Abstract—Nand flash memory is an important part of the electronics products. Over the years scientists and engineers have worked hard to make it faster and cost effective. Recently, there has been a realization that current technologies is not capable of further increasing their capacity and at the same time keep the cost down. Keeping in view this limit and the ever increasing need for Nand flash memory scientists and engineers have proposed 3D integration of Nand flash memory. 3D integration promises to be an excellent replacement of current technologies for the development of Nand flash memory. This paper surveys the 3D integration technologies and does a critical comparison among each of them. In the end their advantages and limitations have been discussed.

Index Terms—Nand flash, 3D integration, vertical channel, vertical gate.

I. INTRODUCTION

Over the years the Nand flash has become an integral part of human life. Everyday consumer electronics are heavily dependent on flash memory to satisfy their memory requirements. Nand flash memory is cheap, compact and cost effective. It comes in different flavors—solid state drives, flash pen drives, mobile phones memory chips, etc.

There has been great advancement in the fabrication technology of Nand flash but recently it has been realized that further advancement to increase capacity is not without increasing the cost per bit and degrading the performance. Recently, a Nand chip of 128 GB using 19nm CMOS has been reported and some believe that one can reach sub 10nm size to fabricate denser memory but achieving a capacity of the order of terabytes seems distant. The current technology is capable of fabricating planar Nand memory (or 2D Nand) and the scaling down of the memory cells is limited to the X or Y direction. What about Z-direction? Some scientists have contemplated this question and recently have successfully come up with fabrication techniques by which the planar Nand can be stacked or integrated on top of each other. This is also termed as 3D integration. 3D integration relieves the burden of planar/lateral scaling down of Nand flash. Instead, with 3D integration one can add multiple layers of planar Nand on top of each other and integrate them to form a unit of memory. A number of novel fabrication techniques have been proposed to achieve this form of integration.

The following sections discuss the factors that limit the scaling down of planar Nand flash, various 3D integration techniques. 3D fabrication technologies have been discussed and compared in the following sections.

II. PLANAR NAND FLASH AND THEIR LIMITATIONS

Planar Nand flash has been commercially available for more than a decade. Starting as small chips of few MBs the latest one reported has a capacity of 128 GB. The reason for such tremendous growth of Nand flash is that with time more advanced fabrication techniques enabled increasing of capacity and reduction of cost per bit. So far the Nand fabrication techniques have been good enough to satisfy our demands but there is a question mark over the capability of these techniques to sustain the same rate of growth and keep the cost down. The following factors are a hurdle in continued growth of Nand flash memory [13].

1. How much can we scale down the cell? This certainly is one of the most important factors for growth of Nand technologies in future. The thickness of the gate determines that operating and programming voltages. It also governs the electrical behavior of the cell and its noise performance. Apart from these the reliability and durability of the memory also depends on it. So it is understandable that there is a limit to which we can scale down the dielectric. Then it also poses the question which way do we scale the cell?

2. How many electrons can we store in the gate?

With ever decreasing size of the floating gate there is lesser area to store the charge. The problem becomes severe in case of multilevel cells (MLC) where there are fewer electrons for each level. Scaling down further is bound to decrease the number of electrons that gate can hold. Figure 1 shows the expected number of electrons against node size and number of levels.

3. Effect on operating voltages and noise performance

Scaling down the size also affects the operating voltages like program and erase voltage, and read current. At smaller size the effect of various stray capacitances also becomes prominent. Interference due to noise and phenomenon like cross-talk become more difficult to control.

The factors indicate that we cannot continue to scale down in the traditional way. One way to achieve higher capacity and
low cost is to use 3D integration which has been discussed in the next section.

![Figure 1. Number of electrons per logic level [13]](image)

III. 3D INTEGRATION TECHNOLOGY

As the lateral scaling is limited by cost of process along with severe effects on electrical properties and performance the only obvious choice is to go vertical. 3D stacking has been there for a long time but its applicability to Nand flash fabrication is to be considered. Using TSV seems is another option but the cost associated with fabrication process is makes it less attractive for Nand flash which has the cost factor attached to it. Figure 2 shows graphically the variation of cost per bit for 3D stacked Nand and BiCS Nand [1] which is discussed in coming sections.

For these reasons a number of Nand architectures requiring less expensive fabrication process have been proposed. Each of these architectures has been briefly discussed in the next sections.

IV. BIT COST SCALABLE TECHNOLOGY (BiCS) [1]

BiCS technology was one of the earliest methods of 3D Nand implementation. It can be used to fabricate a multi-stacked array with a few constant number of critical lithography process steps. It uses a novel “punch and plug process” for fabrication [1]. Using this process it is also possible to keep the area penalty in control without affecting the cost per bit as shown in figure 2. In the BiCS architecture stacks of Nand are layered on top of each other. Holes are punched through these stacks and strings of cells are connected vertically on the plugs located on these holes. The plugs run through the holes and end at the bottom where they are connected to the source diffusion of the substrate. The lowest layer acts as select gate, the upper layers as control gates and the top layer as upper select gate. Edges of the control gate layers are stair shaped for via holes which are connected to the control transistors. These are connected to multiple rows of strings reducing the number of control transistors required and hence the area even if the number of layers is increased. A single cell can be accessed by selecting the desired bit line and row select line. The cell size of BiCS is $6F^2/n$ where $n$ is the number of layers. This type of structure is also termed as vertical channel Nand flash. The bird eye view of the BiCS architecture is shown in figure 3.

The fabrication process of the BiCS architecture is articulated below.

1. Lower select gate transistor, memory string and upper select gate transistor are fabricated individually.
2. Gate material is P+ poly-Si. Holes for transistor channel or memory plug are punched through and LPCVD TEOS film or ONO films are deposited.
3. The bottom of dielectric films are removed by RIE and plugged by amorphous Si.
4. Arsenic is implanted and activated for drain and also source of upper device. Figure 6 shows schematic cross sections of vertical FET and SONOS memory cell on poly-Si gate. It should be pointed out that ONO films are deposited in the opposite order of the conventional SONOS device, i.e. from LPCVD TEOS film as top block oxide (5nm), LPCVD SiN film (11nm) and LPCVD TEOS film as tunnel oxide (2.5nm).
5. Edges of control gate are processed into stair-like structure by repeating of RIE and resist sliming as shown in figure 8.
6. For minimizing disturb, whole stack of control gate and lower select line are etched to have a slit which separates a block of memory plugs from each other.
7. Upper select gate is cut into line pattern to work as row address selector.
8. Via hole and BL are processed on the array and peripheral circuit simultaneously.

The fabrication process is given in figure 4.

A 90nm BiCS flash memory of size $nx512$ kb, select gate FET with $L=150nm$ and $D=90nm$, shows an endurance of $1e5$ cycles and $V_{th}$ shift of less than 0.5V after $1e4$ cycles. The programming and erase voltage are 15V and -13V respectively. With the above results BiCS technology established itself as one of the strongest contenders for future replacement technology for fabrication of Nand flash memory.

![Figure 2. A comparison of cost per bit for stacked Nand and BiCS Nand as layers increase [1].](image)

![Figure 3. (a) Bird eye view of BiCS Nand, (b) top down view [1].](image)
V. PIPE-SHAPED BiCS TECHNOLOGY (P-BiCS) [2]

The BiCS fabrication process has but three issues that were not addressed and for which the pipe-shaped BiCS technology was introduced [2]. These issues are:

1. Small P/E window, read disturb and low data retention capability makes it a poor candidate for MLC.
2. Large total current due to numerous cells on the same string cause variations in the voltage on the source line.
3. Since the lower select gate is placed in heavily diffused source the diffusion profile is not easily controlled.

The P-BiCS is different in structure from BiCS in that it has a folded pipe like string of Nand cells. Like BiCS it has a vertical channel structure. The program and erase operation is based on Fowler-Nordheim tunneling. The folded pipe-shaped connects two Nand strings at the bottom of the structure. One of the terminals of the pipe is connected to the bit line (BL) and the other is connected to the source line (SL). Multilevel metal layers on the SL give it a lower resistance profile. Architecture of the P-BiCS is shown in figure 5. The fabrication process steps have been listed below [2] and have been shown in figure 6.

1. The first step is the PC formation.
2. The next step is the deposition of the sacrificial films followed by memory-hole formation. For multiple layers multiple layers of memory films should be deposited.
3. The SG transistors are formed after the fabrication of the Nand strings.
4. After SG-hole formation the sacrificial films are removed. The removal of the sacrificial film leaves a U-pipe that connects two vertical Nand cells strings.
5. Next the memory films are deposited and silicon deposition is done after that.

[2] reports a larger P/E window and higher speed of P-BiCS over BiCS [1]. Other performance measurements show higher data retention of P-BiCS with no degradation after 10 years, threshold voltage shift of less than 0.3 V after 100k cycles of read operation with the read voltage of 7.5 V, constant threshold voltage for multi-level data during 10 years. These results by [2] are enough to conclude that P-BiCS that data retention and the immunity to read disturb are sufficient for MLC operation.
VI. VERTICAL-STACKED ARRAY TRANSISTOR (VSAT) [3]

VSAT is fabricated through a method termed as planarized-integration-on-the-same-plane or PIPE. One of the key issues in BiCS based Nand architectures is that the stair-like structure provided to connect to peripherals circuits is not area efficient. The VSAT improves removes this flaw by removing the stair structure as shown in figure 8. The architecture of VSAT is shown in figure 9. The fabrication process steps are given below.

1. A Si mesa is prepared by dry etching. Over this Si mesa multiple layers of gate electrodes and isolating films of poly-doped-silicon and nitride are deposited.
2. The active regions are created through lithography followed by dry etching.
3. Multiple WLs are patterned using KrF lithography followed by dry etching.
4. All the gate electrodes are exposed on the same plane after a CMP process.
5. The tunneling-oxide, charge-trapping-nitride, and control oxide films are deposited in turn on the active region, followed by a poly-silicon deposition process of the channel material. Finally, to isolate vertical strings, an etching process is carried out.

The fabrication process is given in figure 10. The author [3] anticipates a memory chip of 128 GB can be implemented using MLC on 50nm node, 16 layers VSAT.

VII. TERABIT CELL ARRAY TRANSISTOR (TCAT) [4]

Like VSAT, TCAT tried to solve two major concerns of BiCS flash. First, difficulty in etching metal/oxide multilayer simultaneously makes it very difficult to implement metal gate structure for BiCS flash. Without this it one cannot achieve good erase speed, wider Vth margin, and better retention characteristics. Second, GIDL erase of BiCS flash requires an extensive circuit change to apply negative bias on word line. Thus there is an unwanted area overhead with limited erase voltage. TCAT can resolve these two issues of BiCS Nand memory. There are various structural differences between BiCS flash and TCAT

i) Oxide/nitride multilayer stack,
ii) Line-type ‘W/L cut’ etched through the whole stack between the each row array of channel poly plug,
iii) Line-type CSL formed by an implant through the ‘W/L cut’,
iv) Metal gate lines have been replaced for each row of poly plug. Gate replacement process has been implemented to achieve the metal gate SONOS structure.

The structure of TCAT is shown in figure 11. W/L cuts are dry etched and sacrificial nitride layer is removed by wet removal process after which gate dielectric layers and gate metal are deposited in the conventional order.
Separation of each gate node is followed by etch processes. The cell string consists of NAND cell transistors with SSL transistor at the top and the GSL transistor at the bottom. The SSL and GSL transistors are formed simultaneously with cell transistors.

The additional ‘W/L cut’ structure does not contribute to any area overhead to the cell of TCAT flash which is inherently smaller than BiCS flash. The extra area penalty of BiCS is due to ‘gate first’ process. The minimum diameter of plug hole of BiCS flash is limited by the gate dielectric layers. The channel poly plug in TCAT flash structure is connected to Si substrate, not n+ common source diffuson layer as in the BiCS flash which makes it possible to implement the conventional bulk erase operation and that too without any major peripheral circuit changes.

VIII. VERTICAL GATE NAND FLASH ARCHITECTURES [5-8]

The earlier 3D Nand architectures-BiCS, P-BiCS, VSAT and TCAT, are also known as vertical channel Nand flash. Although, they tried to match the performance of the planar Nand yet they have certain shortcoming due to architectural flaw. For example, BiCS Nand flash has difficulty with WL interconnect, program disturbance, and channel resistance and they get worse as the number of WL between top BL and bottom CSL increases [5]. P-BiCS and TCAT have structures such that the channel current is conducted through a hole drilled through the layers in the vertical direction, and an additional WL-cut process must be applied to isolate the WL’s in the X direction. Thus they have limited X pitch scalability due to the corresponding lithography overlay issue involved. The cell size of all vertical channel architectures is 6F$^2$ which is relatively large and does not correspond to the traditional planar Nand cell size [7]. Moreover, as the number of layers increases, the read current inevitably degrades due to the increase in the length of the NAND string [6]. The Vertical Gate (VG) 3D NAND architecture [5-8] has no pitch scaling limitations and with a 4F2 cell size is highly scalable.

One of the early VG-NAND architecture proposed [5] includes WL, BL, CSL, horizontal active string with pattern, VG (for SSL, WL, GSL), Charge Trap Layers between active and vertical gate, Vertical Plugs of DC, source and active body. The structure of the VG Nand has been shown in figure 12. WL and BL are formed at the beginning of fabrication before cell array making interconnect between WL, BL and decoder easier. To enable body erase operation the source and active body (Vbb) are electrically connected to CSL. To perform erase operation a positive bias is applied to CSL. The array schematic of each layer is similar to that of a planar NAND flash except for SSL. VG-NAND requires multiple SSLs for multiple active layers because VG-NAND uses common BL and common WL between multi-active layers to select data from a chosen layer out of multi-layers. Cell size of VG-NAND is 4F$^2$ per layer.

The fabrication steps are listed below.
1. Integration sequence of VG-NAND and the integration scheme is based on simple patterning and plugging. BL with n+ poly-Si is fabricated first and then n+ poly-Si WL is formed on top of it.
2. Multi-active layers with p-type poly-Si are formed with n-type ion implants for SSL layer selection and alternated inter-layer dielectrics are inserted between actives.
3. Patterning is done on the multi-active layers and charge trap layers (ONO) are deposited over the patterned actives.
4. Consecutively VG is formed and connected to WL.
5. In the final step, vertical plugs of DC and Source-Vbb are connected to BL and CSL after contact ion implants. N+ doped source and p-type active are electrically tied to CSL.

The fabrication steps have been show in figure 13.

The vertical gate NAND was proposed described above has a simpler process and more lateral scalability but the implementation is done on very large devices with half pitch >0.15um. Moreover, the proposed in-layer array decoding method is complex and inefficient. A vertical gate architecture that takes care of the above mentioned issues uses buried-channel (n-type well) junction-free structure [6]. Each device
is a double-gate TFT BE-SONOS. The channels are all n-type doped poly (buried-channel) thus additional junction implantation is not required. Buried-channel (n-type well) device helps not only to improve the read current of TFT NAND but also allows implementation of the junction-free structure required for 3D stackable devices. The decoding is also easier to implement. The conventional WL’s and BL’s are grouped into planes. The conventional BL contact is replaced by the SSL. The intercept of the three selected planes (WL, SSL, and BL) defines the selected memory cell. However, in this architecture it is difficult to isolate SSL gate in the X-direction and can limit the pitch scalability of the cell. To avoid this one must do away with the need to fabricate plural SSL gates in one block. Rather the decoding circuit and the array structure should be simple and highly scalable.

To achieve this PN-diode decoding method has been proposed [7]. The PN diodes are fabricated self-aligned at the source side of the vertical gate architecture as shown in figure 14. Also the need to fabricate plural string select (SSL) transistors inside the array has been completely eliminated. This allows for a highly symmetrical and scalable cell structure. The PN diodes prevent the leakage of the self-boosted channel potential. Since there are no plural SSL gates in one block the array structure is now simpler than before and highly scalable. The PN diode is fabricated, self-aligned, at the source side by using either P+ ion implantation or P+ poly plug process. The cell structure is made symmetrical and the layout is similar to that of the conventional 2D NAND. This feature of the architecture proves good for scalability and provides 1/2-pitch scalability to 2Xnm node or below. Each device is a double-gate (vertical gate) horizontal-channel device. To increase the read current, channel is fabricated lightly-doped n-type (buried-channel device).

Decoding can be further improved by using a self-aligned independently controlled double gate (IDG) string select transistor (SSL) decoding method [8]. The VG NAND is architecture consists of a double-gate TFT BE-SONOS device, fabricated by the intercept of WL’s and BL’s as shown in figure 15. The self-aligned independent double gate (IDG) SSL is implemented by stripping the top portion of SSL poly gate after WL patterning. Every SSL is independently connected through the interconnection of CONT, ML1, VIA1, and ML2 toward the SSL decoder (in parallel to the WLs). Each unit has 2N channel BLs (poly channel), where N is stacked memory layer. These BLs are controlled by the corresponding 2N SSLs. The 2N SSLs share the same BL pad, where a total of N staircase BL contacts are fabricated. Each BL corresponds to one memory layer. Stacked VIA/CONT connect the staircase BL contacts to ML3 BL’s and the page buffer for sensing. There are many units repeated in X direction to connect the IDG SSLs of plural units in one block. The ML3 BL and the staircase contacts have double X pitch for better process window. A common source line (CSL) is

**Figure 13. Fabrication process of VG Nand [5]**

**Figure 14. Structure of PN-diode decoded 3D Nand architecture [6]**

**Figure 15. (a) Schematic structure of IDG decoded 3D Nand, (b), (c) process flow to form the self-aligned IDG SSL device [8]**
IX. SINGLE CRYSTALLINE STACKED ARRAY (STAR) NAND FLASH MEMORY ARCHITECTURE [9]

With STAR Nand architecture, a new unit of 3-D structure, i.e., “building” has been proposed. Using this new unit, a 3-D block and full chip architecture can be successfully designed. From figure it can be concluded that compared to vertical channel architectures as the number of stacks increases the STAR can attain terabit density more quickly. The implementation of STAR takes care to minimize the additional process steps and extra area in memory array part. The VG Nand [5] has as area overhead penalty of memory unit area since it needs several string select line (SSL) transistors in one channel string. To make the SSLs, additional photolithography and ion-implantation steps are added at each stacked layer. To have good compatibility with peripheral memory functional blocks, the operation methods similar to the conventional Nand Flash memory have been employed. Throughput penalty related with page program, block-erase, and page read have been avoided as much as possible.

In the STAR structure, the BLs are formed on the top floor of a building. Further, the BLs are also perpendicular to lines WLs and SSLs which are parallel with different levels. They fabricated, which connects all sources lines of every memory layers. The decoding is explained through the following page operation.

Each memory device is selected by the intercept of WL (64-WL for one block), ML3 BL (corresponding to each memory layer), and page (corresponding to the sandwich of two SSL’s). The unselected adjacent pages are inhibited by the IDG SSL operation using an inhibit bias (V\textit{ inhibit}) at the other SSL gate, as explained next.

1. Assuming a large-density memory chip with M (such as 16Kb) channel BL’s, and N (such as 8) memory layers. Since the ML3 BL has double X pitch, the total number of ML3 BL number is M/2 (8Kb), while the total unit number is M/16 (1Kb).
2. Every unit has 2N (16) pages, where each page is defined by the sandwich of two adjacent SSL’s.
3. To select one page (such as SSL0/1) for each WL (such as WL30), it selects M/16 (1Kb) SSL devices in parallel units.
4. On the other hand, if we allow all-bitline (ABL) sensing for the 8-layers together, the total selected devices are 1Kb*8=8Kb (M/2), which defines the page size.
5. In the architecture, increasing memory stacks doesn’t decrease the array layout efficiency but just simply change the BL pad layout and the associated page number.

As discussed using various architectures, the VG Nand promises to be a potential technology for future 3D Nand memory fabrication. In the next section we discuss a novel VG Nand architecture.
are connected with each decoder at the opposite end site. The gate of SSL is connected between the BL and the first WL without the underlap region. Leakage current can degrade the self-boosting efficiency, so to reduce it the gate length of the SSL transistor is kept longer than that of the memory cell transistor. The designed SSL can have longer gate length without extra area. The GAA structure of SSL also maintains excellent current drivability of STAR. The source line (CSL) which is N+ doped common and p-type body are electrically tied. Since the gate of the GSL transistor covers the N+ region, sufficient electron supply is possible during read or program operation. Additionally, the channel of STAR is connected to a p-type body. As a consequence, conventional bulk erase operation can be achieved. The widths of A and B should be same for maximum memory density and BL throughput.

The fabrication steps of STAR are given below.

1. In the first step, SiGe/Si layers that are formed sequentially and epitaxially grown on the Si substrate followed by active channel formation. Oxide/poly-Si/oxide layers are used as a hard mask for etch. Then, n- and p-type ion implantations are carried out for the BL region and body, respectively.
2. The screening oxide is deposited for preventing the donor from penetrating into the active channel. Next, high dose n-type ion implantation is performed to make the N+ CSL region.
3. Oxide deposition is followed by patterning the oxide buttress, which sustains the long channels not to be collapsed or adhered. Then, selective SiGe etching process is carried out. If SC1 solution is used for SiGe removal, silicon fin is rounded as NH4OH and H2O2 solution also slightly etches silicon.
4. Oxide re-deposition to fill the gap between Si channels is performed. As a consequence, conductive SiGe is replaced with an oxide insulator.
5. Oxide patterning carried out for making WLs.
6. Isotropic oxide etch is carried out. The key point of this process is that the channel width must be smaller than the oxide buttress width for preventing oxide buttress from being removed during oxide etching. After isotropic recess, the width of oxide buttress decreases and the silicon channel for unit cell is exposed.
7. Growth of ONO dielectrics is followed by tungsten deposition and planarization for the gate of WL, SSL, and GSL. If tunnel barrier engineering is used for tunneling oxide formation, more reliable memory characteristics can be achieved. Using this damascene gate process, all gates of cell, SSL, and GSL transistors are self-aligned.
8. SSL transistors are made by lithography [(h) and (i)].
9. BL region is made by carrying out trench etch. Next, SiGe selective etch is performed for perfect isolation between the BLs [(j) and (k)].
10. Finally, the stair-like BL structure is created.

The fabrication process is graphically depicted in figure 19. Single-crystalline stacked array (STAR) shows to have many advantages over other Nand architectures. It shows better scalability, fabrication of a single-crystal channel. It is less sensitive to 3-D interference, has stable virtual source/drain characteristics, and has better extendibility over other stacked structures. The unit cell size of STAR is larger than that of VG
NAND because O/N/O gate dielectric layers are formed along oxide buttress during damascene gate process. Other advantages are uniform distribution of cell performance because of the absence of the defects associated with grain boundaries, large BL read current and small sub-threshold swing owing to the GAA structure, Stable virtual source/drain (S/D) characteristic. Small intra-layer interference, Immunity to interlayer interference, Small channel–channel (Ch–Ch) coupling.

X. Noise Performance Analysis of 3D NAND Architectures [10]

We have discussed various 3D NAND architectures and the fabrication process involved. The most important issue that remains to be discussed and explored in the future is the noise performance of these 3D architectures and its impact on the performance and scalability. It is observed that most vertical channel NAND technology uses poly-silicon as a channel material. Two intrinsic variation sources of the cell threshold voltage induced by poly-silicon traps have been identified and simulated: Random Trap Fluctuation (RTF) and Random Telegraph Noise (RTN) [10].

RTF is due to fluctuations of the traps location inside Poly-Si channel. These traps follow a Poisson statistics and the traps density can be used as reliable metric for evaluating the electrical performance of Poly-Si channel.

RTN is induced as a result of RTF in poly-Si and is observed to follow an exponential distribution. The RTN energy distribution shows that most of RTN traps are present at Fermi level and these are induced during program operation by the cycling of the cell.

Vertical NAND devices that rely on poly-Si channel consist of a charge trapped device (CTF). Till now models of intrinsic variability poly-Si for sub-10nm 3D generation have been missing. The modeling of RTF and RTN is very important as it can allow early prediction of $V_T$ distribution for 3D NAND devices that implement MLC operation. RTF and RTN have been briefly discussed below.

Random trap fluctuations – Poly-Si is made of silicon grain having different crystalline orientations. It has been observed that they have traps sites making the grain size extraction arbitrary. Poly-Si has been modeled as a silicon material with large traps density distributed uniformly inside the channel. Usually, by fine tuning the IBL/VWL curve of a cell with median characteristics for -20°C to 85°C range can be used to evaluate the traps distribution. It is also observed that $V_T$ and sub-threshold slope temperature dependence is due to traps density at mid gap, in the 1-5×10^18 cm^-3 range. The other variations sources in VNAND are due to structure fluctuation and injection spreading. From figure 20 it can be observed that the modeled $V_T$ distribution after a single program and erase pulse agree with the measurements. Considering its effect on performance and device characteristics, RTF must be
considered and measured when modeling one pulse erase and program distribution. Random telegraph noise – For each cell, RTN is evaluated by measuring \( V_c \) 200 times and compared to the average cell value. RTN distribution presents an exponential tail even if vertical Nand channel is not doped and this is similar to bulk NAND cells. The exponential tail is due to the presence of traps inside the channel. This results in non-uniform current density making the structure more susceptible to generate large RTN tail. As a result, RTN must be considered when modeling RTN in vertical Nand structure. The RTN traps occupancy follows Fermi statistics and thus their occupancy probability will depend on the trap energy level. Thus RTN trap above (below) Fermi level will be mainly empty (filled) inducing a positive (negative) tail. Figure 20 shows the RTN trap energy distribution in no and cycled case extracted. The extraction performed in no cycle case shows that the majority of the switching traps are at Fermi level. It is also seen that after 3k cycles a RTN traps are created that are 0.2 eV above the Fermi level in read condition implying that cycling induces the creation of switching traps thus enhancing the positive RTN tail. This can be explained by the degradation induced by the programming step. Thanks to these RTF and RTN models, VNAND distribution during MLC operation can be accurately modeled.

**CONCLUSION**

Time is running out for planar Nand technology. It will not be long that planar Nand will be completely replaced by 3D Nand. 3D Nand promises to satisfy the growing need of Nand memory. Figure 21 shows the transition of Nand from 2D to 3D [12]. Table compares the most important 3D memory technologies that can potentially the planar Nand. More studies and experiments need to be done to categorically understand the effect of noise and other architecture induced impacts on the performance of 3D Nand.

**TABLE I. COMPARISON OF 3D NAND FLASH MEMORY ARCHITECTURES**

<table>
<thead>
<tr>
<th>Comparison of 3D Nand flash memory architectures</th>
<th>Vertical Channel</th>
<th>Vertical Gate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P-BiCS [1]</td>
<td>VSAT [2]</td>
</tr>
<tr>
<td>Cell Size</td>
<td>6F²</td>
<td>6F²</td>
</tr>
<tr>
<td>Current Flow Direction</td>
<td>U-turn</td>
<td>Vertical</td>
</tr>
<tr>
<td>Device Structure</td>
<td>GAA</td>
<td>Planar</td>
</tr>
<tr>
<td>Possible Minimum F</td>
<td>~50nm</td>
<td>~50nm</td>
</tr>
<tr>
<td>Impact of number of layers of memory</td>
<td>Low read current</td>
<td>Low read current</td>
</tr>
<tr>
<td></td>
<td>No impact</td>
<td>No impact</td>
</tr>
</tbody>
</table>

**REFERENCES**


A Highly Scalable Vertical Gate (VG) 3D NAND Flash with Robust Program Disturb Immunity Using a Novel PN Diode Decoding Structure Chun-Hsiung Hung+, Hang-Ting Lue, Kuo-Pin Chang, Chih-Ping Chen, Yi-Hsuan Hsiao, Shih-Hung Chen, Yen-Hao Shih, Kuang-Yeu Hsieh, Mars Yang, James Lee, Shu-Yu Wang, Tahone Yang, Kuang-Chao Chen, and Chih-Yuan Lu, Symposium on VLSI Technology, 2011

A Highly Pitch Scalable 3D Vertical Gate (VG) NAND Flash Decoded by a Novel Self-Aligned Independently Controlled Double Gate (IDG) String Select Transistor (SSL) by Chih-Ping Chen, Hang-Ting Lue, Kuo-Pin Chang, Yi-Hsuan Hsiao, Chih-Chang Hsieh, Shih-Hung Chen, Yen-Hao Shih, Kuang-Yeu Hsieh, Tahone Yang, Kuang-Chao Chen, and Chih-Yuan Lu, Symposium on VLSI Technology, 2012


Intrinsic Fluctuations in Vertical NAND Flash Memories by Etienne Nowak, Jae-Ho Kim, HyeYoung Kwon, Young-Gu Kim, Jae Sung Sim, Seung-Hyun Lim, Dae Sin Kim, Keun-Ho Lee, Young-Kwan Park, Jeong-Hyuk Choi, Chilhee Chung, Symposium on VLSI Technology, 2012

