

Rino Micheloni *Editor*

3D Flash Memories

 Springer

Rino Micheloni
Editor

3D Flash Memories

 Springer

Editor
Rino Micheloni
Performance Storage BU
Microsemi Corporation
Vimercate
Italy

ISBN 978-94-017-7510-6 ISBN 978-94-017-7512-0 (eBook)
DOI 10.1007/978-94-017-7512-0

Library of Congress Control Number: 2016936991

© Springer Science+Business Media Dordrecht 2016

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Science+Business Media B.V. Dordrecht

Contents

1	The Business of NAND	1
	Rahul N. Advani	
2	Reliability of 3D NAND Flash Memories	29
	A. Grossi, C. Zambelli and P. Olivo	
3	3D Stacked NAND Flash Memories	63
	Rino Micheloni and Luca Crippa	
4	3D Charge Trap NAND Flash Memories	85
	Luca Crippa and Rino Micheloni	
5	3D Floating Gate NAND Flash Memories	129
	Rino Micheloni and Luca Crippa	
6	Advanced Architectures for 3D NAND Flash Memories with Vertical Channel	167
	Luca Crippa and Rino Micheloni	
7	3D VG-Type NAND Flash Memories	197
	Andrea Silvagni	
8	RRAM Cross-Point Arrays	223
	Huaqiang Wu, Yan Liao, Bin Gao, Debanjan Jana and He Qian	
9	3D Multi-chip Integration and Packaging Technology for NAND Flash Memories	261
	Herb Huang and Rino Micheloni	
10	BCH and LDPC Error Correction Codes for NAND Flash Memories	281
	Alessia Marelli and Rino Micheloni	
11	Advanced Algebraic and Graph-Based ECC Schemes for Modern NVMs	321
	Frederic Sala, Clayton Schoeny and Lara Dolecek	

**12 System-Level Considerations on Design of 3D NAND
Flash Memories 349**
Chao Sun and Ken Takeuchi

Index 377

Chapter 7

3D VG-Type NAND Flash Memories

Andrea Silvagni

7.1 Introduction

Solid State Disk applications have been the main drivers for the cost reduction of NAND Flash memory. Technology scaling limits have been almost reached and new directions were investigated for a decade. 3D NAND appeared as the most promising and short term viable solution to the scaling problem. 3D NAND offers the possibility to keep scaling memory density beyond 2D NAND 1z nm technology and opens new landscapes if the technology hurdles can be definitely overcome. Additionally Multi bit per cell technology is applicable to 3D NAND as well; therefore, the storage capacity growth trend is not going to meet the inflection point soon. Big Data storage demand remains certainly the driver for the rapidity of 3D NAND to become industry mainstream.

With respect to the flat 2D NAND Flash, three-dimensional architectures demand a more challenging approach to deal with process, design architectures, parasitic effects, testing. At the time of writing this chapter, 3D technology is moving into production inside solid state storage solutions; however, those NAND chips aren't commercially available as raw components. 3D NAND chips are used by vertically integrated manufacturers and few partners to build the storage systems under a strict control on the usage. Technology solutions are still diversified among the few players in the 3D arena and no unique mainstream solution is emerged yet.

In this chapter the principal architectural pieces of 3D NAND will be reviewed putting the focus on one of them: VG-Type, Vertical Gate 3D NAND.

A. Silvagni (✉)
Macronix, Milan, Italy
e-mail: andreasilvagni@macronix.com

7.2 3D NAND Architectures

Several options for a true 3D NAND have been proposed for a decade in the technical symposiums before the first use in production. The main contributors have been Toshiba [1–3], Samsung [4–7], Macronix [8–10], Hynix [11–13], Micron [14] amongst the Memory companies. The main scope of 3D development was lowering the cost per bit and increase the chip density without shrinking the 2D lithography node. In the following Fig. 7.1, the main concepts presented in technology symposiums are summarized. The key structures have been presented in the range of only nine years, from 2007 to 2015. Most of the structures are based on charge-trap devices, only few are based on floating-gate, and the choice of one specific technology over another remains closely linked to the experience and confidence of each player. The production of products based on 3D NAND Flash cell started in 2013 and by the end of 2015 have been announced 3D-NAND chips with capacity of 256 Gb with 48 layers and 3bit per cell technology. Breakthrough of 1 Tb in one single chip is expected in the coming years. Vertical channel architecture type is currently the main choice, while charge trapping and floating gate technologies seem to be both pursued for production but with a preference on the first for most players.

The common feature among the different 3D NAND solutions is constituted by very deep vertical (z direction) etching steps that define the Flash cells geometries simultaneously. Transistor geometries are formed by the deep trench through a

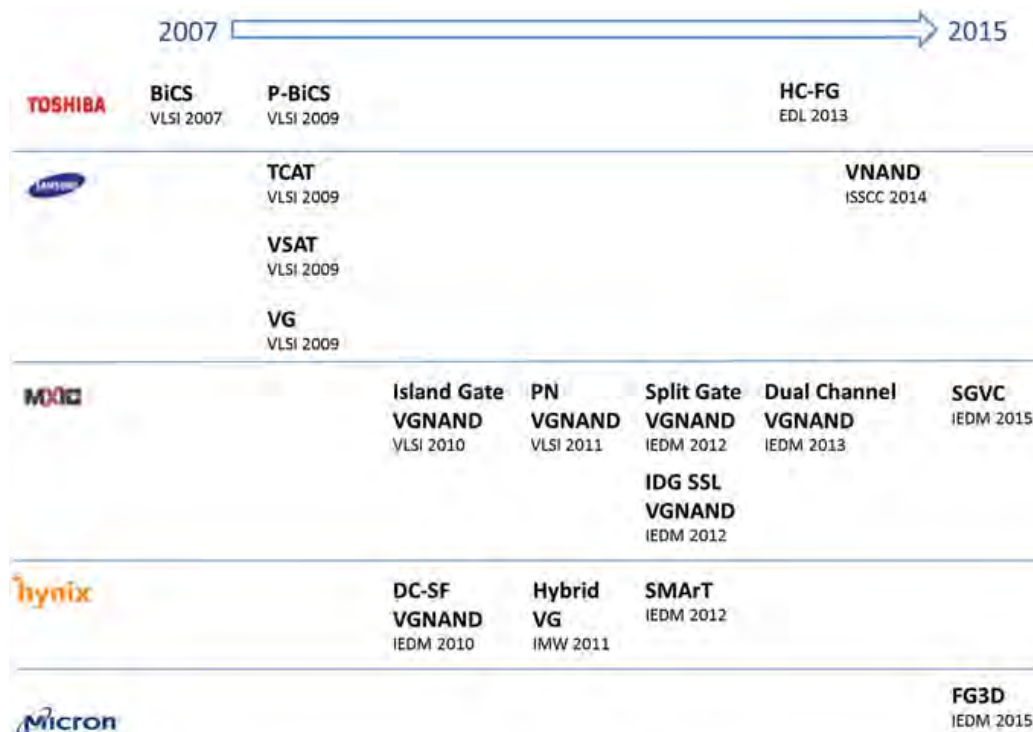


Fig. 7.1 3D Architectures proposed in technical symposiums from 2007 to 2015

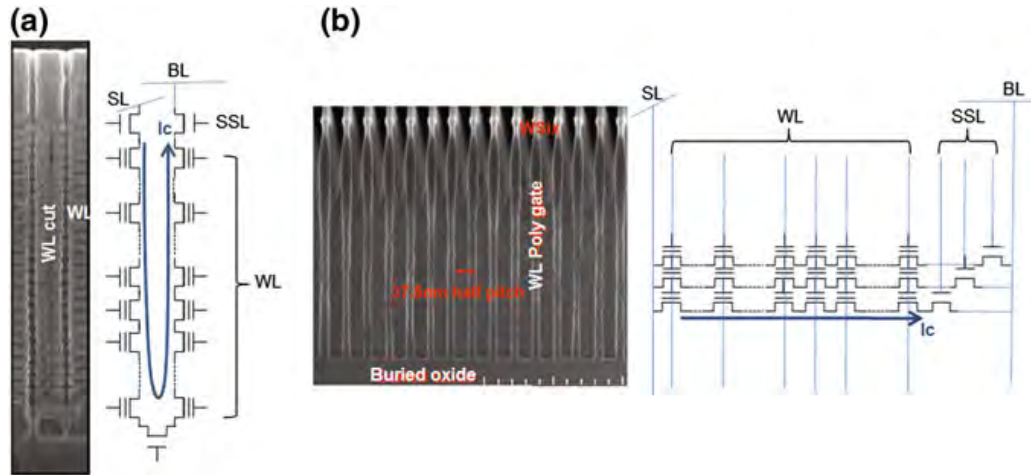


Fig. 7.2 a VC- type P-BICS 3D NAND, b VG-Type 3D NAND [2, 8]

multiple polysilicon/oxide stack. The cells stack shows an impressive high aspect ratio such as the one shown in Fig. 7.2, where one example of *Vertical-Channel* (VC) and one of *Vertical-Gate* (VG) type are shown. In VC gate-all-around type, the channel is realized by etching a hole through the layers stack in one single step, and then forming the transistor structure with deposition of its ONO charge trapping layers, tunnel oxide and the polysilicon channel fill in the middle. The cell gates are constituted by the polysilicon horizontal layer surrounding the vertical channel forming a *Gate-All-Around* (GAA) structure. The string current flows in the cells in the vertical direction. An additional wordline cut is necessary to separate the wordlines in the X direction. The channel is connected by a bridge at the bottom to another adjacent stack so that the resulting string is constituted by two times the stack and thus both the bitlines and the sourceline can be connected through the top metal backend and dedicated selector transistors. In the VG-type the vertical etching is necessary to separate the strings in one direction and to separate the wordlines in the other direction. The wordlines and the horizontal strings can be identified in the figure. The current flows in horizontal direction and each layer must be connected to the top metal bitlines and source lines by a proper connecting structure (not shown here).

3D NAND architectures need to be proven cost effective in terms of process costs and product yield because the process complexity is not trivial if compared to 2D NAND. The cost must be evaluated considering other factors rather than simply the number of layers: the overhead of decoding structures and circuits plus the yield determine the cost advantage. The non-uniformity of the critical dimensions through the layer stack plays a significant role in the determination of the effective cost per bit. It is recognized that in order to reach cost effectiveness in 3D NAND the number of layers must be roughly above 32 which is already an impressive number. Currently 48 layers have reached production stage. About 64 layers might be already a limit, considering the etching process difficulties of a stack with an aspect ratio of 40 [15].

Most of the proposed 3D NAND devices are based on *Charge Trapping* (CT) SONOS (Silicon-Oxide-Nitride-Oxide-Silicon) devices but also traditional floating gate structures have been proposed. CT SONOS devices did not become the mainstream in 2D NAND Flash where Floating gate cells are still dominant; however several drawbacks of the CT technology can be minimized by engineering the charge trap stack, Read/Write algorithms and advanced error correction schemes managed by the controller processor in Solid State Disk products.

Figure 7.3 shows the top view of 3D NAND cells for vertical channel and vertical gate types. The drilled circular hole imposes some limitations on both X and Y directions for VC-type 3D NAND. The constraints are the minimum diameter size of the channel determining the cell current and the minimum thickness of the ONO trapping layers for sustainable reliability. The resulting minimum hole size must be greater than roughly 50 nm, although much bigger hole size such as 120 nm would be necessary to control some negative effects induced by the circular structure and critical dimensions along the stack [16]. This hard limitation in the 2D shrink-ability of GAA VC-type needs to be compensated by the increase of the number of layers. A vertical etch (wordline slit) is also needed to separate the different strings control gates. In order to optimize the flash cell array density, the hole trenches can also be arranged in a staggered fashion [16], controlling two strings with the same wordlines and select gate, thus requiring one less wordline cut.

For VG-type NAND, the cell geometrical limitations are restricted mainly to one direction due to ONO stack and to the polysilicon gate fill-in. Hence, the overall cell size of VG-type can shrink better than VC-type in 2D dimensions. The double-Gate VG-type NAND can achieve a $4F^2$ cell size in $2x$ nm and better scalability than VC-type; therefore, generally speaking, less layers are necessary to obtain the same cost with benefits on the manufacturability.

Figure 7.3c, d show a simplified 3D view of a Vertical-Channel hole-type 3D NAND and a Vertical-Gate 3D NAND. The VC-type cell gate all-around structure with the channel built on a poly structure is shown. Gate all-around structures leverage the very effective electric field enhancement effect resulting from the circular structure, thus obtaining optimized programming conditions. However, field enhancement also enhances disturb effects such as program disturbs and V_{pass} disturb. Due to the field enhancement, the circular cell structure has a strong sensitivity to curvature variations present along the stack that may limit the practical minimum cell size, as already mentioned. *Macaroni* body structures [17], constituted by a thin polysilicon channel with hollow center filled by dielectric, have been utilized for better gate controllability and to reduce the trap density at the boundary.

In the 3D-VG-type NAND shown in Fig. 7.3d the channel is constituted by the polysilicon layer which is defined by first etching of the poly/oxide stack. Afterwards, the tunnel oxide and the ONO charge trapping stack is deposited and finally the wordlines are formed. The limits of scalability in VG-type 3D NAND are only constrained by the ONO structure and the poly channel and not by the field enhancement constraints. Because there is no field enhancement effect due to the planar structure in VG-type 3D NAND, the resulting planar FinFET cell transistor

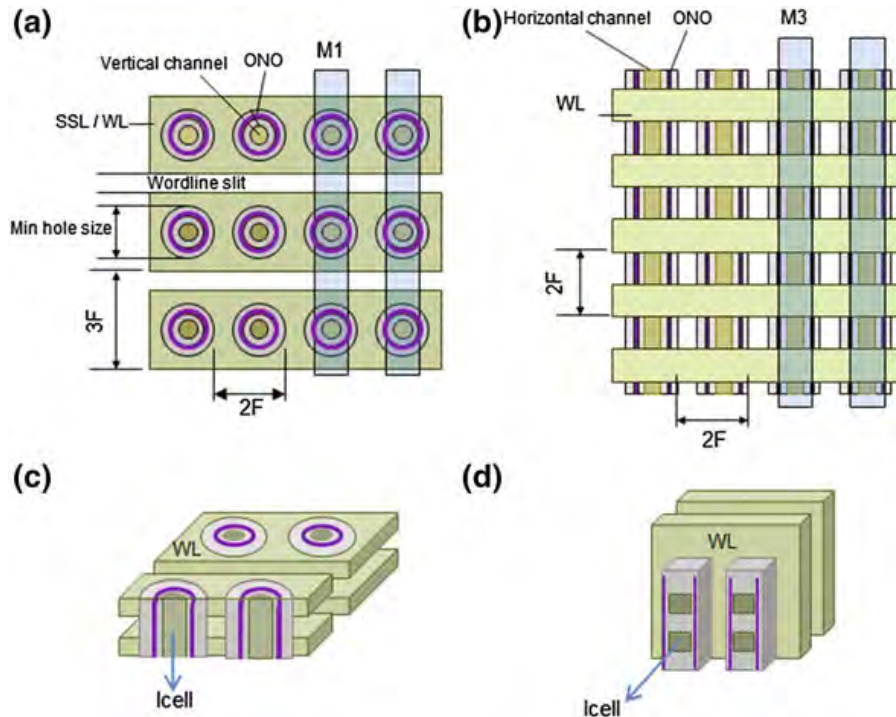


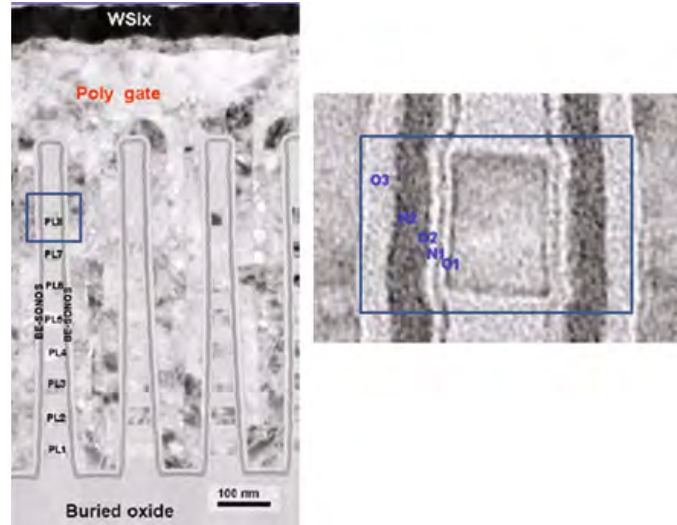
Fig. 7.3 3D-SONOS VC and VG architectures: top view and 3D view

needs to be engineered; however, the excellent planarity of the structure makes the engineering of the oxide and nitride stack easier, by utilizing the extensive knowledge available for barrier engineered BE-SONOS 2D structures. The channel current flows horizontally in VG-type 3D NAND and, therefore, there is no degradation when the number of layers is increased; on the contrary, each layer may exhibit a different behavior from string current performance point of view. Even if the channel lateral dimension of the FinFET cell structure is as small as 8 nm, the channel current is not degraded because the effective channel width is determined by the poly thickness in the Z direction, which does not necessarily need to scale. In contrast, VC-type NAND channel effective width is determined by the planar dimension of the poly channel.

A *Single Gate Vertical Channel* (SGVC) architecture not based on drilled-hole channel with GAA structure has been presented [9]. The SGVC structure shown in Fig. 7.4 can achieve a better memory density than other drilled-hole VC-type structures, with the same advantages of the cell's planar structure of the VG-type 3D NAND.

SGVC implements a single-gate flat CT-TFT (*Charge Trapping Thin Film Transistor*) cell with an ultra-thin body. The vertical etching doesn't need to be precisely controlled because the flat cell is tolerant to critical dimension variations; this is different from GAA structures which are susceptible to field-enhancement variations due to curvature variations. A U-turn string is naturally formed by polysilicon deposition on the trench sidewall. The U-turn structure allows a

Fig. 7.5 Vertical section of 8 layers VG-type 3D NAND [10]



polysilicon gate and the polysilicon channel. The charge trapping medium is a very planar BE-SONOS stack (ONONO).

The NAND cell string runs horizontally and is made of several wordlines plus the dummy wordlines and the select transistors. A silicide layer can be easily deposited on the top of the polysilicon gate to achieve a lower wordline resistance and a very fast access time despite the long wordlines.

A wordline pitch of 2 times 38 nm and a bitlines pitch of 2 times 75 nm have been studied [10]. With such a relaxed pitch and 8 layers, it is already possible to produce an equivalent cell efficiency of a 2D NAND in 1× technology node. A scaled down version already demonstrated the equivalent 1y nm technology node feasibility.

Figure 7.6 shows a 3-dimensional view of the building block of VG-type split-gate 3D NAND structure. Each string is made of 64 wordlines. At the end of the string the channel is selected by an independent SSL (*String Select Line*) transistor (*island-gate* formed transistor) and then connected to a common poly plate (BL pad) indicated as PLn. Each BL pad is used to connect all the strings of the corresponding layer, and then each BLpad is connected to the top Metal 3 bitlines in order to achieve the connection of the BL pad polysilicon plate. Staircase contacts (BL poly plugs) are realized by drilling the BL pads progressively. Poly plugs are finally realized in order to connect each pad to the corresponding top bitline.

On the opposite side of the SSL select transistor, NAND strings are vertically connected to a poly plug. In this way, the source line of the stack is directly connected to the metal backend with a low resistance.

Figure 7.6 is the basic building block of the memory array: it is worth noting that SSL select transistors and source plugs are alternated on both sides of the block in an even/odd fashion. With this arrangement, it is possible to keep a relaxed double pitch between SSL transistors and the corresponding backend connections, at a cost of a slightly lower array density.

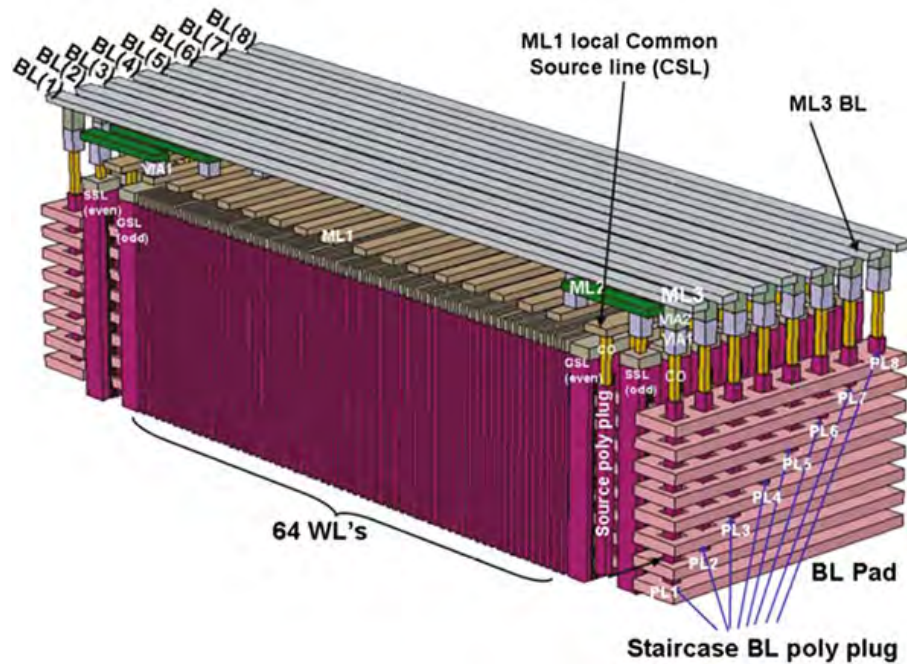


Fig. 7.6 Horizontal channel/vertical gate structure 8-layers split page [10]

Figure 7.7 shows the top view of the VG-type NAND unit of the 8 layer split-bitline structure. Each unit is composed of 16 strings for each layer. Since each string has its own SSL select transistor, the main metal bitlines are connected to a pair of poly BL pads from each side. Each string also has two GSL (*Ground Select Line*) due to the split-page architecture. In total there are 16 pages in a unit. The full page is built by repeating the unit block several times (e.g. 8 k times to obtain an 8 kB page size). When the first page (SSL0) is selected, all the strings corresponding to the same select transistors are connected to the 8 Metal bitlines: cells from all the layers are present in a page.

The layer selector used to connect each layer to the top bitlines is critical for VG-type 3D NAND in terms of area overhead. Other approaches for the layer selector scheme and SSL arrangement have also been proposed: PN diode selection method [19] and staggered SSL arrangement [26]. In particular, the proposed staggered SSL arrangement allows achieving a block efficiency of 80 %, which is similar to that of 2D NAND.

The characteristics of 3D-VG type can be summarized as follows:

1. Planar Scalability: studied down to 25 nm;
2. Equivalent 2D technology feasibility well below 10 nm;
3. Horizontal planar BE-SONOS flash cell that allows mitigate the CD variability along the layers;
4. Low wordline and Source Line CSL resistance;
5. Constant array efficiency independently from the number of layers number;
6. Multilevel functionality.

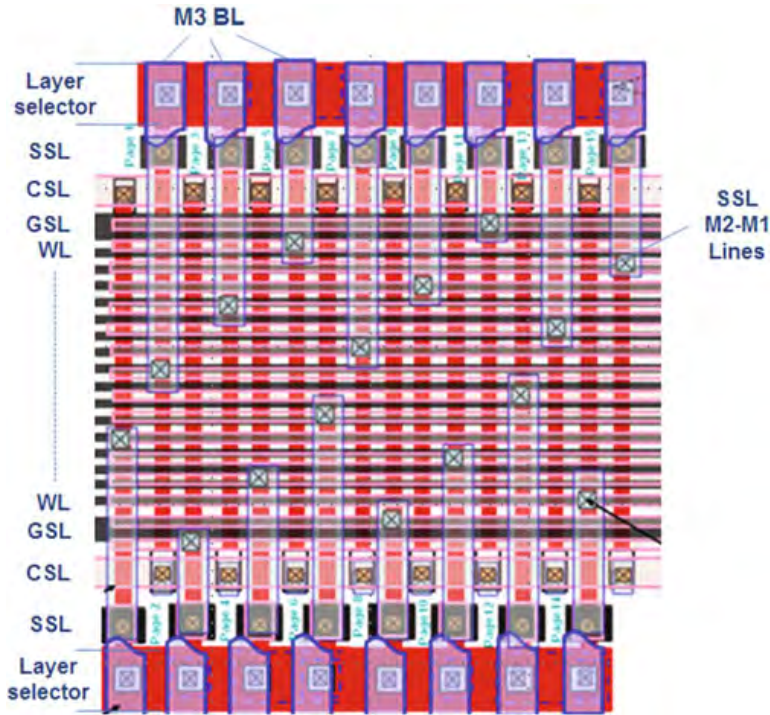


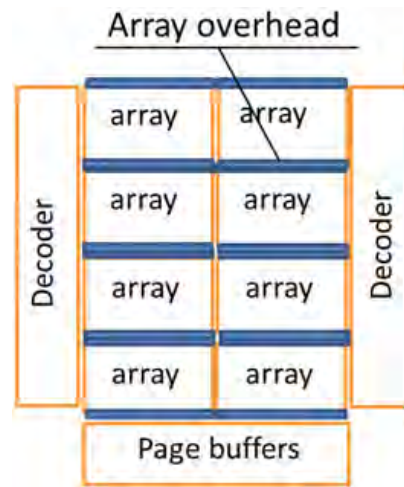
Fig. 7.7 Top view of the VG-3D NAND split-page unit block [10]

7.4 Key Architectural Considerations for VG-Type 3D NAND

The main challenge of 3D NAND is related to vertically stacking more devices in order to increase the bit density per 2D unit area. Obviously, the more layers can be realized, the lower should be the overall cost; however, several aspects contribute to the reduction of the theoretical gain. One of the issues is related to the yield loss derived from complexity of vertical trench or hole etching with many layers. Besides yield issue impacting on cost, the area efficiency of the 3D architecture is dependent on the overhead of side structures such as layer connecting structures or decoders. Architectural solutions are required to optimize the overall array efficiency.

From a performance perspective, there are also few aspects that call for architectural innovations. The use of charge trapping cells requires complex programming and erasing algorithms, and the engineering of the trapping layers. The wordline and source line resistance must be kept low without impacting the array size. RC delays in wordline may not only have an impact on AC performances but also on the efficiency of the programming sequences used to reduce disturbs. The long source lines need to be designed to minimize the parasitic voltage drops, which may reduce the margin for multilevel cell operations. Those problems are specific to each type of 3D architecture and key innovations are required.

Fig. 7.8 Array overhead in 3D NAND



The first technology challenge in 3D structures is related to how accessing NAND strings or wordline layers. Specifically, in VG-type 3D NAND wordlines run vertically and are common to all layers; however, flash cell strings run horizontally on each layer, and each string of cells must be connected to metal bitlines running at the top. On the contrary, in VC-type NAND the horizontal wordlines need to be connected to the decoders. In both cases the connection to each layer is the key step that has an impact on the array efficiency.

Figure 7.8 shows an example of the array architecture in a 3D NAND. Each array block needs a special structure for connecting the horizontal layers (the strings in the VG-type NAND) to the top metal lines. From the pure area point of view it is necessary to keep this overhead as small as possible and in general to avoid a linear increase of its area with the number of layers. Another overhead is represented by the WL Decoders.

A method proposed for connecting the horizontal strings to the main bitline of a VG-type NAND is shown in Fig. 7.9 [18]. All the strings from each layer are simply connected together by vertical poly plugs. In order to select the proper string (layer), a series of SSL is realized in each string. Some of the SSL transistors need to be normally-ON SSL transistors. Each string can be selected by applying proper biasing to the SSL gates, as shown in the decoding table. The series of SSL constitutes the overhead of the structure. If the number of layers is increased, more SSL series transistors are required and this will increase the overhead area. Moreover, the depletion type SSLs require special implantation steps.

Another straightforward method to access the layers is to realize a connecting “staircase” structure. Each polysilicon step of the staircase is connected to one of the layers and it can be connected to the top metals by landing a contact down to the step.

Figure 7.10 shows a progressive “etch and trim” method for realizing the staircase structure [1]. The method consists in multiple etching steps with

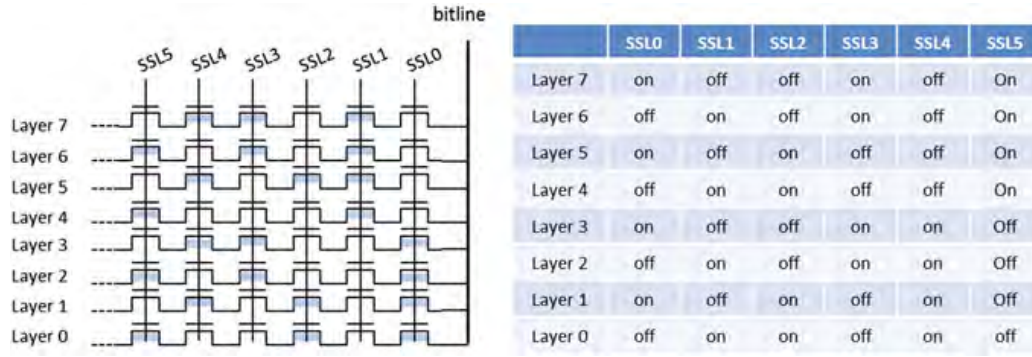


Fig. 7.9 Layer selection in VG-NAND with multiple SSL [18]

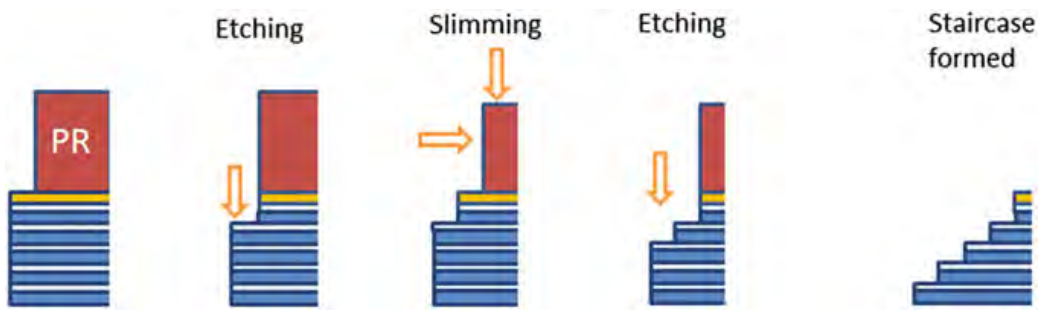


Fig. 7.10 Fabrication method for staircase formation in 3D NAND [1]

photoresist trimming until each layer is exposed and a contact can land on it later on. This method could theoretically be realized with one single lithographic mask; however, the control of this process might become very difficult as the number of layers grows up.

Trim and etch method increases the connecting structure area linearly with the number of layers and, therefore, it is not very efficient.

Figure 7.11 shows another simple method for connecting wordlines in VSAT architecture [5]: layers for multiple gates are deposited over the substrate *mesas* and simultaneously form the “PIPE” structure at the edge, without any additional complex fabrication step. Also in this case the space required to connect the layers is proportional to the number of layers itself.

It has been shown by different sources that the linear increase of die area to build staircase contacts is not cost effective. More in general, the process costs cannot increase linearly with the number of layers [27]. The staircase contact structure is critical because a precise landing of the contacts, and also because it requires multiple process steps including lithographic steps. The tradeoff between complexity of the process flow and the die size must be found.

A proposal to use a *Minimum Incremental Layer Cost* (MiLC) for 3D VG NAND [10] is shown in Fig. 7.12. The proposal is a lithographic method for realizing the staircase structure with minimum lithography steps. When increasing

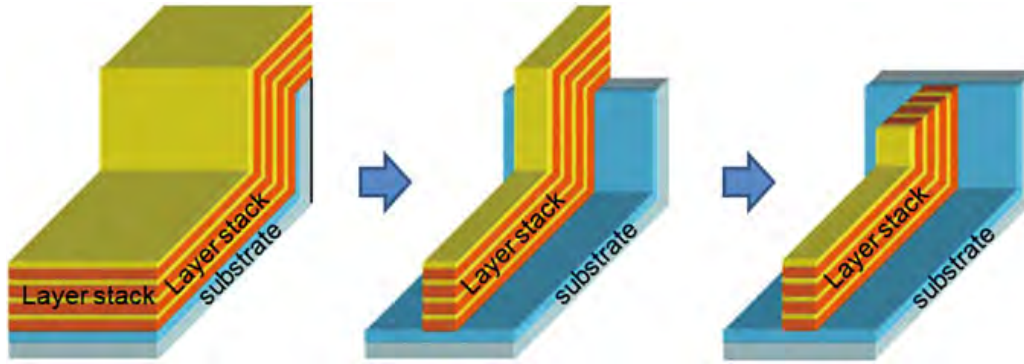


Fig. 7.11 PIPE WL connection in VSAT [5]

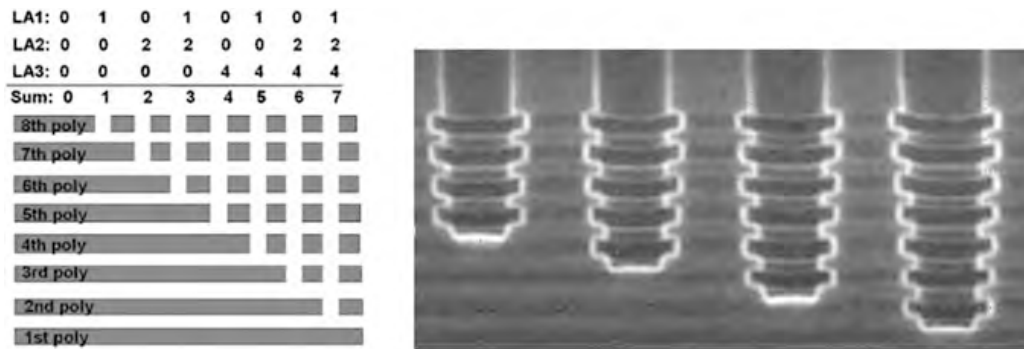


Fig. 7.12 Formation of the MiLC staircase contacts in 3D-VG NAND [10]

memory layers, the array efficiency is kept constant due to the special structure. The cost due to the additional masks and process steps is a logarithmic function of the number of layers, instead of a linear function. The log increase of the cost is sustainable in the overall cost structure when the number of layer is increased. For 8 layers, a superimposition of three masks LA1, LA2, LA3 and related etching steps allow drilling down to each polysilicon layer. The final staircase result is shown in the figure, where the precise landing on the poly pads is shown. An isolation spacer is deposited for isolating the contact from the unwanted layers along the hole. With MILC it is possible to double the number of layers and contacts by adding only one mask.

Even if the polysilicon bitline pad in the VG-TYPE 3D NAND structure can be connected efficiently to the top metal through the MiLC approach, it is still necessary to select one single string out of the many connected to the same poly pad.

A proposal for string selection has been done by using PN diodes as select device, as displayed in Fig. 7.13 [19]. PN diodes can be self-aligned to the source side. Each layer has horizontal source line that can be biased independently (e.g. by using a staircase structure), while the drain side of the strings is vertically connected by poly plugs. String selection is done by proper biasing of horizontal *Source Lines* (SL) and bitlines.

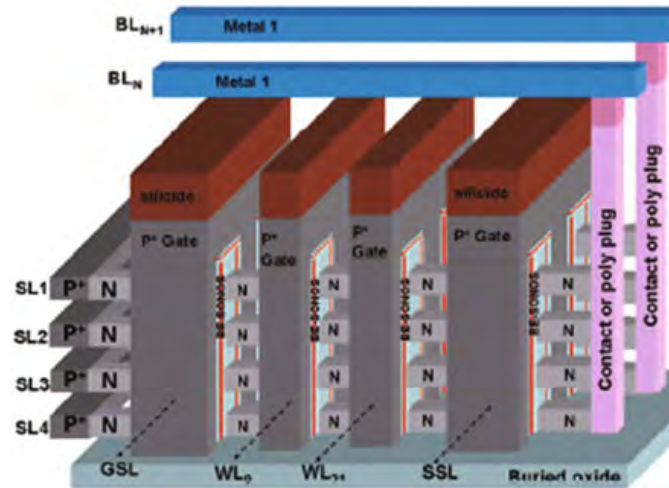


Fig. 7.13 PN-diode decoding structure in 3D-VG NAND [19]

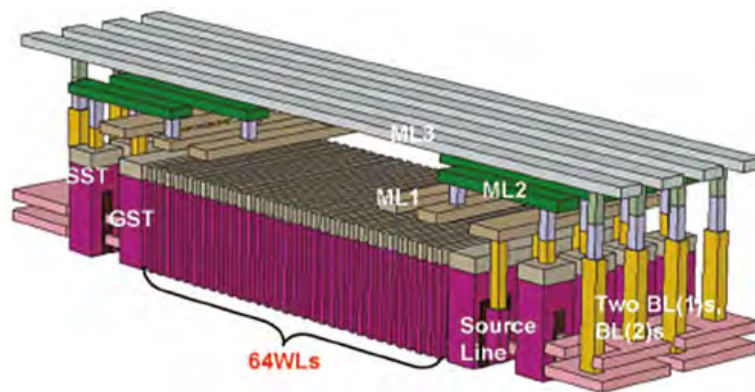


Fig. 7.14 Simplified 2 layer split page VG NAND [28]

A single island-gate SSL selection for VG-type 3D NAND has been realized to individually identify the strings connected to the same poly pad. In order to relax the pitch of SSLs to twice the pitch of bitlines, a split-bitline layout has been proposed [10] and it is shown in Fig. 7.14. The split BL layout architecture consists in having SSL of even and odd strings on the opposite sides of the block; thus, the string current flows in opposite directions for even and odd strings. The description of this architecture is done in the next section. The split-bitline approach improves the process window as SSL are realized with double pitch compared to bitlines. BL contacts are also at double pitch of the channel BL.

In 3D NAND, the string select transistor and the ground select transistor gate dielectrics may make use of the cell's trapping dielectric for simplicity of the process architecture. In this case, the select transistors may suffer unwanted charge injection that causes V_{TH} shift during operation. For example, holes injection during erasing may lower the V_{TH} of the select transistor. If V_{TH} is lowered, then it

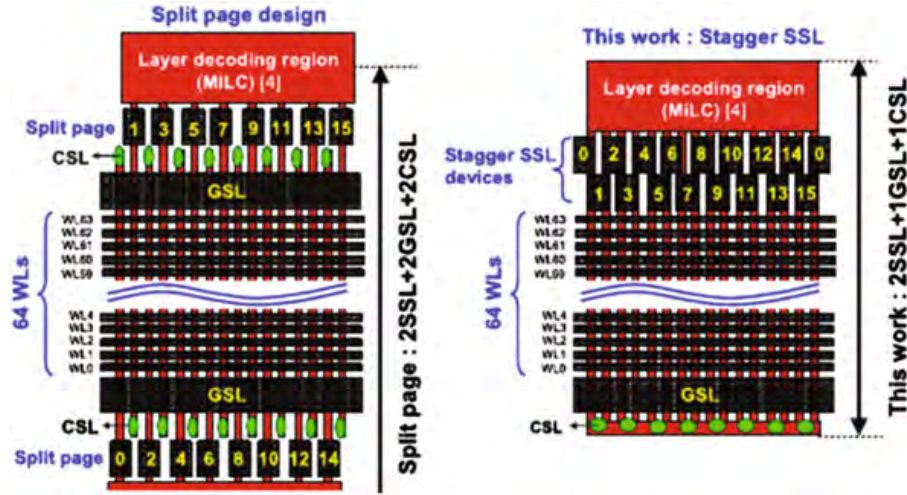


Fig. 7.15 Optimization of split-page string decoding in 3D-VG-Type NAND [26]

is not possible to properly isolate the strings during self-boosting (programming), thus causing failures. To avoid the malfunction, it is necessary to properly bias the select transistor for deselection, by using slightly negative voltages. Moreover, it might be possible to implement V_{TH} adjustments of the select transistors by special algorithms. Because the use of negative voltages complicates the decoding of the control lines, the use of charge-trapping free gate dielectric has been proposed [29]. Charge trapping free dielectric can be realized by removing the trapping oxide (O3) in the BE-SONOS Stack (O1/N/O2/N/O3) and replacing it with an additional oxide (O4) to reinforce the gate dielectric.

In order to reduce the overhead of the additional CSL and GSL lines of the split-page design and optimize the space occupied by the island-gate SSLs, a new decoding scheme with staggered SSL was proposed and it is shown in Fig. 7.15. In this arrangement, each SSL transistor selects 2 strings and a combination of 2 SSL in series is used to decode one string out of 16. Having SSL only on one side has also the benefit of eliminating one GSL and one source line connecting area, thus saving additional space; the resulting array efficiency is close 80 %, similar to that of 2D NAND. By using staggered SSL arrangement, VG-type 3D NAND becomes the most efficient among the 3D NAND solutions.

7.5 VG-Type 3D NAND Array Operations

In this section we will analyze the principal working modes of a VG-type 3D-NAND structure. 3D VG NAND with split page architecture [10] will be used for this purpose. In order to simplify the understanding of the 3D structure, an example with only two layers will be used. In Fig. 7.16 a 3D drawing of the two-layer structure is shown. The figure shows a structure having two basic units

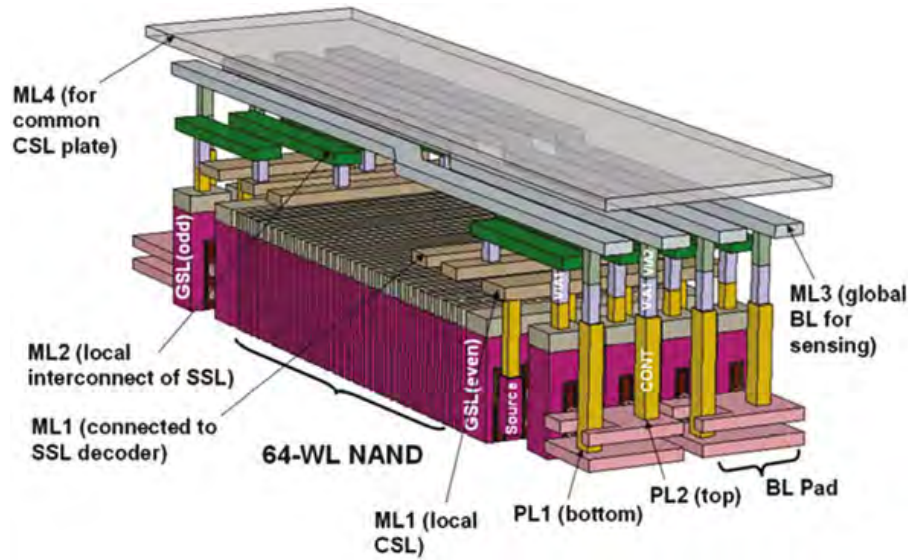


Fig. 7.16 2-layers structure example (2 units) [28]

side by side. Each unit is composed of four strings for each layer. Due to the split page architecture, SSLs of even and odd strings are placed on the opposite sides of the unit, thus allowing larger process window for SSL formation. Each single unit contains 4 pages that can be accessed from the two M3 bitlines. Each page contains cells that are on both layers. In the general case, each page contains N cells belonging to all N layers. Staircase bitline contacts are fabricated for the array decoding by using the MiLC approach [10]. ML1 and ML2 are used to decode the eight SSL devices. The source contact is built at the line end of each channel, and it is directly connected to a local ML1 common source line (local CSL). A top Metal layer can be used to connect M1 local CSL to reduce the overall resistance. Each memory cell is accessed by selecting the corresponding WL, ML3 BL (corresponding to the memory layer), and SSL (corresponding to one channel BL).

Page programming and reading are performed by simultaneously operating SSLs in parallel, in different units for a larger page. A simplified schematic of this basic unit will be used in the next sections to describe the 3D NAND flash operations.

7.5.1 Read Operation

In Fig. 7.17 it is shown a simplified circuitual schematic of 3D-VG-type Flash array. There are two layers and four strings for each layer, in a split page organization. The polysilicon BL pad structure is visualized in the schematic for a better understanding of the architecture. Reading a Flash cell is achieved by using read schemes as in a standard 2D NAND. Forward read sensing can be used also in 3D NAND; the Flash string must be properly biased in order to read out the current

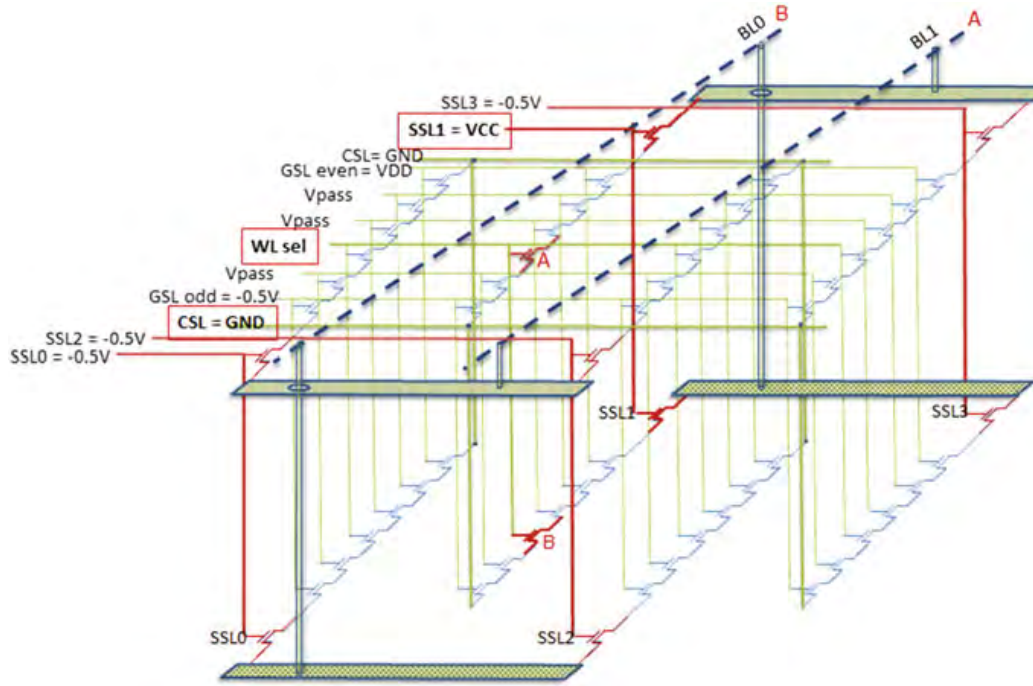


Fig. 7.17 3D VG-type NAND read biasing

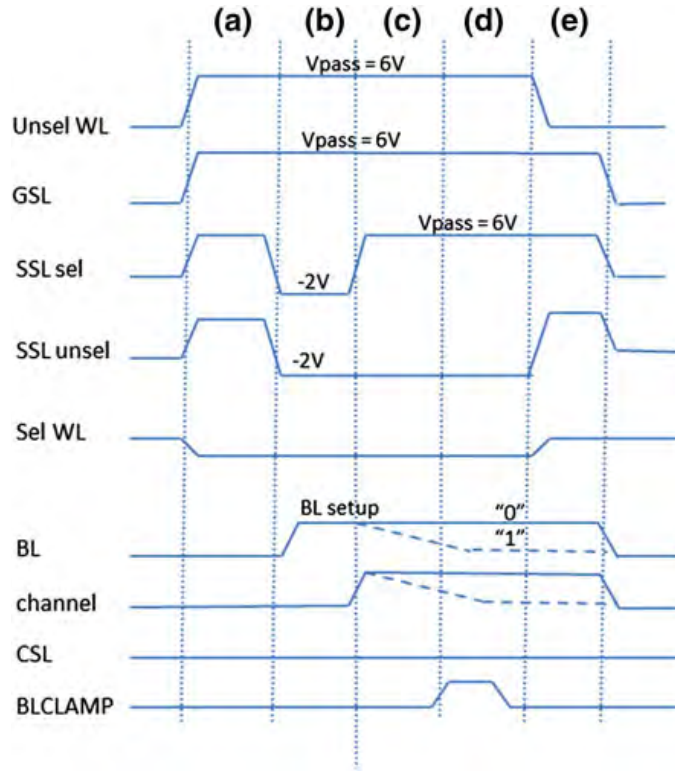
value of the selected cell. The forward-read sensing method is not described here but is deeply studied in literature [30].

Reading a cell requires the corresponding SSL to be turned ON with VCC voltage, while other SSLs need to be turned OFF. In the example of Fig. 7.17, the selected cells are A and B of Page1. The corresponding SSL1 is at VCC. Turning OFF of other SSLs is achieved by applying a negative voltage, e.g. -0.5 V, so that SSLs are turned OFF even if their V_{TH} has drifted. The VGSL even transistors are turned ON by VCC and the selected wordline stays at read voltage of 0 V or even slightly negative. All other wordlines are turned ON at the Vpass voltage of about 6 V. As shown by labels, cell ‘B’ is read out through bitline BL0, while cell ‘A’ is read out through bitline BL1.

Figure 7.18 shows a detailed timing diagram of the read phase with forward voltage sensing. To avoid “local self-boosting” at the cells between SSL and selected WL, that may lead to hot-carrier induced read disturb, a proper discharge of the channel is done by pulsing SSL during wordlines turn on [28]. The selected wordline is biased slightly negative during the read phase.

Other read methods such as “reverse read” can be used as well in VG-type 3D NAND. Reverse read can be useful to cope with specific characteristics of the VG-type 3D NAND described in this section. In VG-type 3D NAND, cells belonging to each of the layers are in the same page and, hence, are read out at the same time. Every layer is connected to a separate M3 bitline, but every layer also share the same SSL select line in the vertical direction. Due to the etching profile of the stack there are geometrical differences among layers. As a consequence, the

Fig. 7.18 Forward read with bitline discharge to avoid read disturb [28]



V_{TH} distribution of a specific layer is intrinsically different from the one of other layers, and each layer also shows a different BL bias dependence during read. The proposed layer-aware read method uses reverse read with different judging level for each bitline, to compensate for the layer differences.

Figure 7.19 shows a schematic example of the reverse-read sensing with multi V_{TH} reading [28]. BL compensation consists in biasing the BLCLAMP transistors of the page buffer at a proper voltage, depending on the layer being read. Reverse read consists in the following sequence: firstly, bitlines BL are discharged to Ground (GND) through the BLBIAS transistor; afterwards, CSL is biased at a positive voltage (reverse read) and BLC is turned OFF to isolate the sensing node SEN from the bitline; WL is raised to the Read or Program Verify voltage (in the example the read voltage), then SSL is turned ON, and BL capacitance charges through the cell and CSL until it reaches voltage $V_{WL}-V_{THC}$, where V_{THC} is the selected cell's threshold voltage. During the sensing phase, BLC is set to the discriminating voltage V_{BLCL} (which varies according to bitline/layer L). If the cell's V_{TH} , V_{THC} , is lower than the target V_{TH} , then the BL voltage is higher than $V_{BLC}-V_{THBLC}$, thus the BLC transistor is OFF and the voltage of SEN node remains high. Otherwise, if the V_{THC} is equal or greater than target V_{TH} , then the BLC transistor is ON, SEN discharges, and the sense amplifier output switches to '0' value.

The proposed method for read and verify can optimize the reading conditions based on the characteristics of the cells in each layer. Therefore, it is possible to

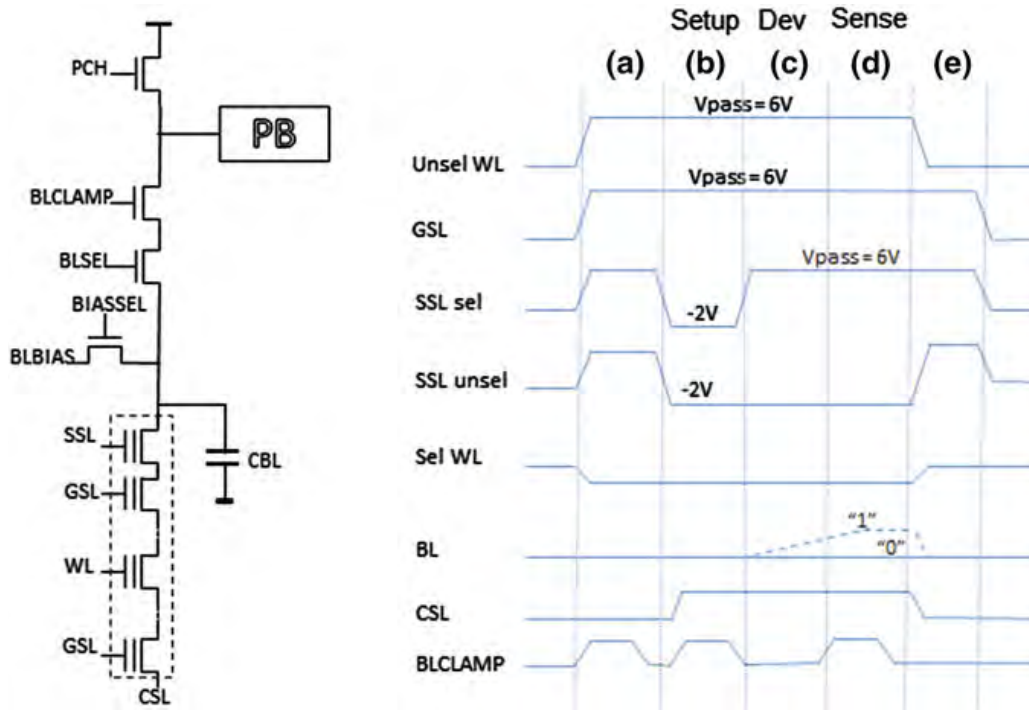


Fig. 7.19 Reverse read operation with BL clamp compensation [28]

optimize the target Program Verify PV for each layer instead of using a common PV level; the programming time can significantly be reduced, and also the program disturbance is reduced by avoiding unnecessary over programming of the lower layers. A drawback of the layer-aware read/verify method is the difference in the bitline voltage swing between layers. The top layer will have higher V_{TH} than other layers and, therefore, the greatest BL swing. This happens because, during reverse read, the BL is charged from GND to the target voltage, which is different for each layer/bitline. The layers with the higher operating point will result in higher background pattern dependency. In order to level this kind of dependency, a method was proposed where each bitline is pre-charged to specific layer-dependent voltage instead of discharging all the bitlines to GND during the setup phase [31].

7.5.2 Program Operation

Programming 3D-VG NAND is obtained by optimized ISPP (*Incremental Step Programming Pulse*) algorithms. In Fig. 7.20 it is shown the equivalent circuit to illustrate the programming of page 0 (SSL1) and the corresponding program-inhibit method used on the other cells not being programmed. When programming page 1, the select line SSL1 is at VCC, while a slightly negative voltage is applied to other SSLs to guarantee turn-off. GSL (even) is turned-on, while GSL (odd) is turned-off.

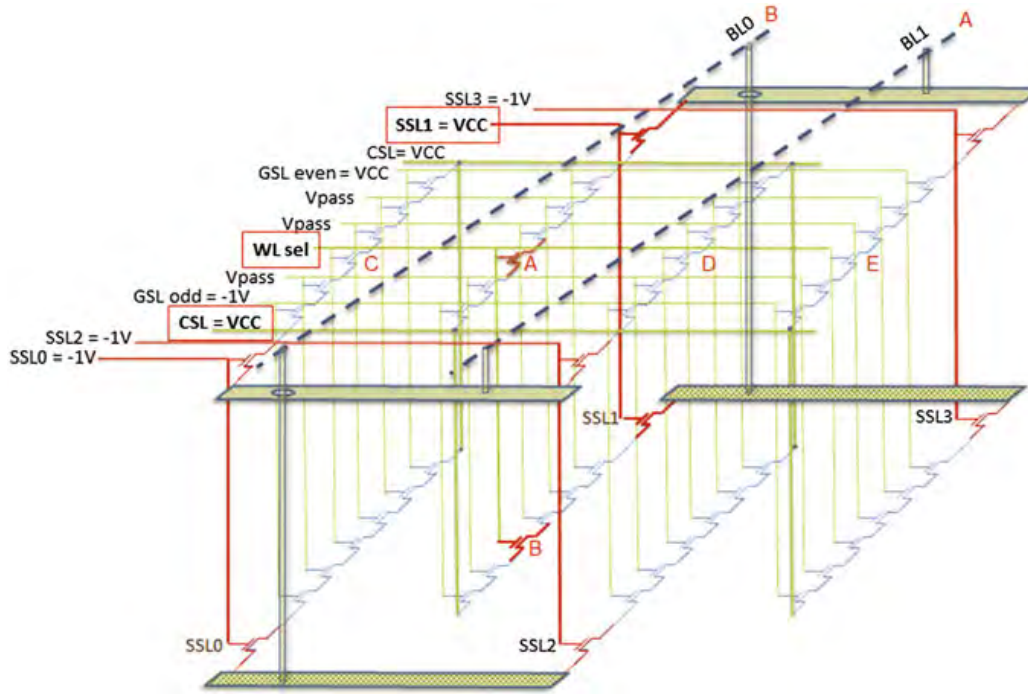


Fig. 7.20 3D VG-type program biasing

CSL is set to VCC in order to perform the inhibit operation. The selected wordline is biased with the programming voltage according to the ISPP scheme, while other wordlines are biased with Vpass voltage. Program Inhibit of unselected pages is achieved by using different methods. The cells belonging to Page 3 (SSL3) bitlines (e.g. Cell 'E') are inhibited by entirely floating the NAND string, because SSL3 and GSL (even) are both deselected; the cells belonging to even pages (SSL0 and SSL2), i.e. Cells 'C' and 'D', are inhibited by charging the channel via CSL and GSL, which are both at VCC [10].

7.5.3 Erase Operation

A *Barrier Engineered SONOS* (BE-SONOS) device is adopted in 3D VG-type NAND. Erasing of BE-SONOS devices is achieved by holes injection from the channel. BE-SONOS can overcome Electron Gate injection due to the high electric field during erase and reduce erase saturation with an excellent retention. Vertical Channel devices, which usually have a gate-all-around structure, can benefit from the *Electric Field Enhancement* (FE) due to the curvature effect which reduces the electric field through the gate, thus mitigating the electrons injection from the poly gate. However, FE method has several drawbacks and imposes limitations on the cell size [32, 23]. Another way to reduce gate injection would be through the use of Hi-K material which can minimize the electric field in the oxide. BE-SONOS

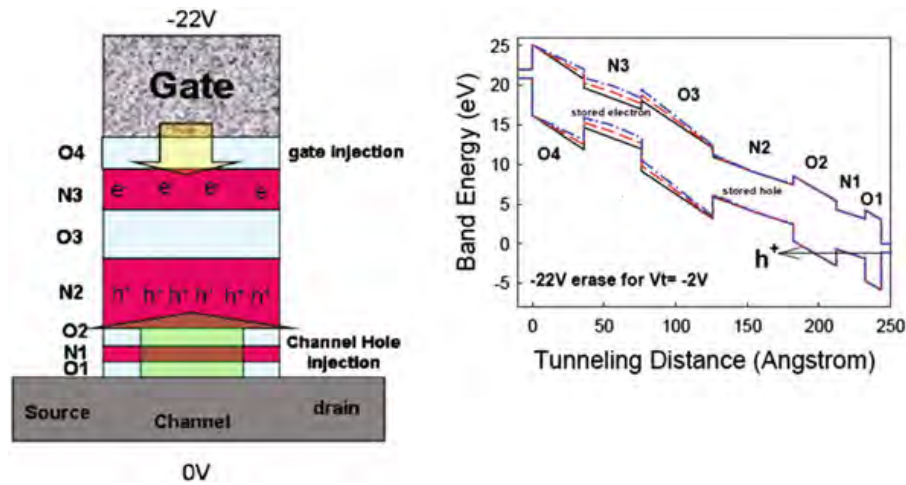
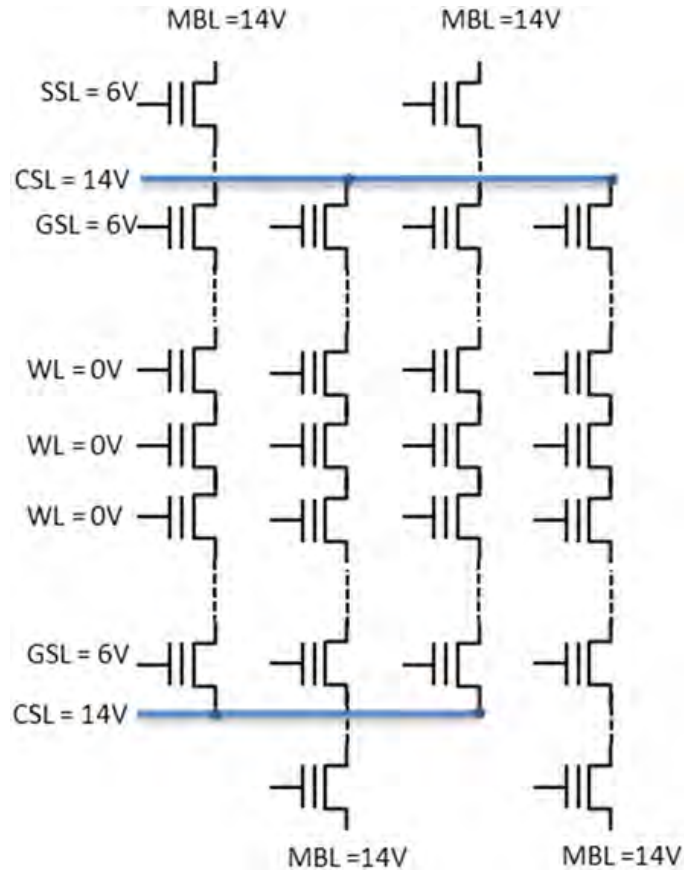


Fig. 7.21 Erase Mechanism in a BE-SONOS CT device [23]

devices allow avoiding the complication and drawbacks of FE or the use of Hi-K materials. In 3D-VG NAND, further engineering of the BE-SONOS structure is possible, thanks to the planar and very uniform structure [23]. In Fig. 7.21 erase through holes injection from the substrate in a BE-SONOS device and the parasitic electron injection from the polysilicon gate are shown. The displayed structure shown is BE-SONOS with additional nitride and oxide layers specifically added to stop electron injection at the additional nitride layer.

In 3D-NAND a junction-less cell with floating body is generally employed; therefore, there is no holes source for erasing from the substrate. Holes generation for erase can be obtained by GIDL (*Gate-Induced Drain Leakage*) at the select transistors. Figure 7.22 shows that erase is achieved by applying a large voltage to the CSL (*Common Source Line*) and MBL (*Main Bitline*), while an intermediate voltage of about 6 V is applied to the SSL (*String Select Line*) and GSL (*Ground Select Line*). The select transistors should be biased to pass the erase voltage (13 V) from the source lines to the channel; at the same time, the pass transistors must be guarded against parasitic charge injection that may damage them or make their threshold voltage drift. With appropriate biasing of SSL and GSL, the GIDL induced current that is built up at the drain junctions of the select transistors can transfer the erasing voltage to the string channel and minimize the unwanted hole injection that would result in erasing of the SSL and GSL themselves. To Erase a Block in 3D-VG Type NAND, a high voltage is applied to the source lines while the wordlines are grounded. To inhibit erasing of the unselected blocks, the wordlines are left floating so that they are self-boosted to a inhibit voltage. It is important to reduce any leakage on the wordlines that could reduce the efficiency of the boosted potential and, hence, the erase inhibit.

Fig. 7.22 VG-type block erase diagram (selected block)



7.6 Disturbs of VG-Type 3D NAND

The most peculiar disturbs and interference effects in 3D NAND are along the vertical (Z) direction, which does obviously not exist in 2D architectures. In VG-type 3D NAND, bitlines are completely shielded by wordline polysilicon plugs; therefore, there is no electrostatic interference in the X-direction (along the wordline direction) or even in diagonal direction between different layers, as shown in Fig. 7.23. Instead, Z interference can be observed between adjacent vertical layers: it is a back-gate bias effect caused by electrostatic interference between adjacent cells. The electrostatic interference is caused by electrons stored in the trapping layers during programming: the V_{TH} shift of the adjacent victim layer can be in the range of 150 mV. The interference is obviously dependent on the inter-layer distance and it can be optimized [24, 33].

Wordline interference has also been observed in VG-type 3D NAND when programming adjacent wordlines in the same layer due to the tight pitch. The WL interference is associated to channel potential variations in the junction free channel of the selected cell, which are induced by the charge accumulated in adjacent pass cells. WL interference results in 400 mV V_{TH} shift; this shift can be contained by proper algorithmic countermeasures. Programming algorithm with Pre-PV

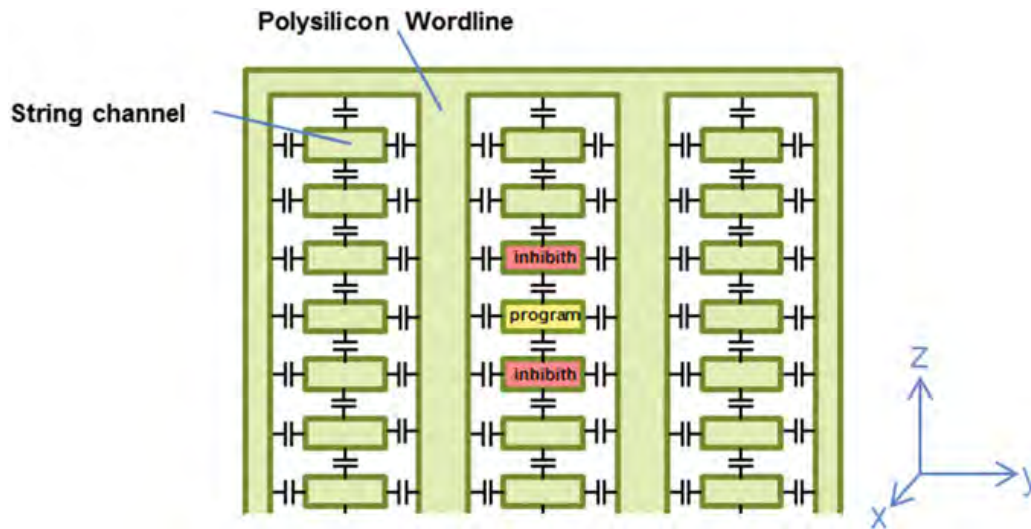


Fig. 7.23 Z disturbs in VG-type 3D NAND

(*Pre-Program Verify*) level can be used to alleviate WL-interference. The algorithm consists of programming WL(n) to a pre-PV level lower than the target PV level, then programming WL(N + 1) to the pre-PV level, going back to program WL(n) to the final PV value, and so on.

Since all the layers are programmed simultaneously, as part of the same page in 3D-VG NAND, the interference and other effects caused by the programming pattern must be considered.

A program disturb effect has been studied. As shown in Fig. 7.23, when a certain layer is programmed, the related channel is kept at 0 V, while adjacent channels, if not programmed, are boosted to the inhibit voltage. Due to the capacitive coupling between layers, the resulting boosted value is lowered with respect to the case where all the channels are in the inhibit condition. The inhibit function is therefore degraded and a program disturb can be observed. The threshold shift due to Z-disturb can result in a shift of 0.6 V of the V_{TH} distribution, because of the reduced inhibit efficiency [24]. Moreover, when different layers are programmed at different times, an impact of the first programmed cells on the following programming operations can be observed, due to the electrostatic potential change. This effect is substantially changing the V_{TH} distribution.

Another Z-effect, called enhanced programming Z-disturb, can be observed in VG-type 3D NAND. When adjacent layers are programmed (biased to 0 V), the programming speed is reduced with respect to the case where adjacent layers are inhibited (boosted to inhibit high voltage). This effect is explained by the inversion electron density function at the polysilicon channel that is increased in the second case, while in the first case it limits the F-N tunneling current.

It is necessary to consider the impact of the Z-effects when the programming pattern changes during the ISSP steps. The described Z-disturbs will cause the

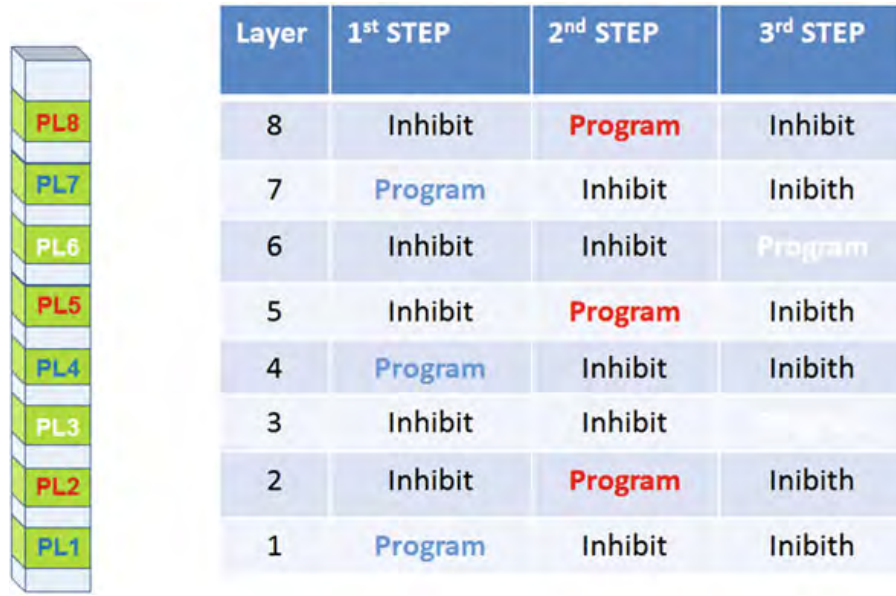


Fig. 7.24 An algorithmic programming method to level out Z-disturbs [24]

enlargement of the programmed distribution and reduce the margin for multilevel operation.

Z-interference (electrostatic) effects are obviously dependent on the thickness of the buried oxide between layers; therefore, they are also controllable, to some extent, by augmenting the inter distance between layers [8]. Algorithmic methods can also be adopted to control the Z-disturbs. In the example shown in Fig. 7.24 the programming sequence is split into a sequence involving only certain layers at a time. In this example eight layers are divided into three groups and programming is accomplished in three programming steps. During each of the programming steps, each inhibited layer has only one adjacent layer (in Z-direction) which is programmed, thus limiting the interference to one neighbor only. At the same time, each programmed layer has two inhibited adjacent layers on both sides so that the enhanced Z-programming effect is constant [25].

Vpass disturb and Read disturb are also critical in 3D Flash. In VG-type 3D Flash there are $2N$ pages per each wordline, where N is the number of layers. Thus programming the entire block results in a very large Vpass disturb time. Read disturb is also worsened due to the large number of pages in a WL.

Compared to Vertical Channel nanowires type of 3D NAND, substantially, the effect of disturbs is similar. In VC-type 3D NAND the Field Enhancement Effect allows achieving shorter programming time; however, enhanced field increases the Vpass disturbance. On the contrary, the planar structure of VG-type 3D NAND can minimize the Vpass disturb.

By implementing proper algorithmic sequences, as described in this section, it is possible to obtain a V_{TH} window large enough for MLC operations, as shown in

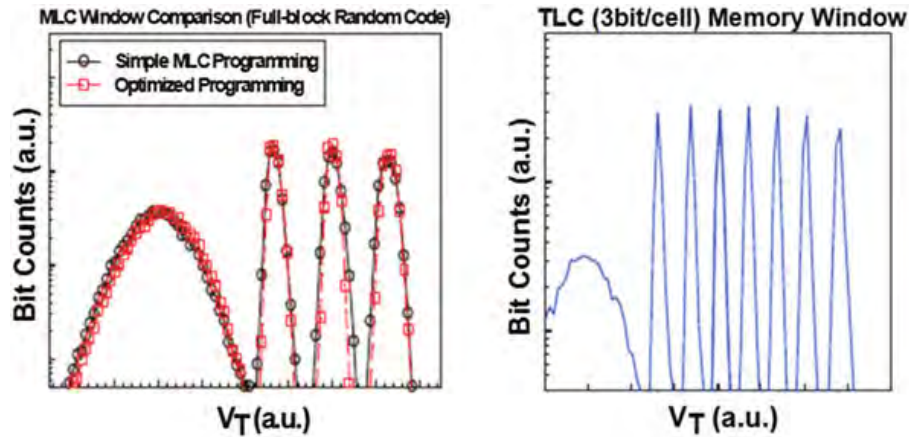


Fig. 7.25 MLC and TLC capability of VG-type 3D NAND [24]

Fig. 7.25. TLC window has been also demonstrated. With proper optimization of the structure and Error Correcting techniques, VG-type 3D NAND is a viable candidate for realizing high density products.

7.7 Conclusions

Vertical Gate (VG-type) 3D NAND technology capability has been demonstrated with several studies and test chips. The technology can achieve cost effectiveness through multiple memory layers and multilevel operations. The planarity of the CT SONOS Flash cell is the key element because it avoids all the issues of the non-ideal vertical etching required by Vertical Channel (VC-type) 3D NAND architectures based on gate-all-around structures.

Acknowledgements The author would like to thank Macronix International for the use of publications which have been cited in this chapter.

References

1. 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)
2. R. Katsumata, M. Kito et al., Pipe-shaped BiCS flash memory with 16 stacked layers and multi-level-cell operation for ultra high density storage device, in *Symposium on VLSI Technology* (2009)
3. K. Sakuna et al., Highly scalable horizontal channel 3-D NAND memory excellent in compatibility with conventional fabrication technology, in *EDL*, vol. 34 (2013)
4. J. Jang, H.S. Kim et al., Vertical cell array using TCAT(terabit cell array transistor) technology for ultra high density NAND flash memory, in *IEDM* (2009)

5. J. Kim, A.J. Hong 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 *Symposium on VLSI Technology* (2009)
6. K.T. Park, D.S. Byeon, A world's first product of three-dimensional vertical NAND Flash memory and beyond, in *NVMTS* (2014)
7. 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, in *ISCC* (2014)
8. 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)
9. H.T. Lue, T.H. Hsu et al., A novel double-density single gate vertical channel (SGVC) 3D NAND that is tolerant to deep vertical etching CD variation and possesses robust read-disturb immunity, in *IEDM* (2015)
10. S.H. Chen, H.T. Lue et al., A highly scalable 8-layer vertical gate 3D NAND with split-page bit line layout and efficient binary-sum MiLC (minimal incremental layer cost) staircase contacts, in *International Electron Device Meeting (IEDM)*, session 2-3 (2012)
11. S.J. Whang, K.J. Lee et al., A novel three-dimensional dual control-gate with surrounding floating-gate (DC-SF) NAND flash cell, in *IEDM* (2010)
12. E.S. Choi et al., A novel 3D cell array architecture for terra-bit NAND flash memory, in *IMW* (2011)
13. E.S. Choi, S.K. Park, Device considerations for high-density and highly reliable 3D NAND Flash cell in near future, in *International Electron Device Meeting (IEDM)*, session 9-4 (2012)
14. K. Parat, C. Dennison, A floating gate based 3D NAND technology with CMOS under array, in *IEDM 2015*
15. J. Choi, K.S. Seol, 3D approaches for non-volatile memory, in *VLSI* (2011)
16. K.T. Park, J.M. Han et al., tree-dimensional 128 Gb MLC vertical NAND flash memory with 24WL stacked layers and 50 MB/s high speed programming, in *ISSCC* (2014)
17. Y. Fukuzumi, R. Katsumata, Optimal integration and characteristics of vertical array devices for ultra-high density, bit-cost scalable flash memory, in *IEDM* (2007)
18. W. Kim, S. Choi et al., Multi-layered vertical gate NAND flash overcoming stacking limit for terabit density storage, in *Symposium on VLSI Technology* (2009)
19. C.H. Hung, H.T. Lue et al., A highly scalable vertical gate (VG) 3D NAND flash with robust program disturb immunity using a novel PN diode decoding structure, in *VLSI Symposia on Technology*, session 4B-1 (2011)
20. C.P. Chen, H.T. Lue et al., 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), in *VLSI Symposia on Technology* (2012)
21. K.P. Chang, H.T. Lue et al., An efficient memory architecture for 3D vertical gate (3DVG) NAND flash using plural island-gate SSL decoding and study of its program inhibit characteristics, in *International Memory Workshop (IMW)* (2012)
22. T.H. Yeh, P.Y. Du, Increasing VG-type 3D NAND flash cell density by using ultrathin polysilicon channels, in *IMW* (2013)
23. H.-T. Lue, R. Lo, A novel double-trapping BE-SONOS charge-trapping NAND flash device to overcome the erase saturation without using curvature-induced field enhancement effect or high-K (HK)/metal gate (MG) materials, in *IEEE IEDM* (2014)
24. C.C. Hsieh, H.T. Lue et al., Study of the interference and disturb mechanisms of split-page 3D vertical gate (VG) NAND flash and optimized programming algorithms for multi-level cell (MLC) storage, in *VLSI Symposia on Technology*, session 11-3 (2013)
25. W.C. Chen, H.T. Lue, Study of the programming sequence induced back-pattern effect in split-page 3D vertical-gate (VG) NAND flash, in *VLSI-TSA* (2014)
26. T.H. Yeh, C.J. Wu et al., A new string decoding scheme for enhancing array block efficiency of vertical gate type (VG-Type) 3-D NAND. *EDL* **36**(4), 2 (2015)
27. H.T. Lue, Tutorial 3D NAND, in *IMW* (2014)

28. C.H. Hung, H.T. Lue et al., Design innovations to optimize the 3D stackable vertical gate (VG) NAND flash, in *International Electron Device Meeting (IEDM)*, session 10-1 (2012)
29. Y.-R. Chen, C.-J. Wu, Trapping-free string select transistors and ground select transistors for VG-type 3D NAND flash memory, in *IMW* (2014)
30. R. Micheloni et al., *Inside NAND Flash Memories* (Springer, Netherlands, 2010)
31. C.H. Hung, M.F. Chang et al., Layer-aware program and read schemes for 3D stackable Vertical-Gate BE-SONOS NAND flash against cross-layer process variations. *IEEE JSSC* **50**(6) (2015)
32. A. Maconi, C.M. Compagnoni et al., A new erase saturation issue in cylindrical junction-less charge-trap memory arrays, in *IEDM* (2012)
33. Y.H. Hsiao, H.T. Lue, A critical examination of 3D stackable NAND flash memory architectures by simulation study of the scaling capability, in *IMW* (2010)
34. E.S. Choi, H.-S. Yoo, A novel 3D cell array architecture for terra-bit NAND flash memory, in *IMW* (2013)