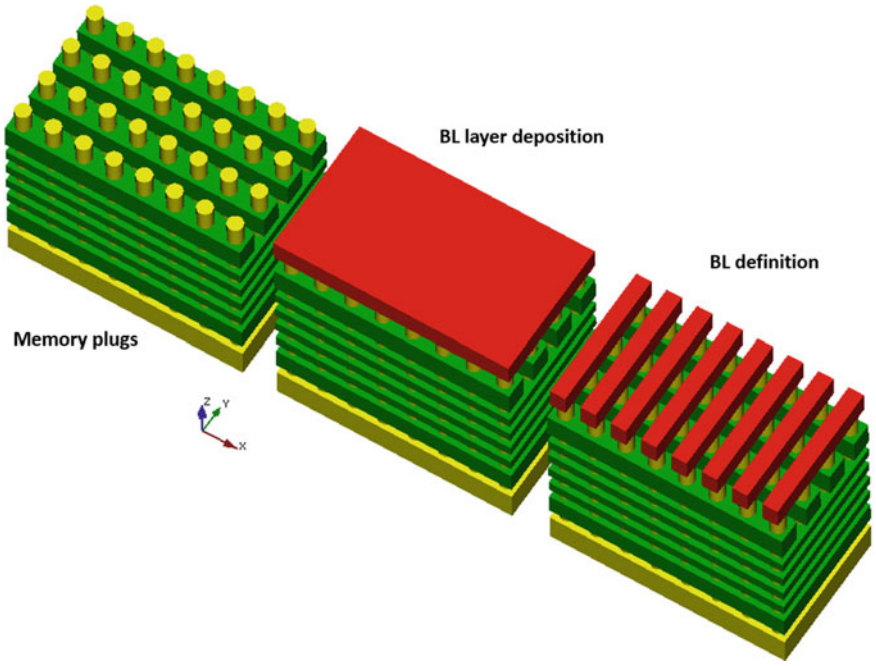

Fig. 4.11 Simplified BiCS fabrication sequence: memory plugs, BL layer deposition, and BL definition

In the following, in order to simplify the drawings, memory cell's structure is defined by the Channel layer (yellow) only.

Edges of control gate plates form a double-sided staircase structure as sketched in Fig. 4.12 [4, 5, 12, 13]. All the plates (CGs, SLS, SL) are contacted on the same side (right hand side in this example), because the other one (left hand side) is used for connecting bitline selectors. Orange rectangles represent metal connections. Figures 4.13 and 4.14 are the top and bottom bird's-eye views of Fig. 4.12.

As we have seen in Fig. 4.5, multiple Flash Pages live on the same CG plate, thus amplifying the impact of disturbs, especially during the programming phase. For minimizing disturbs, the whole stack of control gates, SLS, and SL is etched to form a slit, which, as a matter of fact, creates blocks of memory plugs. A vertical cross section of 3 adjacent blocks is shown in Fig. 4.15, while Fig. 4.16 is the corresponding bird's-eye view [12].

Historically, BiCS evolved in a different architecture called P-BiCS, which is the subject of the following section.

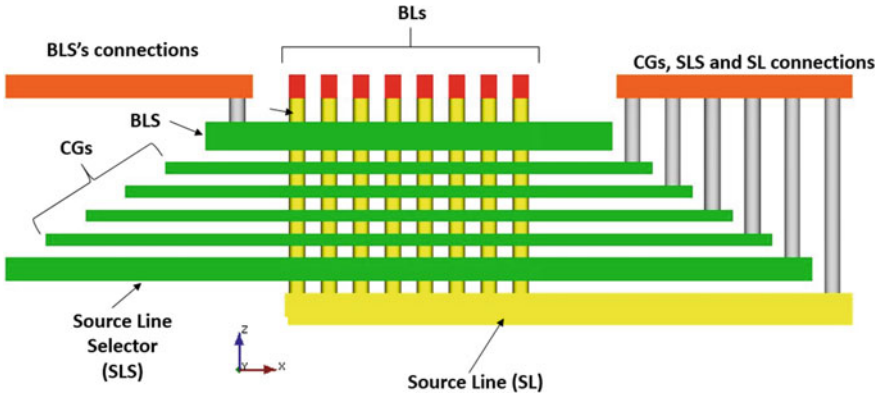

Fig. 4.12 Vertical cross section of BiCS with gate connections

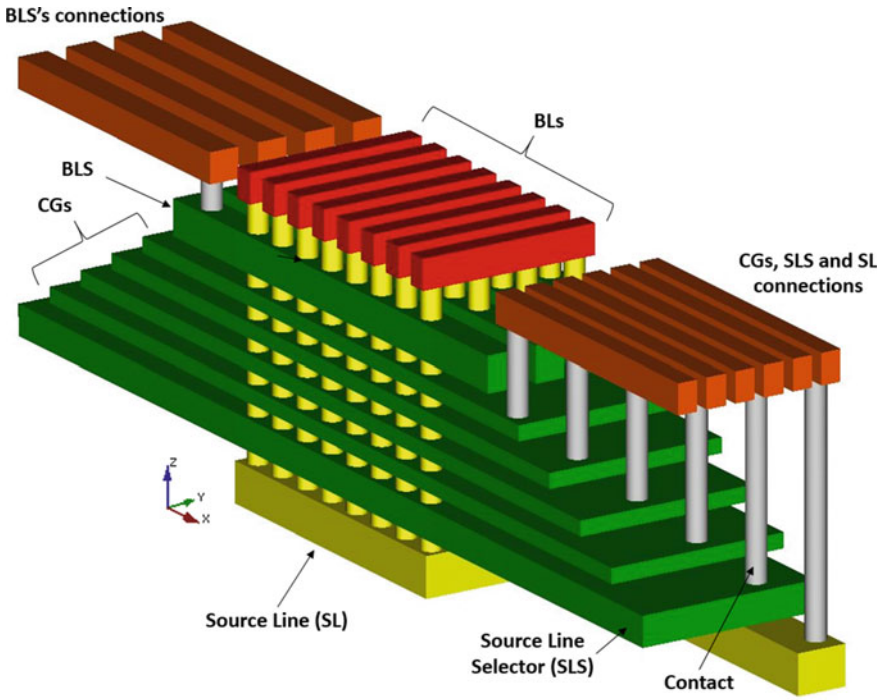

Fig. 4.13 Top bird's eye view of BiCS with gate connections

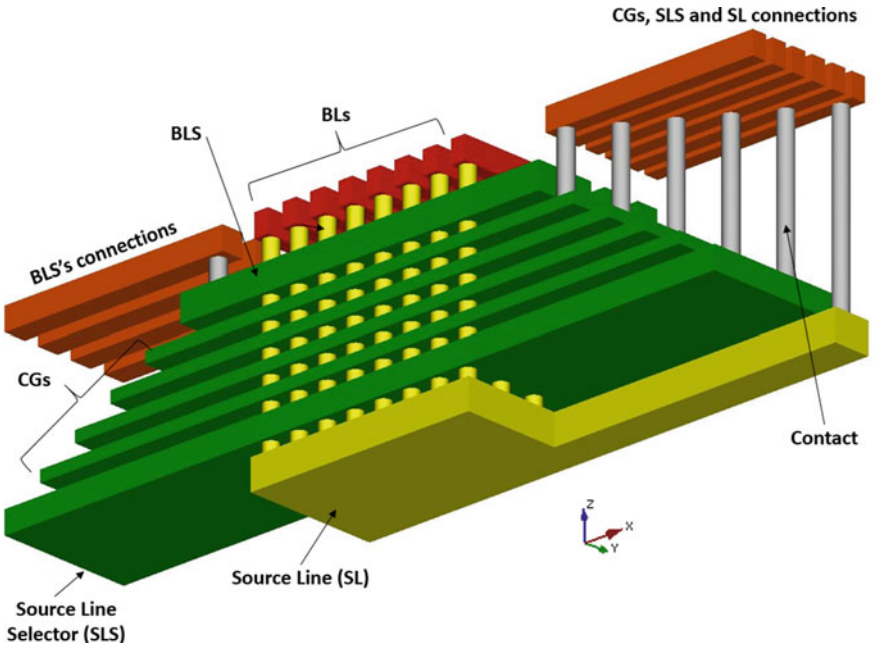

Fig. 4.14 Bottom bird's-eye view of BiCS with gate connections

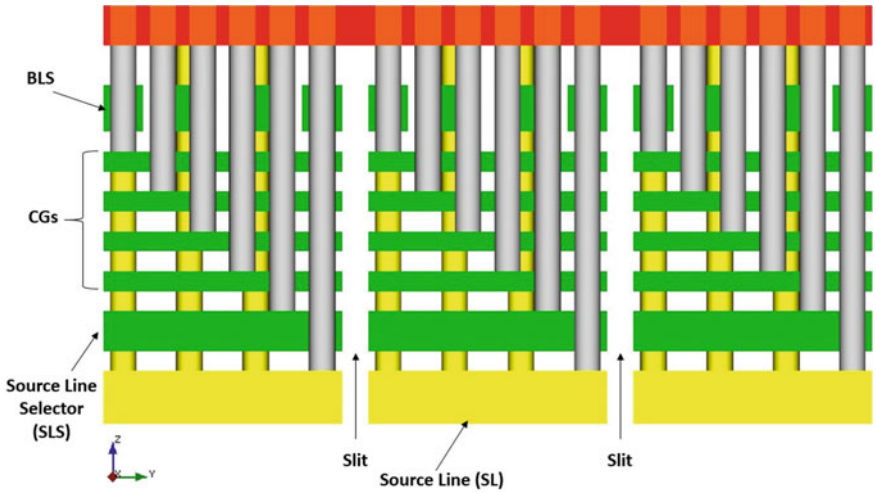

Fig. 4.15 Front view of 3 adjacent blocks to highlight the presence of slits

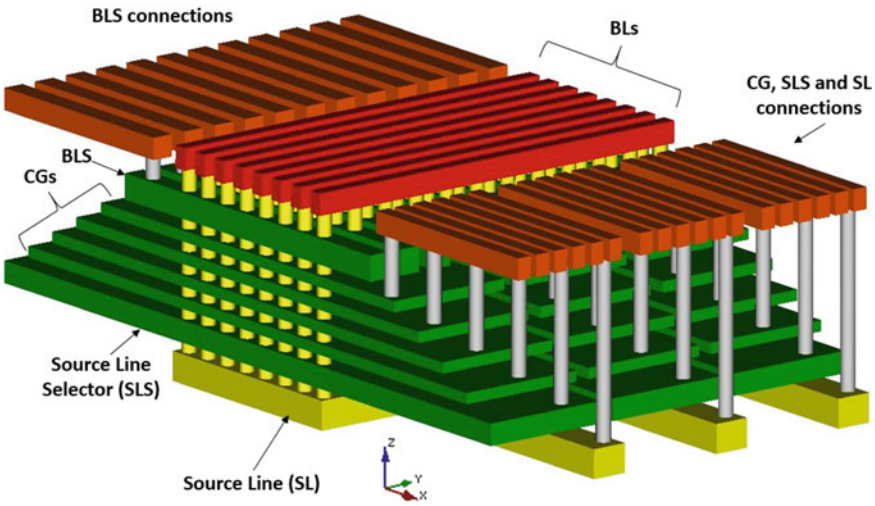

Fig. 4.16 Top bird's-eye view of 3 adjacent blocks

4.3 P-BiCS

P-BiCS (*Pipe-Shaped* BiCS) Flash was developed to solve some critical problems of BiCS, such as poor cell's reliability, poor cut-off characteristics of the Source Line Selector (placed at the bottom of BiCS), and high resistance of the Source Line [6, 7]. By adopting a U-shape vertical NAND string, as shown in Fig. 4.17, P-BiCS solves all the above mentioned problems. In particular, P-BiCS shows three main advantages compared to a straight-shaped BiCS:

- P-BiCS has better data retention and a wider V_{TH} window because the fabrication process is less stressful for the tunnel oxide;
- Because the Source Line is placed at the top of the stack, it is easier to connect it to a metal mesh, which results in a lower parasitic resistance;
- The Source Line Selector is at the same height of the BLS and, therefore, its cut-off characteristics can be tightly controlled, thus improving the functionality of the memory array itself.

A P-BICS array is built starting from the basic structure shown in Fig. 4.17 [7]. Two NAND strings are connected via the so-called Pipe Connection at the bottom of the structure, thus forming a U-shaped string (highlighted in cyan): one of the terminals of this "U" is connected to the bitline, while the other one is tied to the Source Line. It is worth pointing out that the 2 NAND strings are mirrored such that they can share the same source line plate (blue parallelepiped).

The basic building block of Fig. 4.17 can be replicated multiple times along the X direction to form a memory array as shown in Figs. 4.18, 4.19 and 4.20.

Figure 4.21 is the schematic of the equivalent circuit [8].

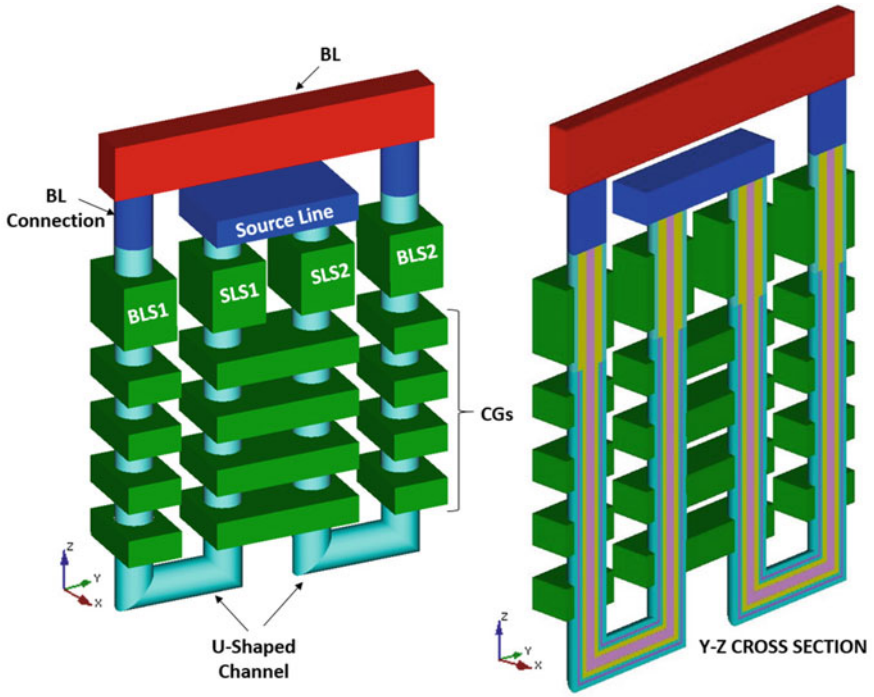

Fig. 4.17 Basic building of a P-BICS NAND flash and its vertical cross-section

For the sake of clarity, Blocking Oxide and Tunnel Oxide layers have been removed from Fig. 4.18 onward; only Charge Trap (pink) and Channel (yellow) are visible.

At this point we are ready to build a bigger array by replicating the basic building block of Fig. 4.17 along both X and Y directions, as sketched in Figs. 4.22 and 4.23.

For a better understanding of Fig. 4.22, it is worth removing some of the backend structures, as shown in Fig. 4.24. Starting from the right hand side, the reader can see how the array would look like by removing bitlines, and bitlines and Source Line at the same time (far left in Fig. 4.24).

As already pointed out, the Source Line plate is shared among all the NAND strings which are adjacent along both the X and the Y directions. This mirrored architecture minimizes the parasitic resistance of the source connection.

Next part of the manufacturing process is designed to build all the necessary connections for biasing CGs, BLs, BLS's, SLS's, and SL [8]. As we have seen in the previous section, in the BiCS architecture each control gate plate is shared by several neighboring rows of NAND strings, in order to reduce the silicon area. In P-BiCS this is not possible as 2 different control gates of the same NAND string belong to the same layer within the stack. As a result, a branched control gate

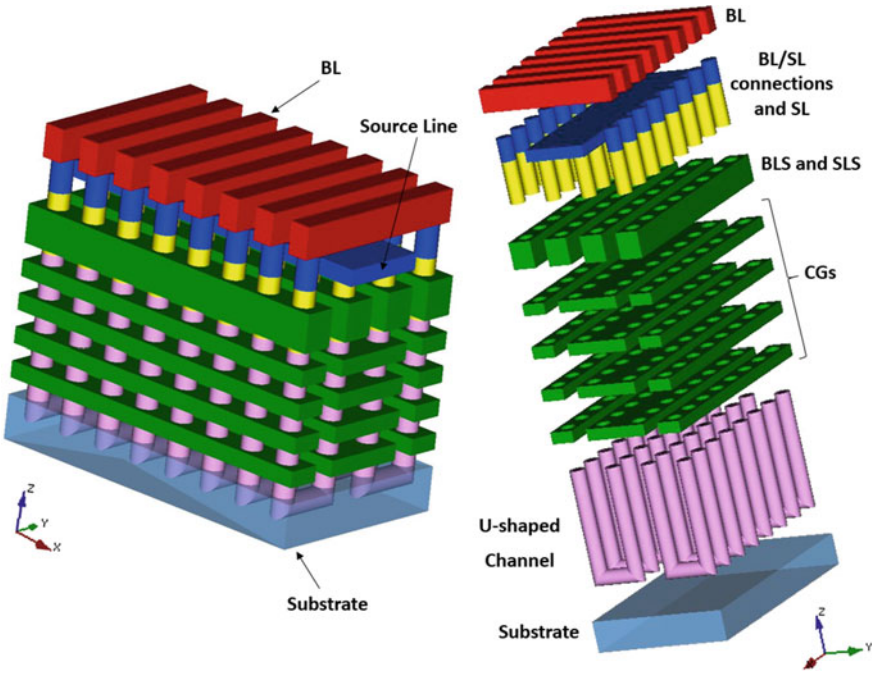

Fig. 4.18 Top bird's-eye view of a P-BiCS NAND array

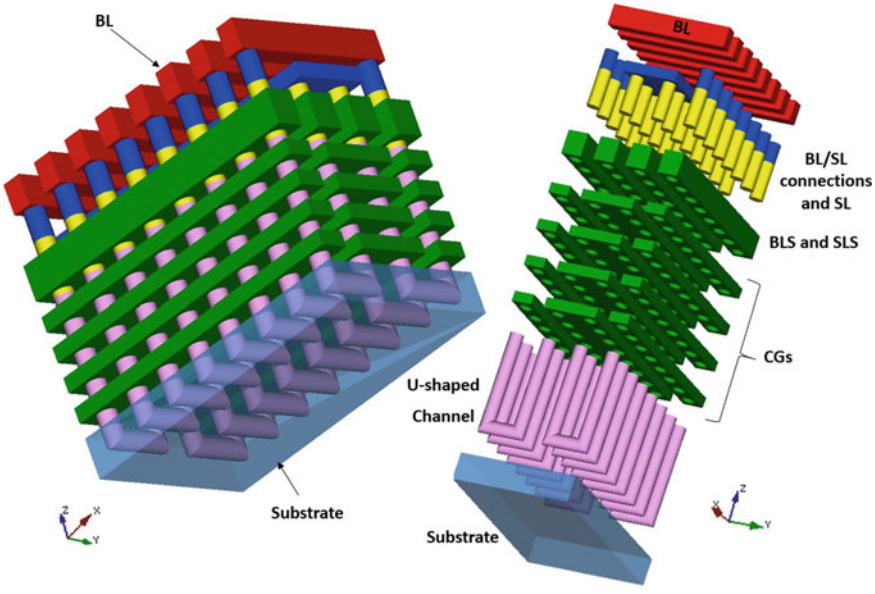

Fig. 4.19 Bottom bird's-eye view of a P-BiCS NAND array

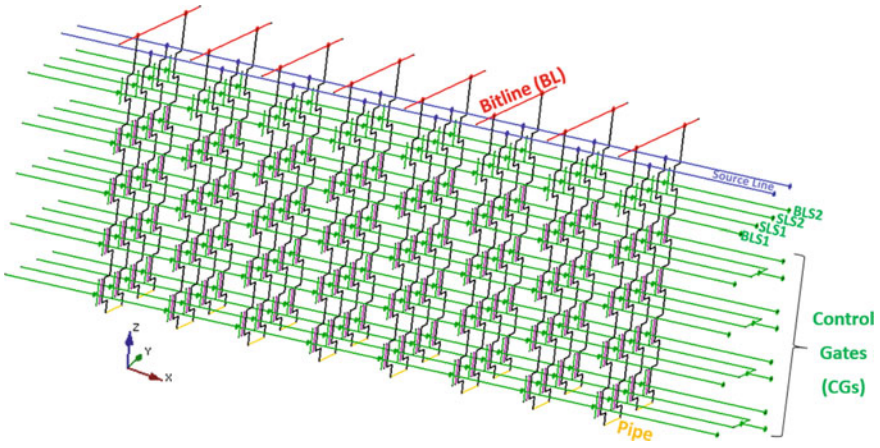

Fig. 4.20 P-BiCS circuit schematic

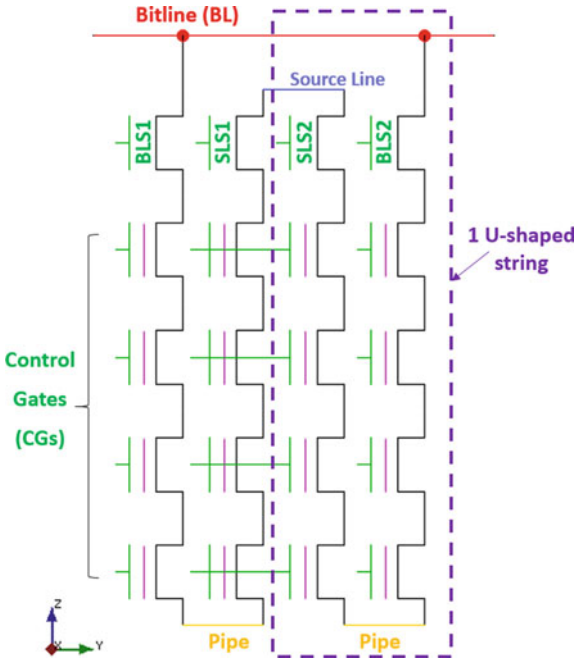

Fig. 4.21 Y-Z view of P-BiCS circuit schematic

configuration is adopted: Figs. 4.25 and 4.26 are bird's-eye views of P-BiCS with all the array connections, while the corresponding X-Z cross-section is drawn in Fig. 4.27. In this example, a NAND string of 8 Flash cells is considered.

To better appreciate the construction details of Figs. 4.25, 4.28 and 4.29 show a zoom of the right hand side from different angles.

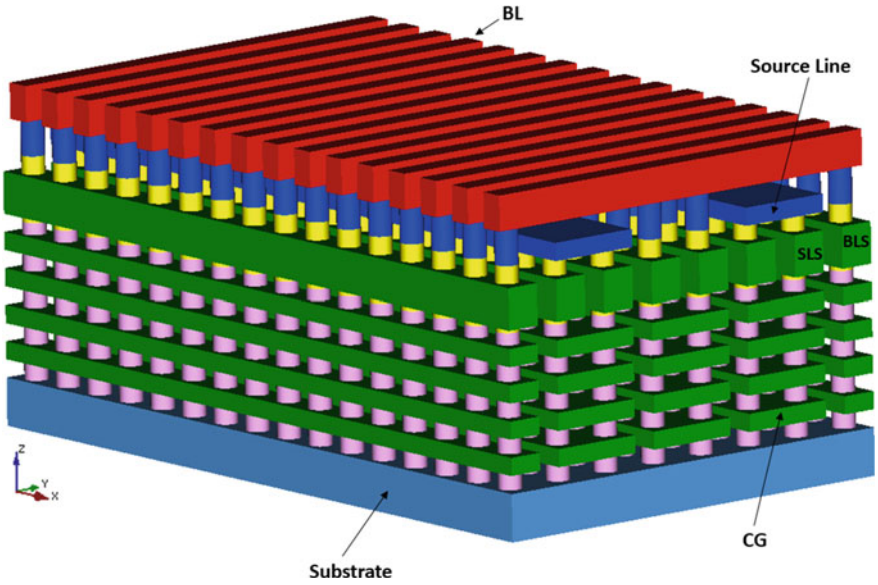

Fig. 4.22 P-BiCS NAND array

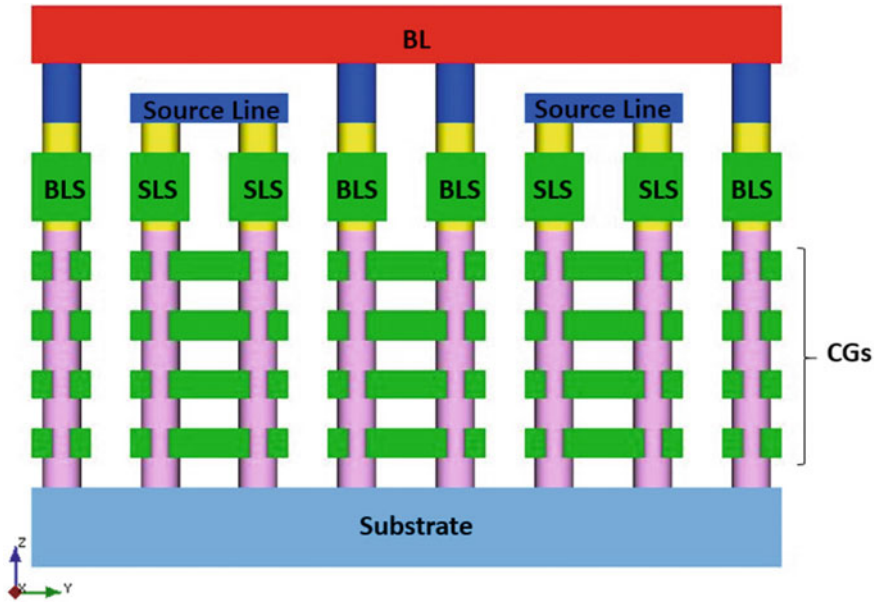

Fig. 4.23 Y-Z cross section of Fig. 4.22

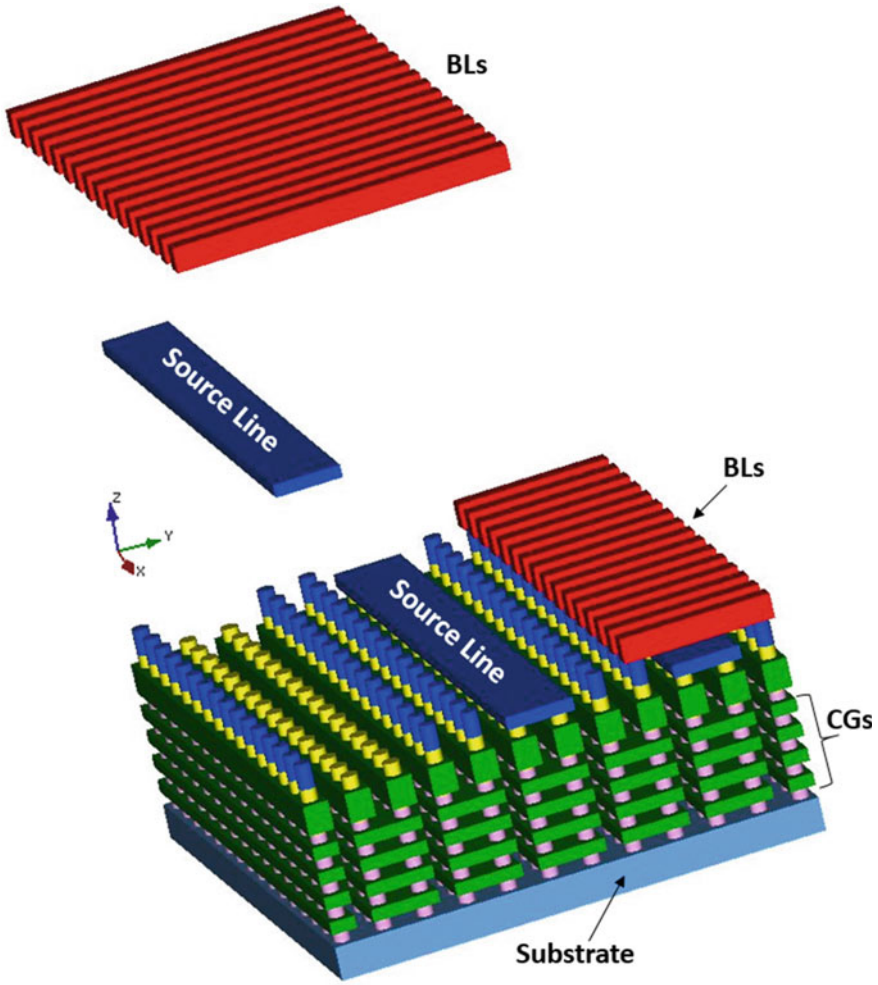

Fig. 4.24 P-BiCS array with partial removal of source line and bitlines

Let's now have a closer look to the control gates, which are realized by using fork-shaped plates [8]. Please refer to Figs. 4.30 and 4.31. Each branch of the fork controls cells of two adjacent pages. Indeed, a P-BiCS NAND string of 8 cells is formed by vertically stacking 4 pairs of control gates. Each pair of control gate plates is arranged in a staggered layout (please refer to CG0 and CG7 in Fig. 4.25 as an example). Of course, this architecture can be scaled to include more cells in the NAND string. For example, with 16/24 layers we can build strings of 32/48 cells.

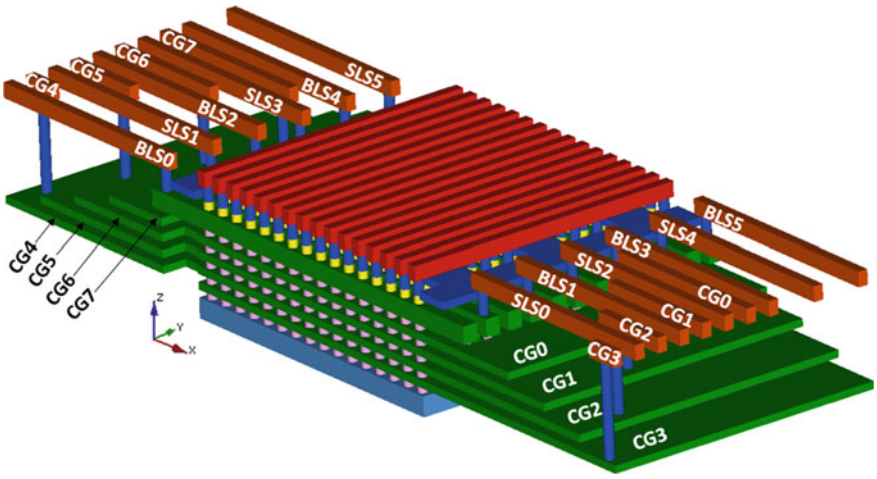

Fig. 4.25 Bird's-eye view of P-BiCS with array connections

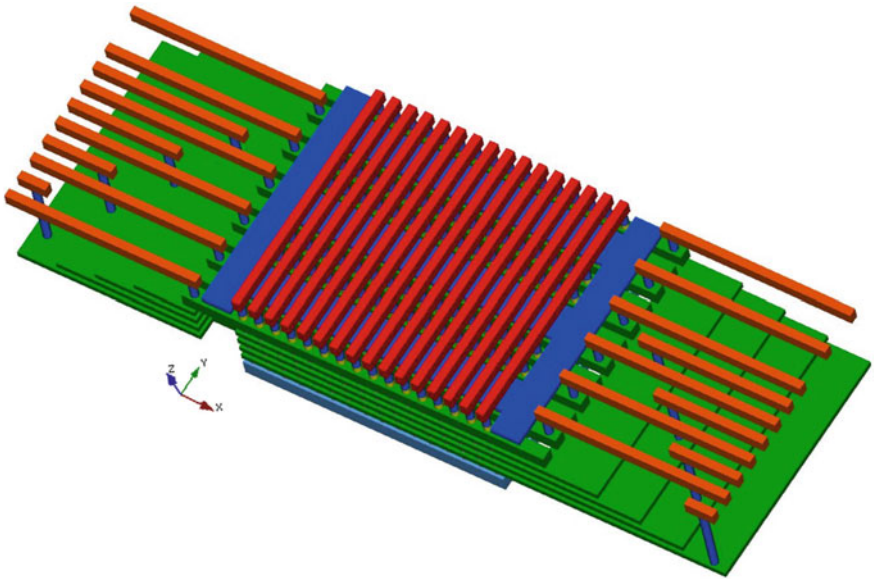

Fig. 4.26 Top view of Fig. 4.25

At the bottom of Fig. 4.30 it is possible to see some U-shapes (in pink) that have been placed to better identify NAND strings.

One other major difference of P-BiCS versus BiCS is the fact that the source line is placed at the top of the 3D structure [6]. In order to enhance noise immunity as much as possible, it is very important to have a source line with a very low resistance. For this reason, another level of source line connection is used on top of

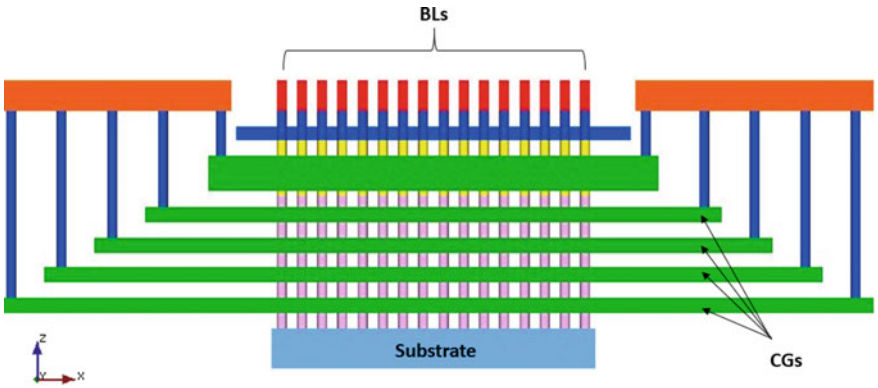

Fig. 4.27 X-Z cross section of Fig. 4.25

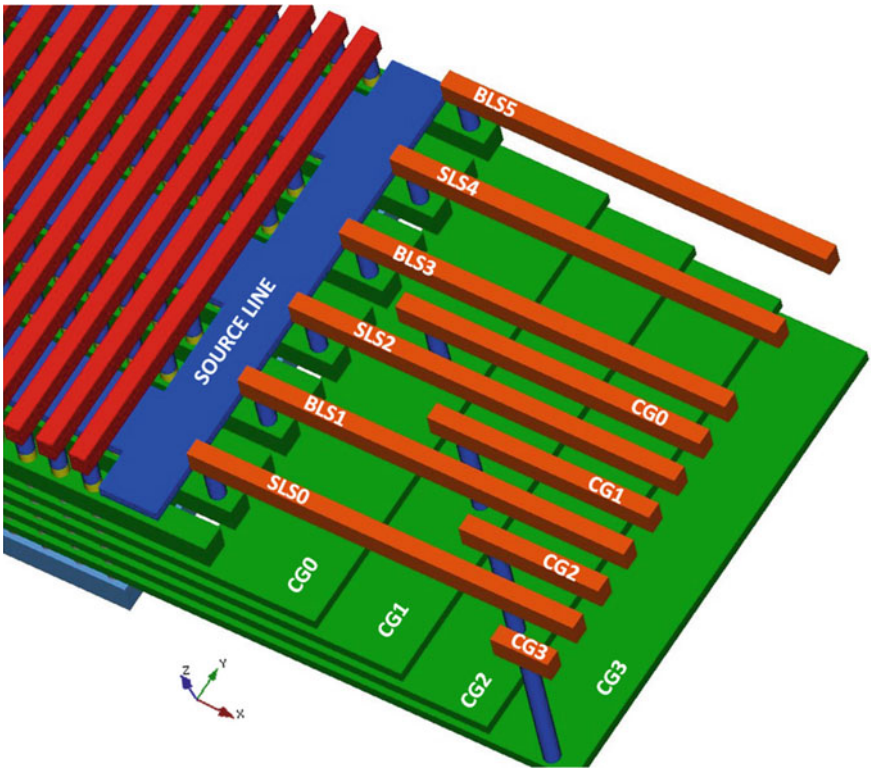

Fig. 4.28 Zoom of Fig. 4.25—top view

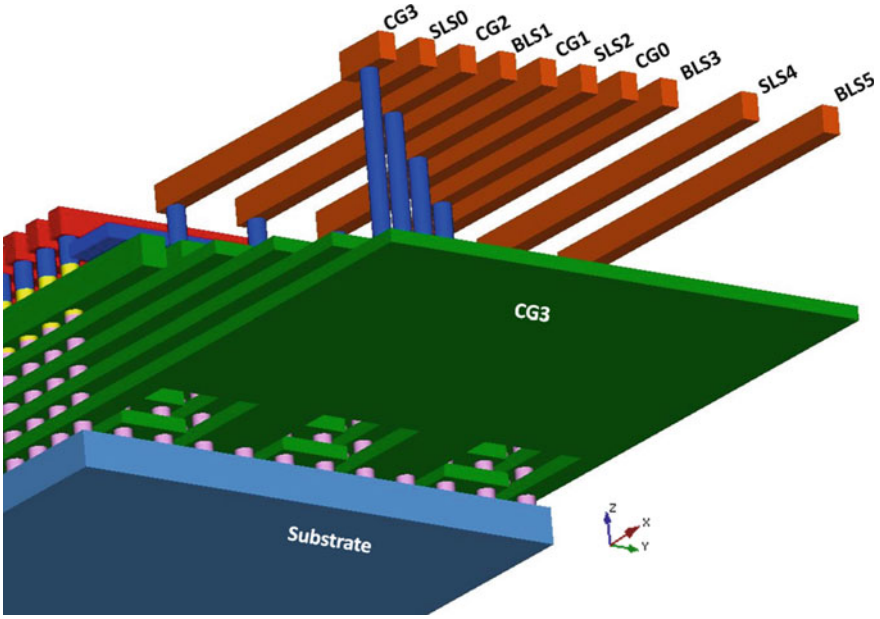

Fig. 4.29 Zoom of Fig. 4.25—bottom view

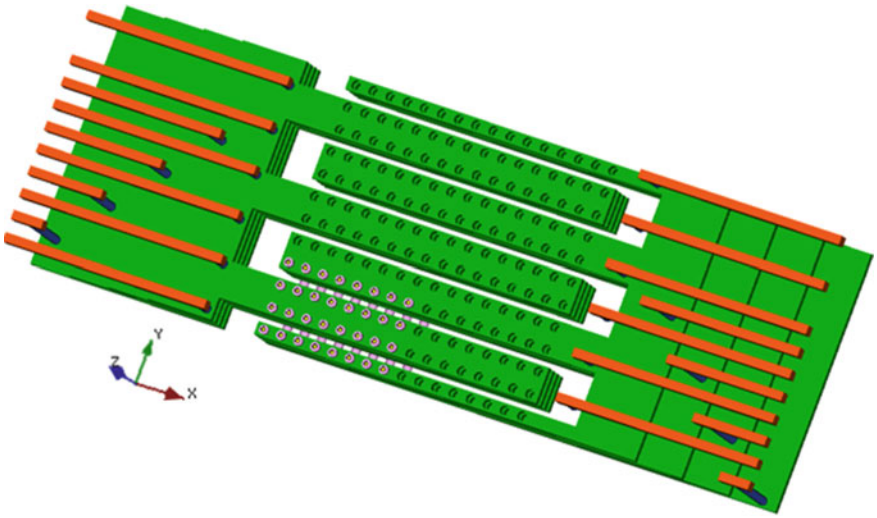

Fig. 4.30 P-BiCS: control gates are realized by fork-shaped plates

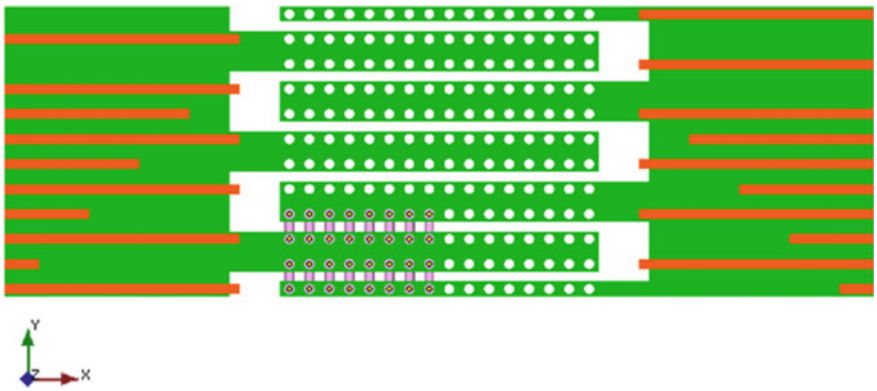

Fig. 4.31 Top view of Fig. 4.30

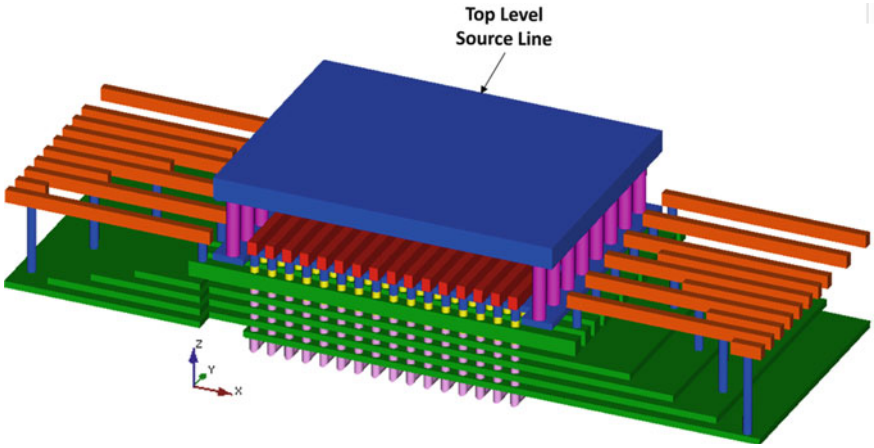

Fig. 4.32 Bird's-eye view of P-BiCS with top level source connection

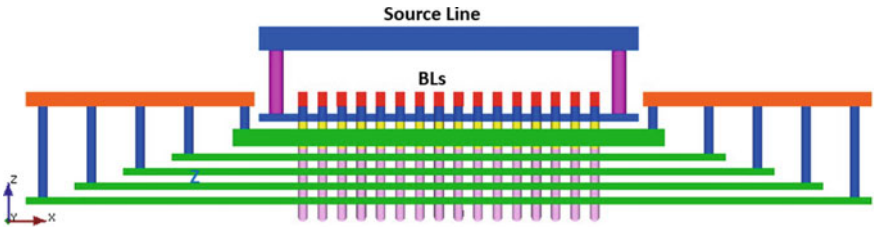

Fig. 4.33 X-Z cross section of Fig. 4.32

the one shown in Fig. 4.25. This additional layer is referred to as "Top Level Source Line" in Fig. 4.32. X-Z cross section and additional bird's-eye views are sketched in Figs. 4.33, 4.34 and 4.35.

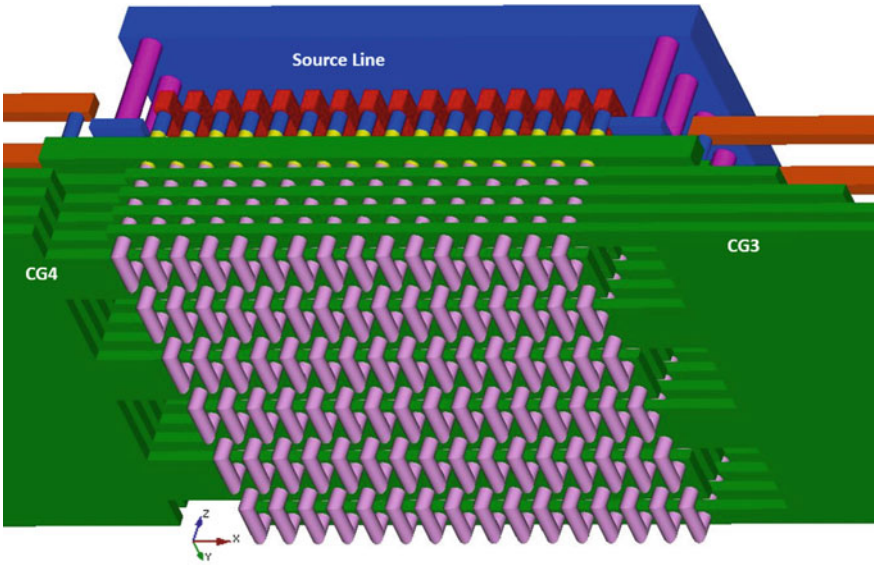

Fig. 4.34 Bottom view of Fig. 4.32

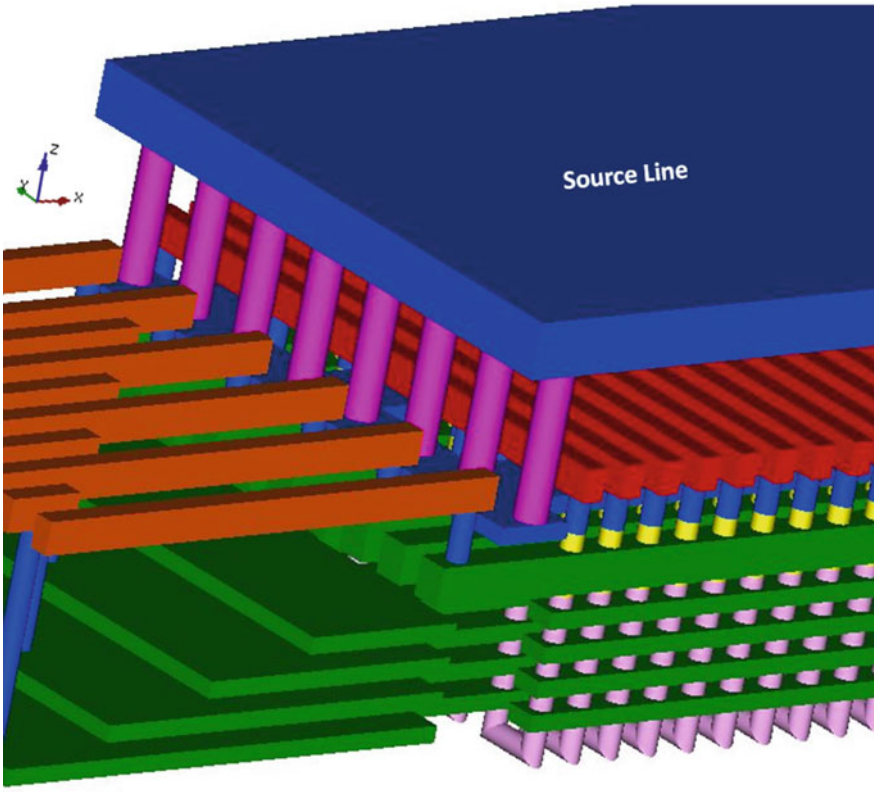

Fig. 4.35 Zoom of the left hand side of Fig. 4.32

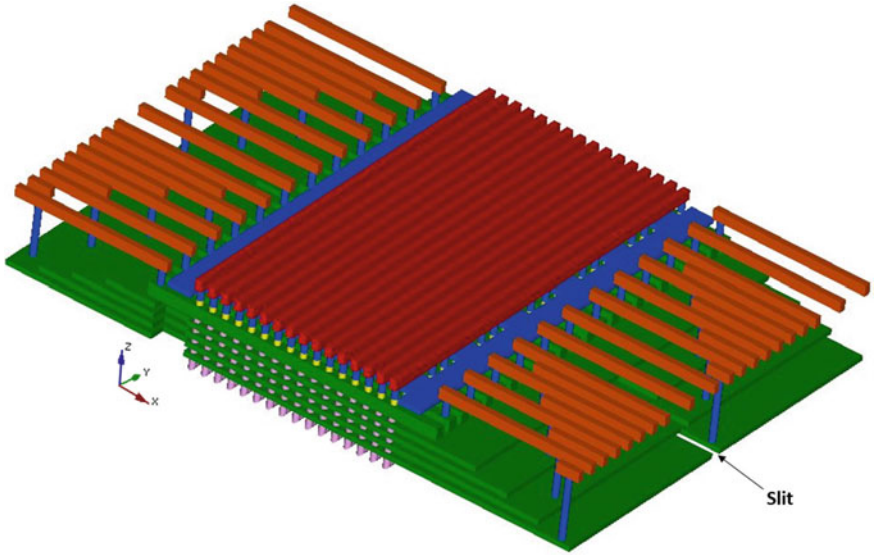

Fig. 4.36 P-BiCS array with slit

As we have seen with BiCS, it is important to regularly cut control gate plates by adding slits, in order to reduce program disturb and read disturb, as shown in Figs. 4.36 and 4.37.

Figure 4.38 is the top view of Fig. 4.36: it is clearly visible how the slit cuts the control gate plates in sub-plates. The corresponding Y-Z cross section is reported in Fig. 4.39.

4.4 VRAT and Z-VRAT

BiCS and P-BiCS were introduced by Toshiba. Samsung has followed a different path before reaching the architecture called V-NAND, which is the first 3D architecture with vertical channel brought to volume production. This section plus the following two describe the intermediate steps. Sect. 4.7 is devoted to V-NAND.

VRAT stands for *Vertical Recess Array Transistor* and was presented in 2008 [14]. Equivalent circuit schematic, vertical cross section and bird's-eye view of VRAT are shown in Fig. 4.40. Storage material is a nitride Charge Trap layer sandwiched, as usual, between tunnel and control oxides. 3D integration process for VRAT is called PIPE (*Planarized Integration on the same PlanE*) and the name has nothing to do with the pipe concept of P-BiCS. As this was one of the first attempts to build 3D structures, developing a simple unique integration scheme for vertical cells and interconnects was key; in other words, one of the most critical issues to

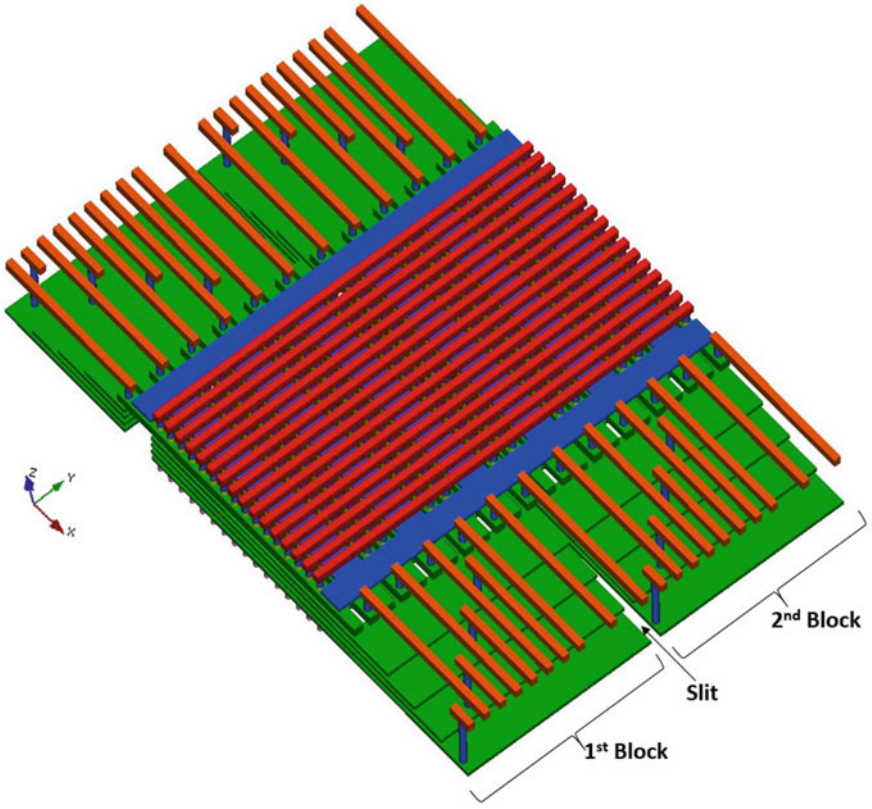

Fig. 4.37 P-BiCS array split in memory blocks

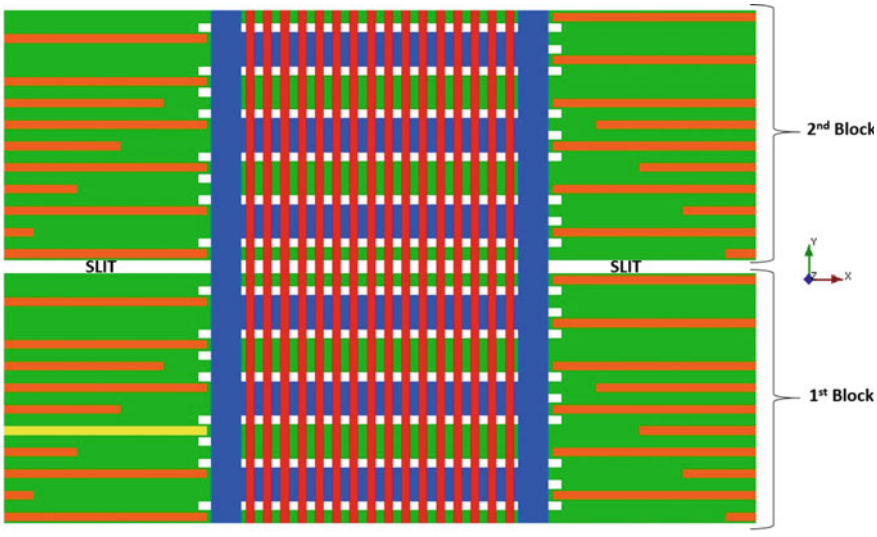

Fig. 4.38 Top view of Fig. 4.37

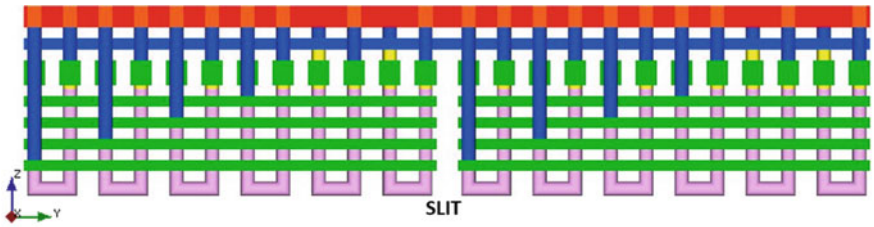

Fig. 4.39 Y-Z cross section of Fig. 4.37

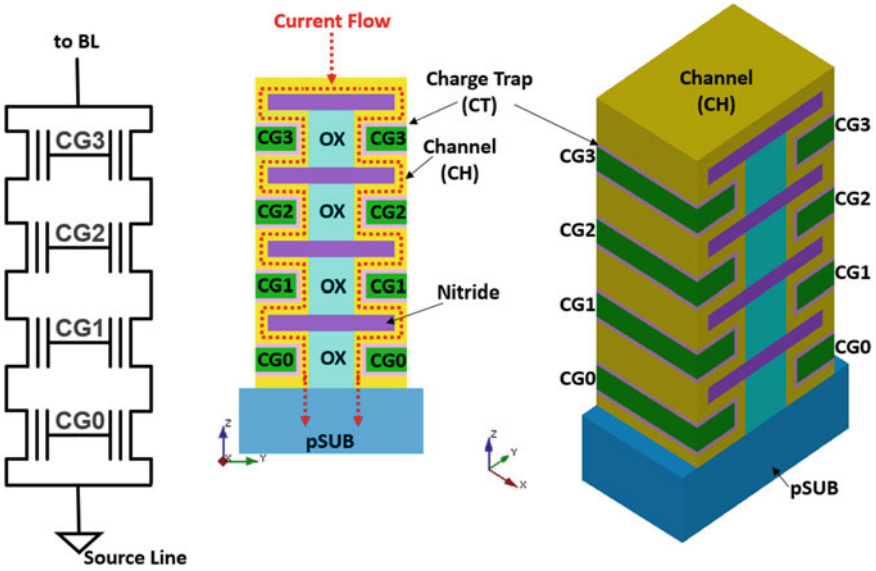

Fig. 4.40 VRAT NAND string

solve for 3D integration is how to connect memory cells from the matrix to the peripheral circuits.

Let's have a look at the main process steps for building VRAT structures. After deposition of multiple layers of nitride and oxide films, cell region is formed by wet etching process. Wordline electrode is deposited after cell's stack formation (oxides and CT material); at this point, an etch-back process removes the electrode from the sidewall, such that wordline layers are separated, as displayed in Fig. 4.40. At the end of the process, vertical strings are isolated (Fig. 4.41), and bitline and wordline contacts are added for connecting the array to the peripheral circuits (Fig. 4.42). Please note that wordline fan-out is not based on a stair-like structure, which can be seen as one of the merits of this solution; in fact, it doesn't require additional lithography steps (for staircase formation) and it is, therefore, cheaper (at least, in principle).

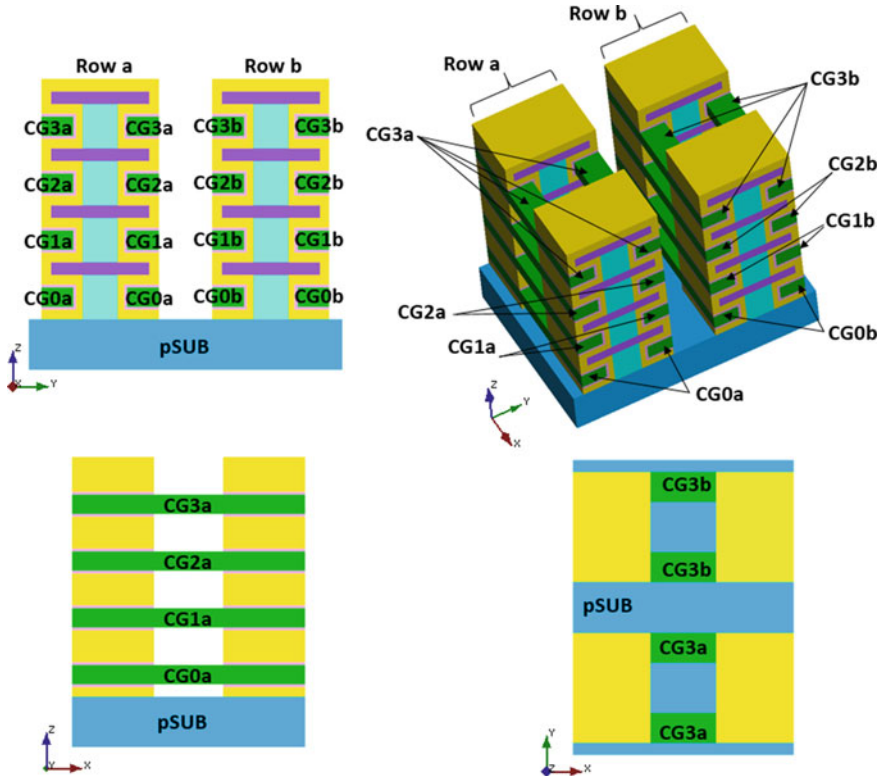

Fig. 4.41 Array of VRAT strings

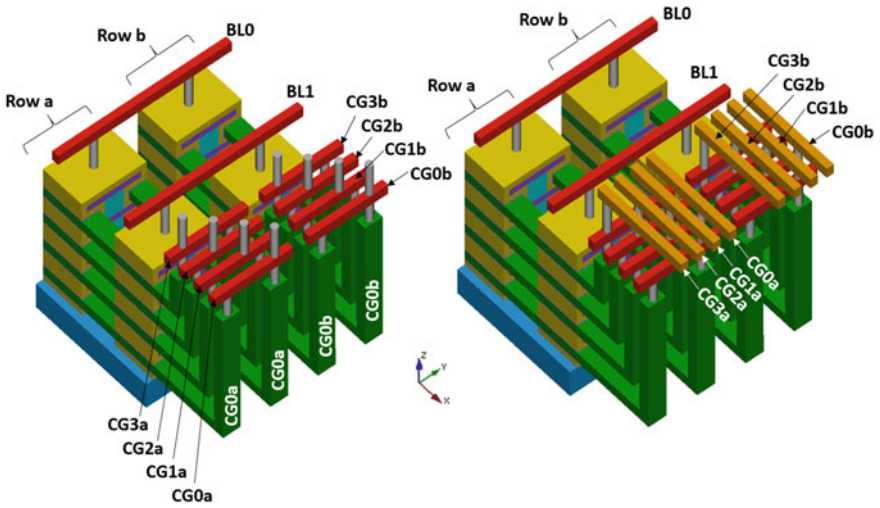

Fig. 4.42 VRAT array with BL and WL connections

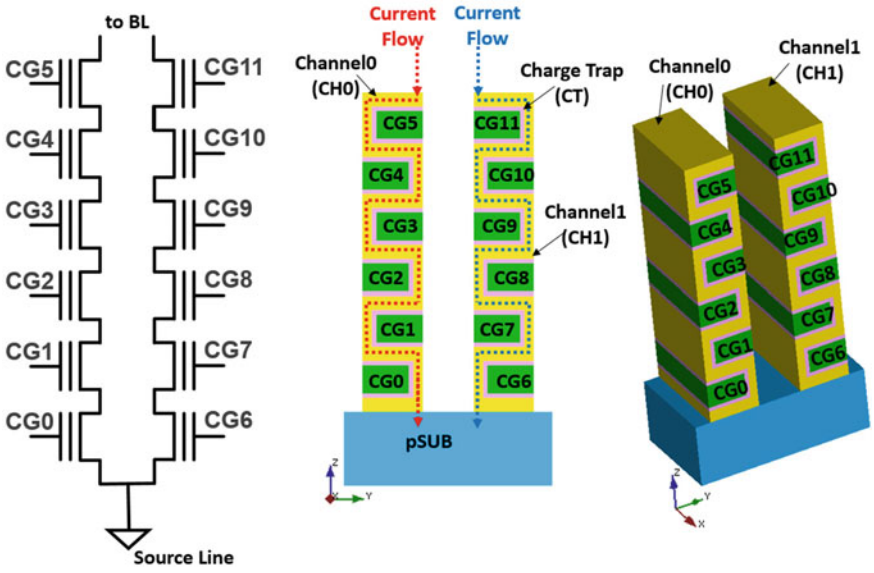

Fig. 4.43 Z-VRAT NAND strings

As a matter of fact, VRAT is a twin structure in the sense that current flows through 2 different paths because of the symmetry of the structure. Z-VRAT (*Zigzag* VRAT) is a further step towards a higher bit density [14]. In fact, each VRAT vertical structure is split into 2 narrower strings, as sketched in Figs. 4.43 and 4.44. The concept of the PIPE integration scheme stays intact; of course, the separation of the strings requires some additional process steps.

VRAT evolved in a different structure called VSAT, which is analyzed in the following section.

4.5 VSAT and A-VSAT

VRAT manufacturing is based on 2 basic concepts: gate-last and channel-first [15]. The challenge of this approach is in the undercut space for gate electrodes, which needs to be created and filled, thus requiring complex fabrication steps. VSAT (*Vertical Stacked Array Transistor*) revers the order: gate-first and channel-last. The basic building block of VSAT is displayed in Fig. 4.45. Let's look at the fabrication process. Multiple layers of polysilicon and nitride films are deposited one on top of each other; polysilicon acts as a gate, while nitride is the isolation material. The active region is defined after patterning the multiple layers, and a subsequent etching process. All gate electrodes are exposed on the same plane after a CMP process, thus allowing easy access to gate electrodes. Tunnel oxide, CT material,

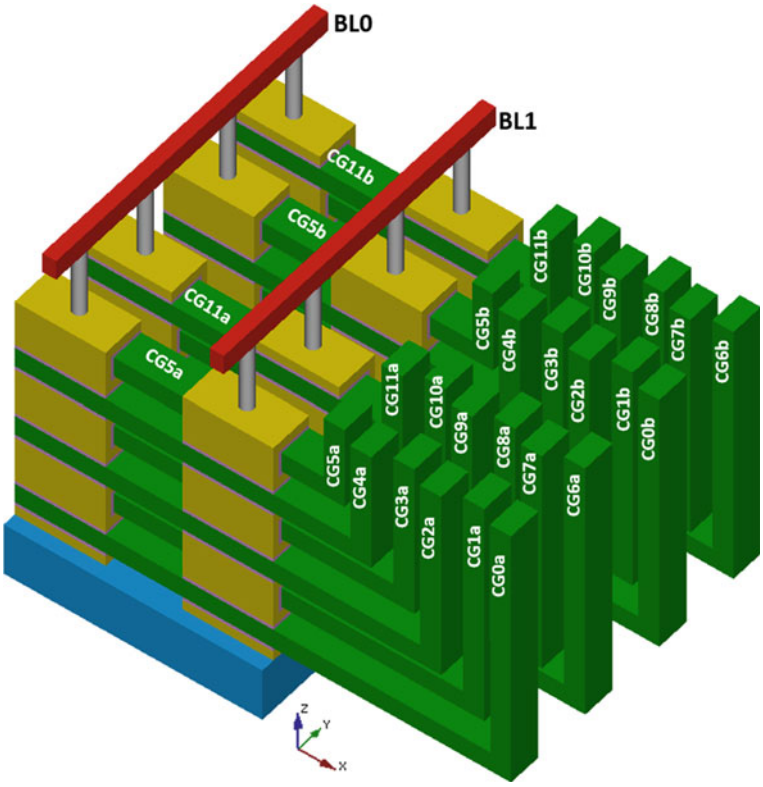

Fig. 4.44 Z-VRAT array with BL and WL connections

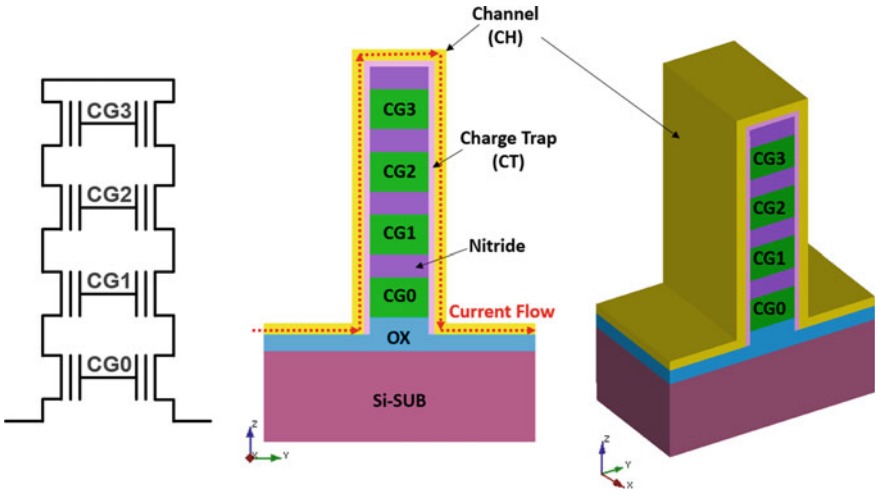

Fig. 4.45 VSAT basic building block

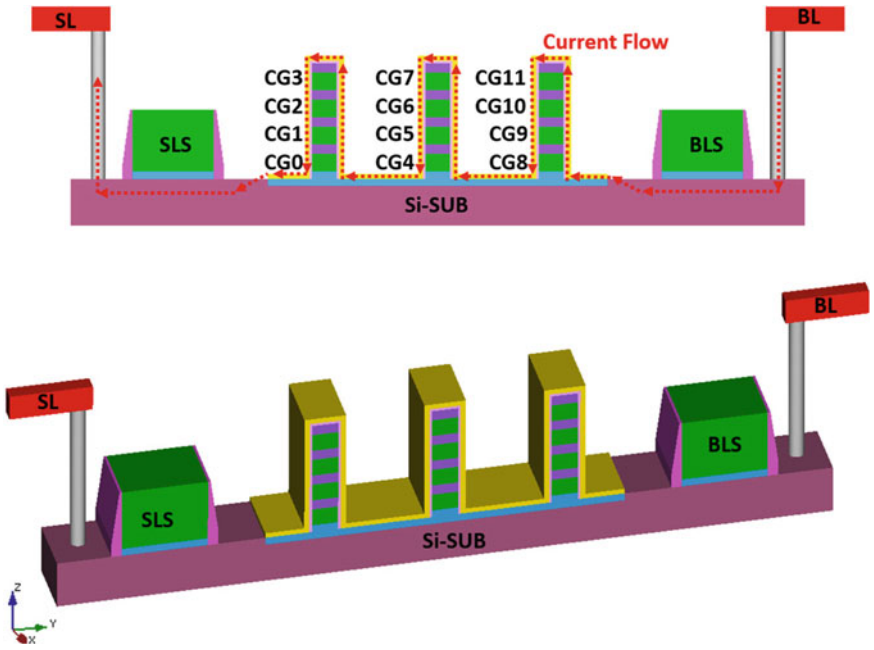

Fig. 4.46 VSAT NAND string

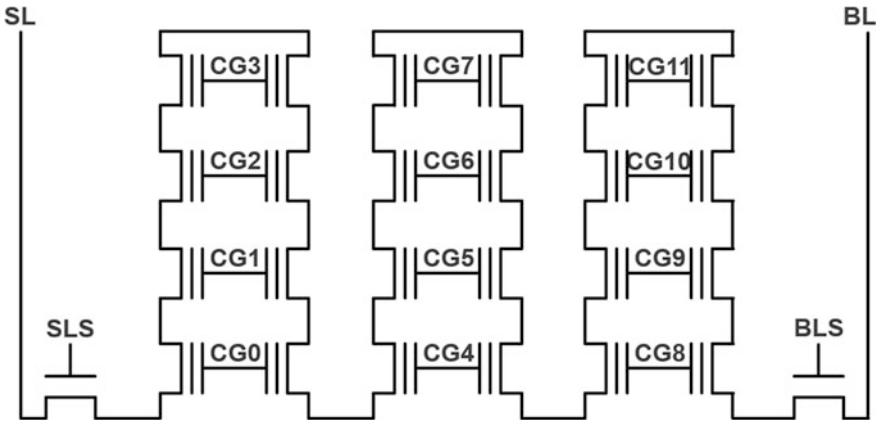

Fig. 4.47 Equivalent circuit of Fig. 4.46

and blocking oxide are then deposited on the active region. Channel is formed by a polysilicon deposition. Of course, final step is an etching process which is used to isolate vertical strings.

Multiple Fig. 4.45 can be combined together to form a NAND string, as shown in Figs. 4.46 and 4.47.

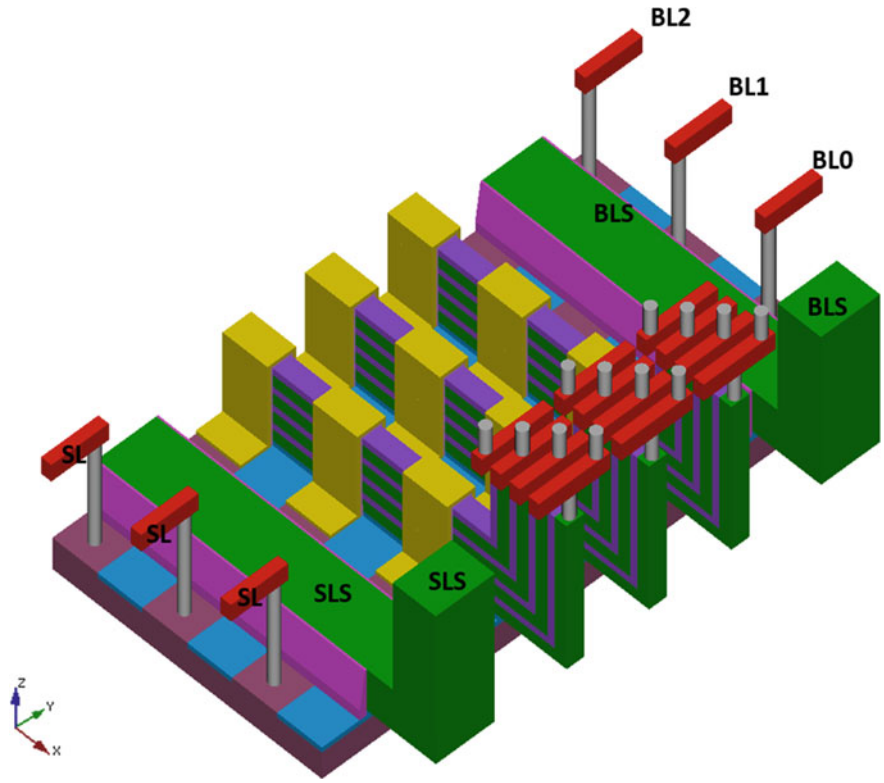

Fig. 4.48 VSAT fan-out

Vertical channel architectures usually have SL selectors at the bottom of the 3D stack; as we have seen in previous sections, this poses some challenges as SLS might have different gate oxide, gate length, and doping density than the memory cells along the string. Figure 4.46 shows that, in VSAT, SL selectors are not part of the vertical string as they are located in the peripheral region, like in the 2D case, thus simplifying the whole manufacturing process. All the merits of the PIPE process in terms of fan-out remain unchanged, as clearly visible in Fig. 4.48.

By comparing the schematic of VRAT and VSAT, we can notice that they are both "twin" cells in the sense that they are split in 2 parts. The difference is that in VRAT the 2 halves are in parallel, while they are in series in VSAT. In other words, this is like saying that in VSAT the current flows through the "same" cell twice, thus degrading the bit density. To solve this problem, an additional wordline cut process was proposed by Macronix in 2015 [16]. This improved vertical architecture is called A-VSAT, which stands for *Asymmetrical VSAT*: basic building block and A-VSAT NAND string are shown in Figs. 4.49 and 4.50, respectively.

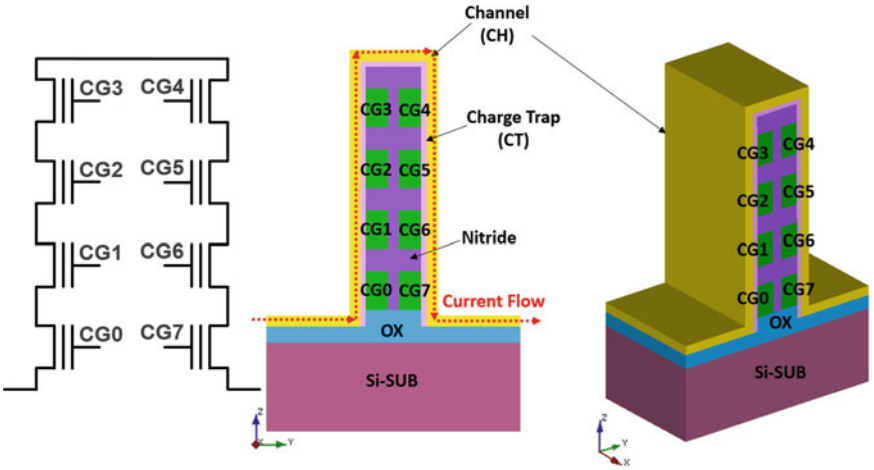

Fig. 4.49 A-VSAT basic building block

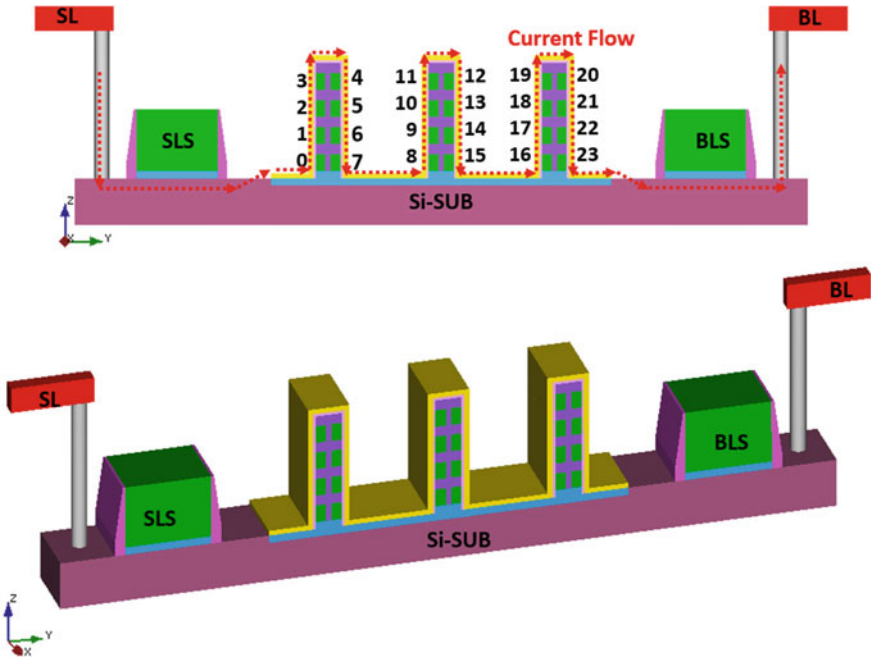

Fig. 4.50 A-VSAT NAND string

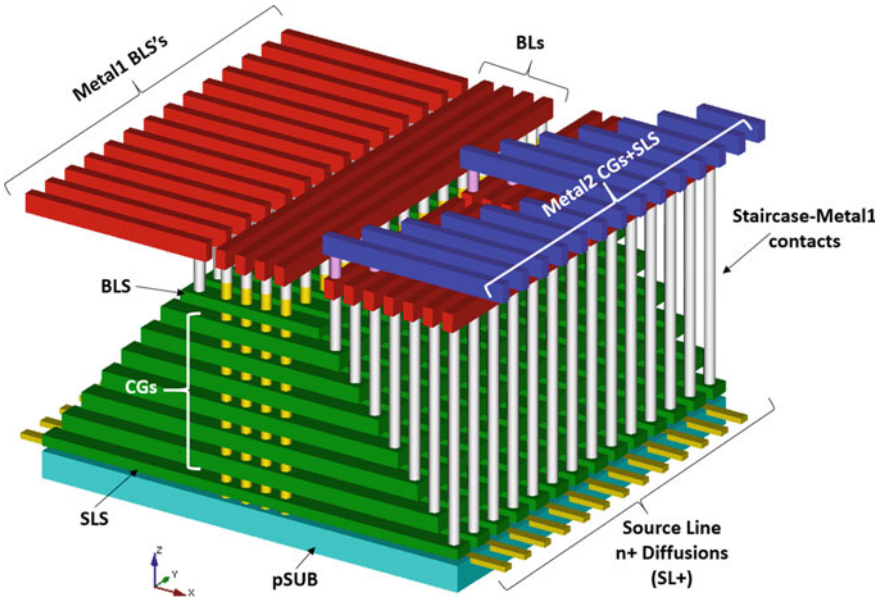

Fig. 4.51 TCAT NAND flash array

4.6 TCAT

TCAT (Terabit Cell Array Transistor) was proposed in 2009 [17]. A bird's-eye view of the TCAT array is presented in Fig. 4.51. The equivalent circuit is sketched in Figs. 4.52 and 4.53. Besides SL+ lines, the TCAT equivalent circuit is identical to the BiCS equivalent circuit shown in Fig. 4.4. All SL+ lines (which are n+ diffusions) are shorted together outside the array to form a common Source Line. Two level of metals are used to fan-out BLS's and CGs+SLS, respectively. Top and side views of TCAT can help understanding better the overall architecture (Figs. 4.54, 4.55, and 4.56).

There are 6 CG layers and 2 NAND blocks. Each block has 7 wordlines per layer. These 7 wordlines are shorted together at the Metal1 level, as displayed in Fig. 4.57. In order to clearly understand how many contacts are underneath the Metal1 layer, another layer removal is needed, as shown in Fig. 4.58. Metal2 is used for wordline decoding, while Metal1 is used to decode the NAND string. It is worth highlighting that, compared to BiCS, in the TCAT architecture the slit is a cut in the bitline layer, as it can be seen in Fig. 4.57; in fact, wordlines are already separated by construction (Fig. 4.59).

After the review of the architectural aspects of TCAT, we can now look at the differences with respect to BiCS. First of all, TCAT makes use of *gate-replacement* [17]; in other words, the gate layer is deposited only at the end of the stack formation, while BiCS is based on a gate-first process. Let's briefly analyze the

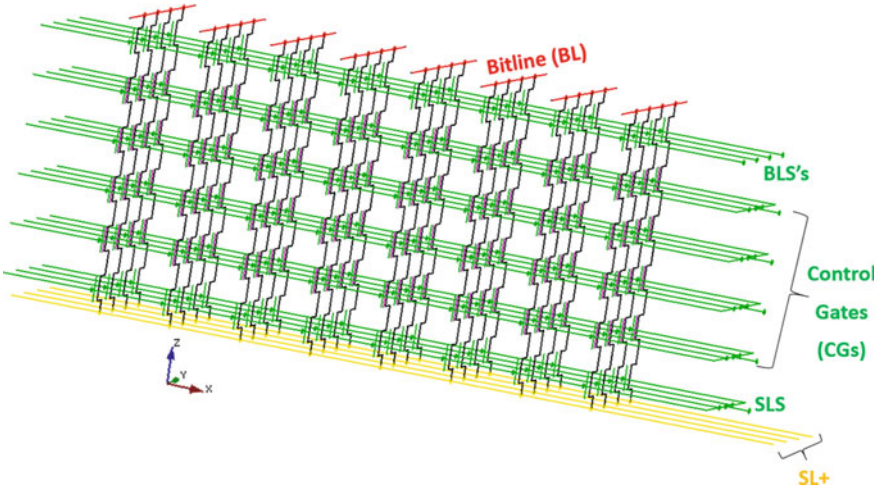

Fig. 4.52 TCAT circuit schematic

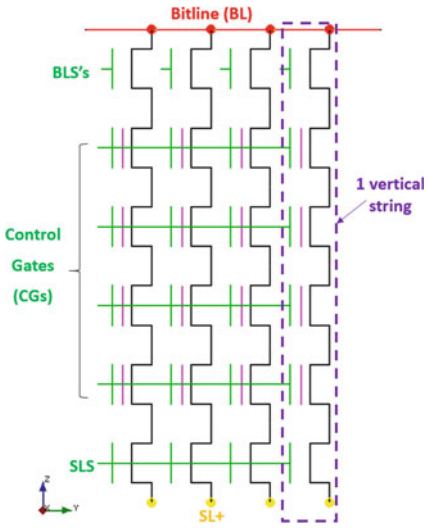

Fig. 4.53 Y-Z view of TCAT circuit schematic

gate-replacement technology. Multiple layers of oxide and sacrificial nitride layers are initially deposited. The whole stack is then etched between each row of pillars (holes) and nitride is removed. At this point, gate dielectric layers, and metal gates are deposited in the conventional order by filling the space between wordlines with tungsten. After that, an etching process is used to separate the gates. In this way, a metal gate SONOS memory cell is fabricated; this kind of cells allows faster erase speed, longer retention, and a wider V_{TH} window. Of course, the other big advantage of the metal gate is the reduced parasitic resistance of the wordline, which translates in faster operations.

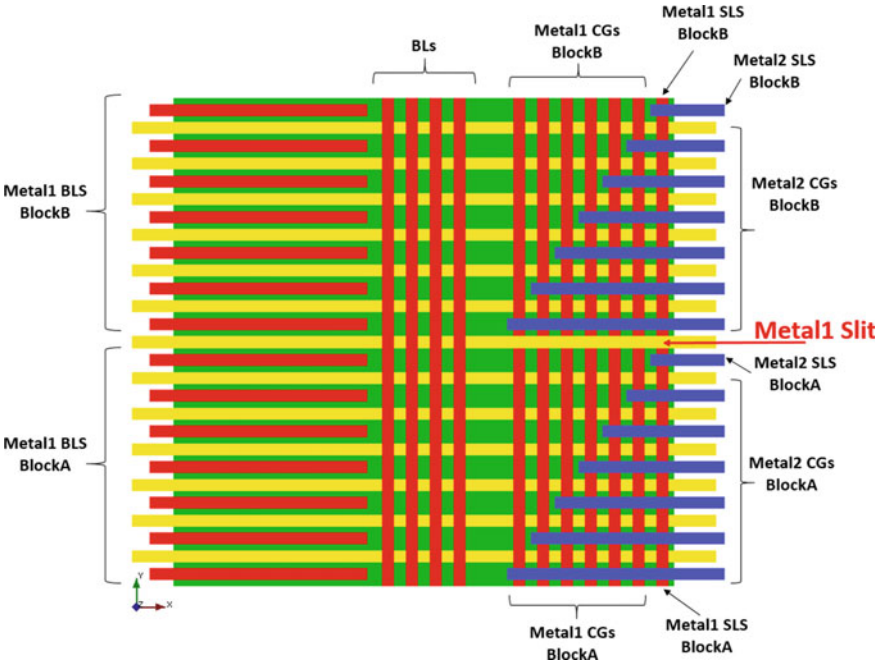

Fig. 4.54 Top view of Fig. 4.51

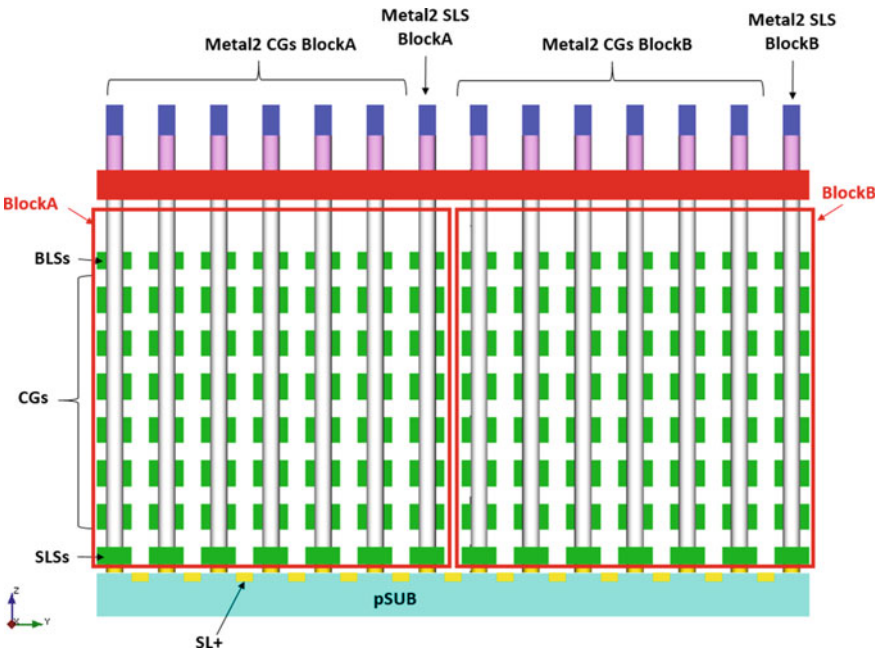

Fig. 4.55 Z-Y side view of Fig. 4.51

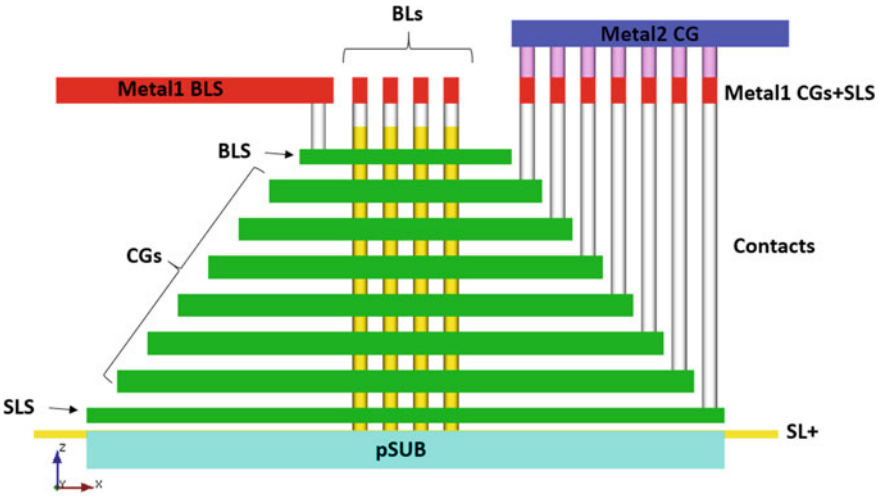

Fig. 4.56 Z-X side view of Fig. 4.51

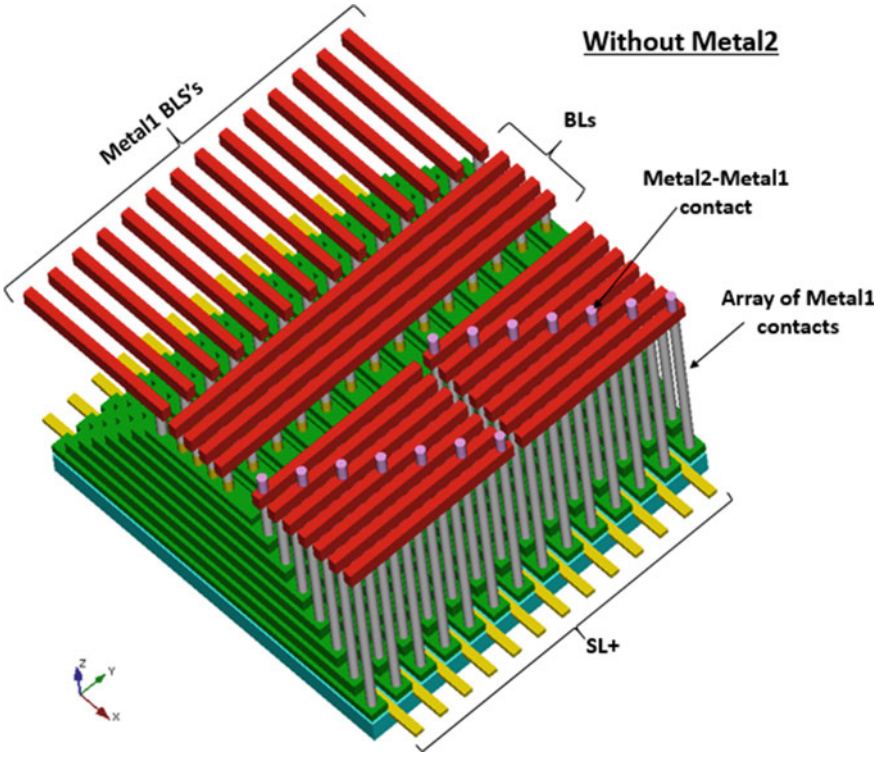

Fig. 4.57 TCAT NAND array without Metal2 layer

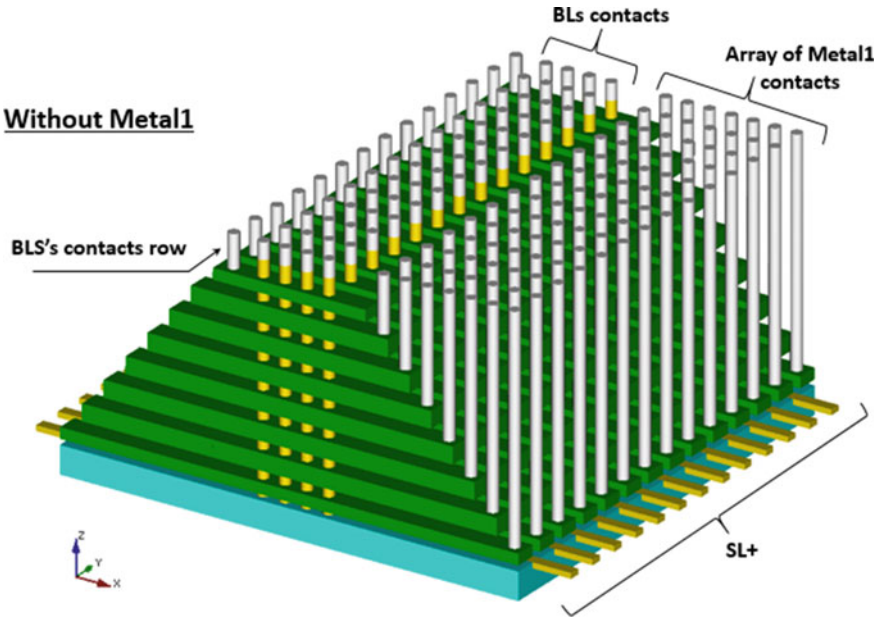

Fig. 4.58 TCAT NAND array without metal layers

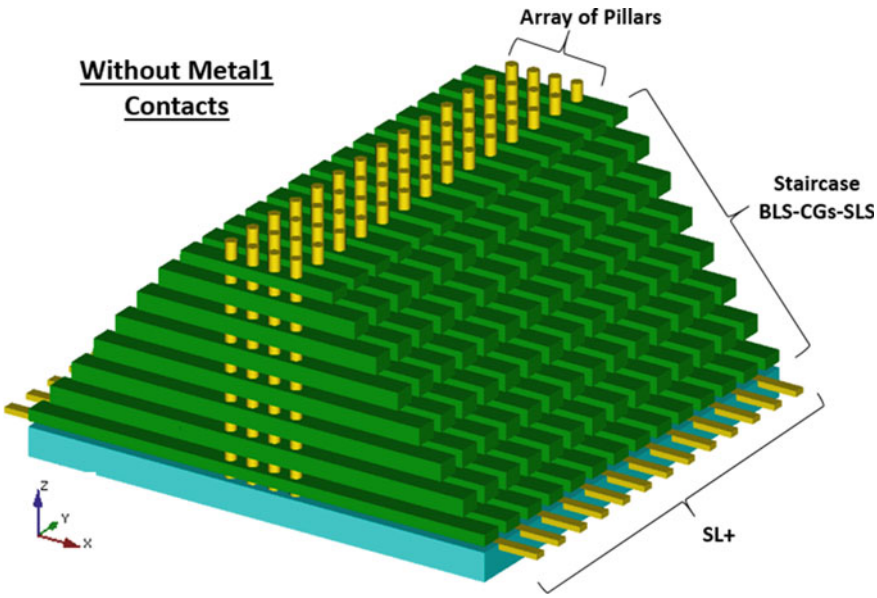

Fig. 4.59 TCAT NAND array without metal layers and bitline contacts

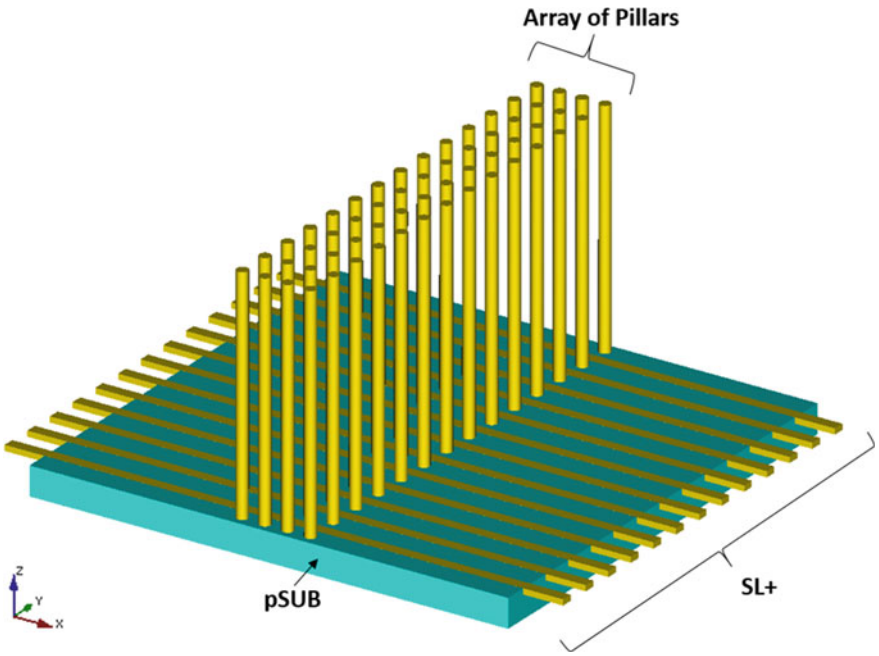

Fig. 4.60 TCAT pillars

Another difference is the bulk erase operation. As shown in Fig. 4.60, the vertical pillar is directly connected to the p-substrate and not to a n+ diffusion. Close to each NAND string there is a n+ region used to drain the cell's current. During erase, holes are directly provided by the substrate, without the need for GIDL (*Gate Induced Drain leakage*) generation at SLS, which is one of the main concerns of the BiCS architecture.

Last but not least, we can take a look at the NAND memory cell: a direct comparison between BiCS and TCAT is reported in Fig. 4.61. Because of the gate-last process adopted by TCAT, the charge trap layer is biconcave, which results in a reduced charge spreading effect. In fact, in a string of the conventional 3D BiCS, the charge trap nitride layer is continuously connected across all gates along the channel side and, as a matter of fact, it acts as a charge spreading path. As discussed in Chap. 2, this causes degradation of data retention characteristics. Thanks to the biconcave shape of the charge trap layer, in the TCAT case it is harder for electrical charges to move from one cell to another, as there is not straight connection of the charge trap layers between them [18].

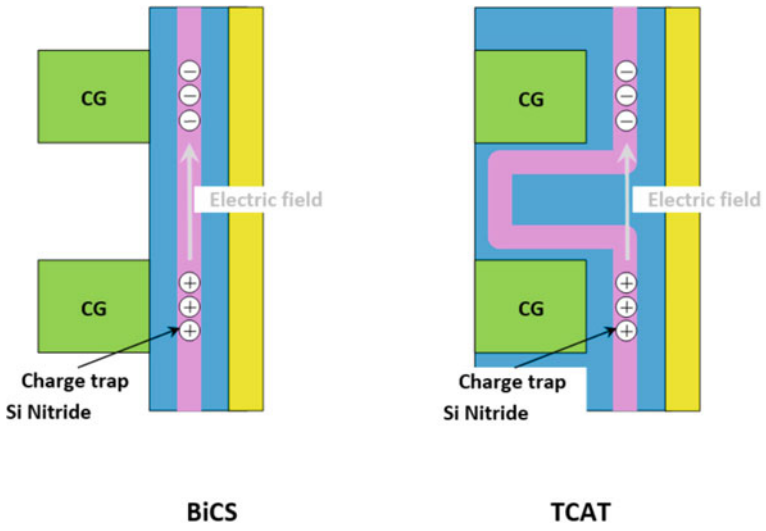

Fig. 4.61 BiCS and TCAT flash memory cells

4.7 V-NAND

TCAT turned out to be the foundation for the first 3D NAND architecture brought to volume production, which is called V-NAND. This technology was launched at Flash Memory Summit 2013 [19]. As summarized in Fig. 4.62, Samsung officially introduced the first generation of V-NAND in 2014 [20–22]. The first product was a 128 Gb MLC NAND, based on damascened metal-gate SONOS-type (CT) cell. The vertical stack of V-NAND Gen1 is made of 24 wordline layers, whose equivalent circuit is shown in Figs. 4.63 and 4.64. In terms of schematic diagram, the main difference with respect to Figs. 4.4 and 4.53 is the presence of dummy wordline layers (dummy CG).

In 3D architectures with vertical channel, memory cells have floating body. During programming, high channel boosting can generate hot carries at the edge of the string, because of the high lateral electric field. As such, the channel potential does not boost up as it should when WL0 is programmed (i.e. hot carriers discharge the channel). Of course, this behavior translates into Program Disturb. Therefore, dummy WLs are inserted between the selection transistors and WLs to prevent the above mentioned effect [23, 24].

To save silicon area, V-NAND Gen1 adopts a special layout for the array of pillars, known as *Staggered Pillars* (or Holes). In practice, even and odd rows of pillars are staggered, without changing the center-to-center distance between adjacent pillars. With reference to the left hand side of Fig. 4.65, please note that in Gen1 each bitline has to fit in 1 channel hole (pillar) pitch.

In V-NAND Gen2 layout of bitlines is different, as sketched on the right hand side of Fig. 4.65: we refer to this layout as *Staggered Bitline Contacts*. In this case,

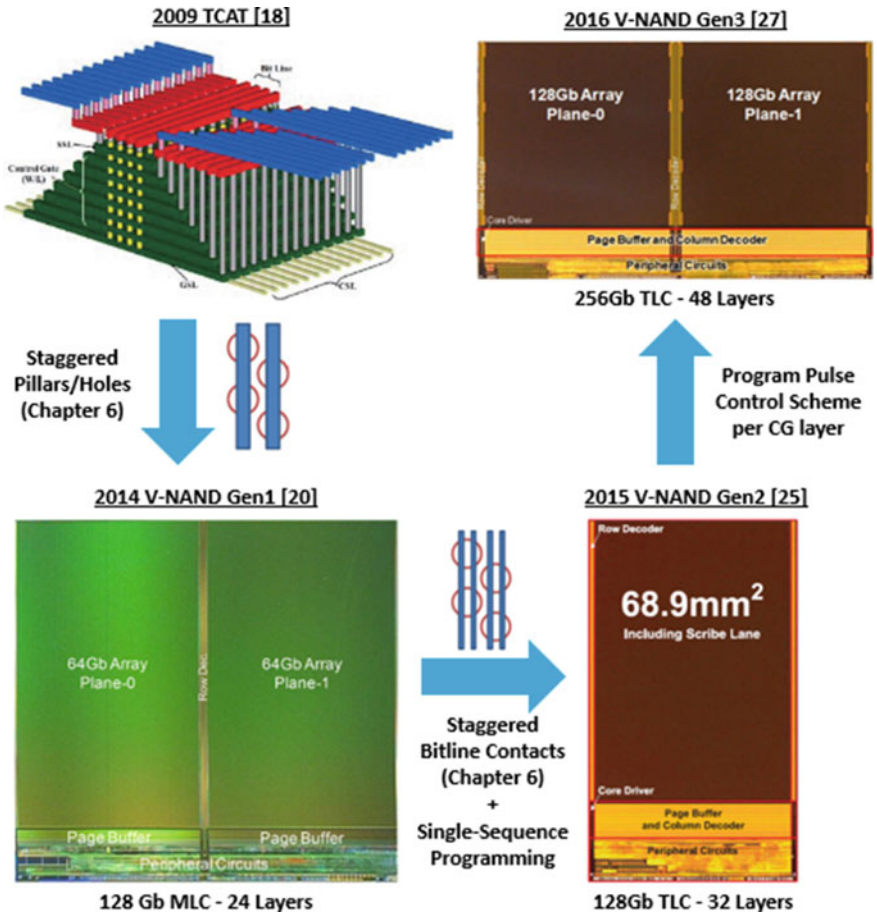

Fig. 4.62 From TCAT to V-NAND (pictures not in scale) [17, 20, 25, 27]

2 bitlines are arranged in 1 pillar pitch [25]. Of course, BL density doubles (NAND page size goes from 8 to 16 kB), but the number of contacts to the bottom Source Line plate is halved. The overall number of pillars is the same of the previous generation. Figure 4.66 displays a side by side comparison of the array cross sections of the 2 generations (not in scale).

The reader can find a detailed explanation of the above mentioned layout techniques in Chap. 6, which is devoted to the analysis of the most advanced 3D architectures for vertical channel Flash arrays.

V-NAND Gen2 was presented in 2015 [25, 26] in the form of 128 Gb TLC. Compared to the previous generation, there hasn't been macroscopic changes in the memory cell itself, but it is worth mentioning that the number of control gate layers is 32 instead of the previous 24.

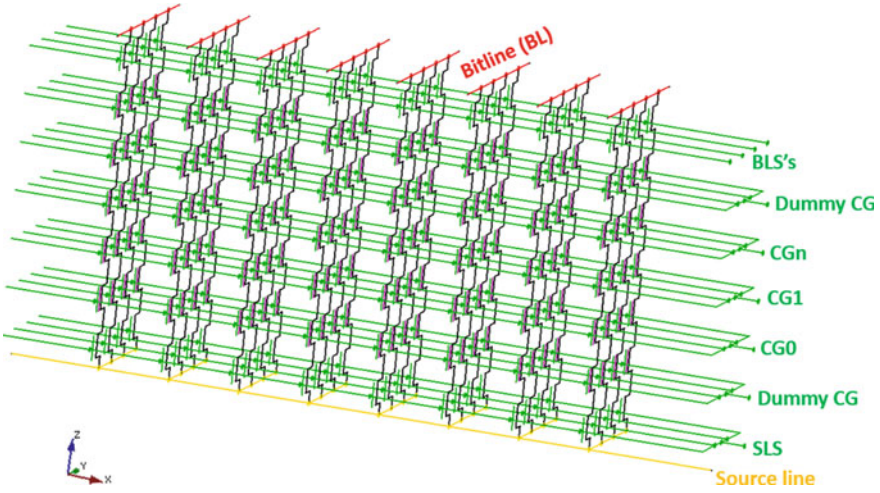

Fig. 4.63 Circuit schematic of a 3D V-NAND array

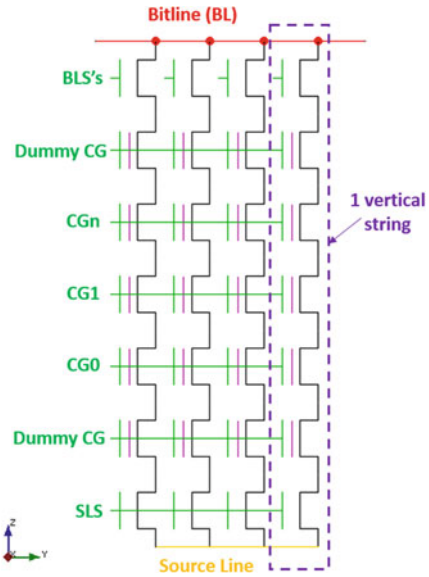

Fig. 4.64 Y-Z view of V-NAND circuit schematic

The other highlight of this device is the Single-Sequence Programming operation. TLC functionality (i.e. 3 bit/cell) requires the ability of squeezing 8 V_{TH} distributions instead of 4, within almost the same V_{TH} window. In planar NAND, conventional TLC programming algorithms are based on repeating the programming sequence 3 times per wordline: V_{TH} distributions become smaller and smaller after each sequence. In V-NAND Gen2, tight cell's V_{TH} distribution is achieved by

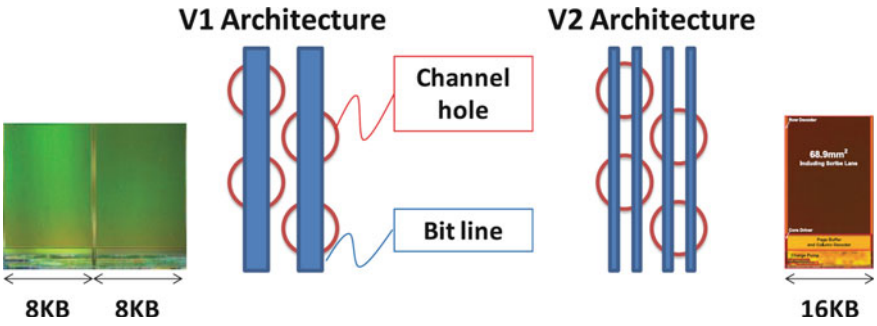

Fig. 4.65 First and second generations of V-NAND bitline architectures

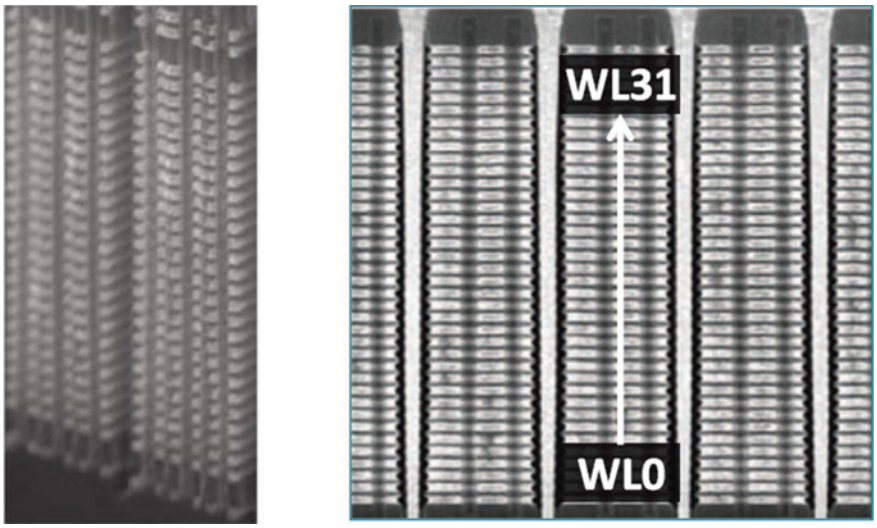

Fig. 4.66 Cross section of V-NAND Gen1 (*left*) and Gen2 (*right*), not in scale

exploiting the advantages of CT, i.e. small cell-to-cell interference and narrow natural V_{TH} distribution. This is coupled with an improved programming algorithm: V-NAND Gen2 asks for the 3 TLC pages of data at the start of programming and it writes the 3 pages at once. Of course, this approach translates into faster programming operations and lower power consumption.

V-NAND Gen3 becomes public at the *IEEE International Solid State Circuits Conference* (ISSCC) in 2016 [27]. It is again a TLC device, but this time it is a 256 Gb based on a vertical stack of 48 control gate layers. When increasing the number of layers, the etching technology can become a serious issue because of the aspect ratio of the pillar. Therefore, it is almost unavoidable to reduce the thickness of each single control gate layer. While the overall stack manufacturability benefits from thinner layers, the situation is totally different for wordlines: in fact, parasitic

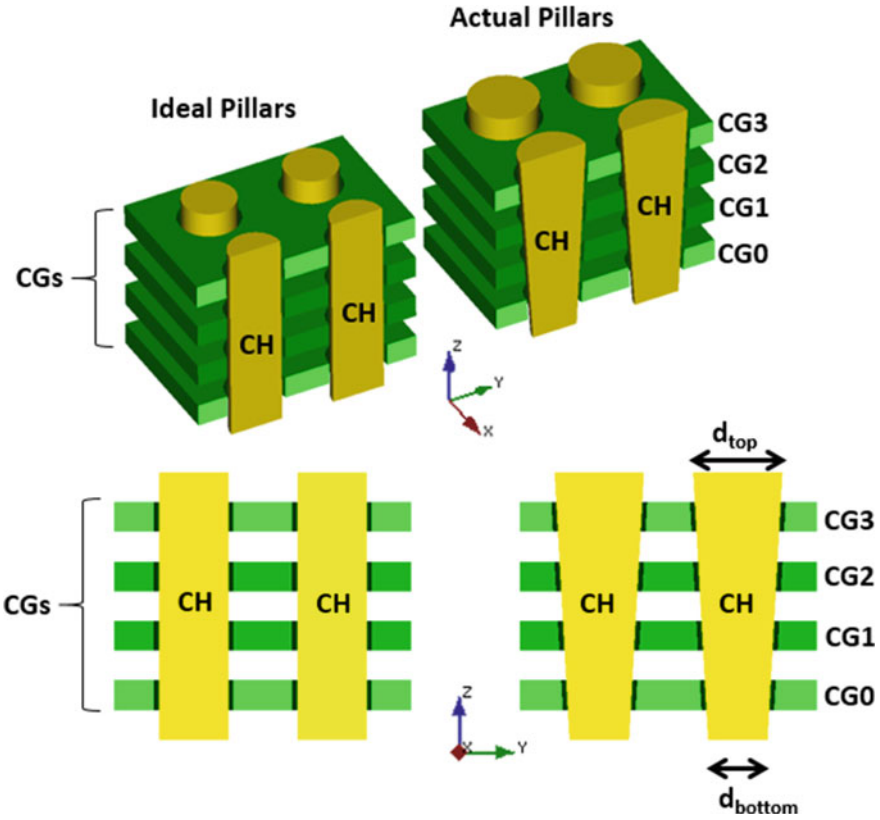

Fig. 4.67 Ideal versus actual shape of pillars

resistance and capacitance increase, thus slowing down both read and program operations. Not only that. Because of the increased resistance, channel hole size fluctuations become even more important. As a matter of fact, channel hole can be seen as a physical barrier to the charge flow along the wordline: a variation of the hole size translates into a variation of the parasitic resistance. This fact needs to be coupled with the actual shape of the pillar (Chap. 2), which is sketched in Fig. 4.67. As a result, the voltage transient is a function of the layer. By using the conventional approach of applying the same program step duration to each single wordline (Chap. 3), the NAND raw BER would become higher for the slower wordlines. As such, an adaptive program pulse scheme is adopted: basically, it varies the program pulse duration according to the target WL's characteristics.

As the number of layers in the 3D NAND stack grows up, we expect these non-uniformity effects to become more and more the focus for circuit designers and process technology engineers.

At this point the reader should have all the architectural elements to understand the challenges of stacking several memory layers, one on top of each other.

Even though BiCS and V-NAND with Charge Trap storage have been a major breakthrough in going to 3D, they are not the only possible approach, and Floating Gate is still alive. The 3D world is full of options! In the next chapter we'll see many different types of 3D Flash memory cells based on Floating Gate.

References

1. http://www.samsung.com/us/business/oem-solutions/pdfs/V-NAND_technology_WP.pdf. Samsung V-NAND technology, White Paper, Sept 2014
2. 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)
3. 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, Berlin, 2011)
4. 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
5. 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
6. 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
7. T. Maeda et al., Multi-stacked 1G cell/layer pipe-shaped BiCS flash memory, in *Digest Symposium on VLSI Circuits*, June 2009, pp. 22–23
8. 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 *2009 Symposium on VLSI Technology* (2009), pp. 136–137
9. 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
10. Y. Yanagihara et al., Control gate length, spacing and stacked layers number design for 3D-Stackable NAND flash memory, in *IEEE IMW* (2012), pp. 84–87
11. 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
12. A. Nitayama et al., Bit Cost Scalable (BiCS) flash technology for future ultra high density storage devices, in *2010 International Symposium on VLSI Technology Systems and Applications (VLSI TSA),* Apr. 2010, pp. 130–131
13. 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
14. 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 *2008 IEEE Symposium on VLSI Technology* (2008)
15. 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 *2009 IEEE Symposium on VLSI Technology* (2009)
16. Y.-H. Hsiao, Ultra-high bit density 3D NAND flash-featuring-assisted gate operation. IEEE Elect. Dev. Lett. **36**(10), 1015–1017 (2015)
17. J. Jang et al., Vertical cell array using TCAT (terabit cell array transistor) technology for ultra high density NAND flash memory, in *2009 IEEE Symposium on VLSI Technology* (2009)

18. W. Cho et al., Highly reliable vertical NAND technology with biconcave shaped storage layer and leakage controllable offset structure, in *2010 Symposium on VLSI Technology (VLSIT)* (2010), pp. 173–174
19. J. Elliott, E.S. Jung, Ushering in the 3D memory era with V-NAND, in *Proceedings of Flash Memory Summit*, www.flashmemorysummit.com, Santa Clara, CA, Aug 2013
20. 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*, Feb 2014, pp. 334–335
21. 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)
22. K.T. Park, A world's first product of three-dimensional vertical NAND flash memory and beyond, in *NVMTS*, 27–29 Oct 2014
23. 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
24. 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
25. 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*, Feb 2015, pp. 130–131
26. J.-W. Im, 128 Gb 3b/cell V-NAND flash memory with 1 Gb/s I/O rate. J. Solid-State Circ. **51**(1) (2016)
27. 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*, Feb 2016, pp. 130–131