INVITED PAPER

Reviewing the Evolution of the NAND Flash Technology

This paper reviews historical trends of NAND Flash technologies, explaining why scaling of planar arrays below 1x nm is less favorable than vertical integration.

By Christian Monzio Compagnoni, *Senior Member IEEE*, Akira Goda, Alessandro S. Spinelli, *Senior Member IEEE*, Peter Feeley, Andrea L. Lacaita, *Fellow IEEE*, and Angelo Visconti, *Member IEEE*

ABSTRACT | This paper reviews the recent historical trends of the NAND Flash technology, highlighting the evolution of its main parameters and explaining what allowed it to become not only the most important integrated solution for nonvolatile storage of high volumes of data but also a strong rival eroding the market share of hard-disk drives. The scaling trend followed by planar arrays will be discussed with close attention, along with the major physical constraints impacting the performance and the reliability of modern deca-nanometer technologies. This will make clear why the development of further planar nodes with feature size below ~15 nm, representing today's state of the art, can be considered less favorable than turning all the efforts toward the integration of 3-D arrays. The most promising 3-D architectures will then be reviewed, discussing their benefits and issues and addressing the impact of the change of the integration paradigm from the standpoint of the major NAND applications.

KEYWORDS | Flash memories; Moore's law; NAND Flash technology; semiconductor device reliability; solid-state drives

I. INTRODUCTION

The NAND Flash technology represents today the most prominent integrated solution for the nonvolatile storage of high volumes of data. The success of this technology is the result of its uninterrupted scaling since its introduction in the late 1980s [1], which has recently resulted in the possibility to outperform magnetic recording on the platters of hard-disk drives (HDDs) not only in terms of volumetric but also of planar storage density, breaking the 1-Tb/in^2 barrier [2]–[6]. This march allowed the NAND technology to constantly reduce its cost per gigabyte (GB), addressing ever new applications and markets. Starting from portable devices, requiring high volumetric storage density, mechanical shock immunity, and low power consumption, and passing through solid-state drives (SSDs) offering an extremely high number of random read/write input/output operations per second (IOPS) and low latency, the technology seems now ready to target enterprise storage and big data applications, thanks to its cost becoming not far from that of high-performance HDD systems. The never ending evolution of the technology along with the increase of its pervasiveness in everyday life and business applications will surely make it the most important memory solution in the next decade.

In this paper, the recent historical scaling trends of the planar NAND Flash technology will be reviewed, addressing, first, the evolution of the technology node feature size and of the bit storage density over the years. Besides, the trends of array capacity, page size, throughput, and chip area will be discussed in detail to provide a comprehensive picture on how the scaling process affected technology performance, design, and manufacturing. Along with this discussion, reference will be made also to some major NAND Flash applications, such as SSDs, addressing their performance parameters and requirements. Then, the process-related and physical issues setting the most relevant constraints to the performance and reliability of state-of-the-art deca-nanometer technologies will be reviewed, explaining why these issues have made a further miniaturization of the planar array less favorable than moving to 3-D integration schemes. The vertical-channel structure will then be discussed as the most promising 3-D array architecture to keep the historical growth rate of the technology storage density through the next decade. Finally, the performance, limitations, and prospects of this integration solution will be addressed, along with the impact of the transition from planar to 3-D arrays on the most relevant NAND applications.

Manuscript received October 26, 2016; revised January 11, 2017; accepted January 27, 2017. Date of publication March 28, 2017; date of current version August 18, 2017. *(Corresponding author: Christian Monzio Campagnoni.)*

C. Monzio Compagnoni, A. S. Spinelli, and **A. L. Lacaita** are with the Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, 20133 Milano, Italy (e-mail: christian.monzio@polimi.it).
A. Goda is with Process R&D, Micron Technology Inc., Boise, ID 83716 USA.
P. Feeley is with SSD Architecture, Micron Technology Inc., Boise, ID 83716 USA.
A. Visconti is with Process R&D, Micron Technology Inc., 20871 Vimercate, Italy.

Digital Object Identifier: 10.1109/JPROC.2017.2665781

II. NAND FLASH FUNDAMENTALS

In this section, a general introduction on the NAND Flash technology will be given, discussing, first, the array architecture and layout and, then, the program/erase and read operations. Aim of this section is to provide the reader with the background on the NAND Flash technology that is needed to understand the subjects dealt with in the following sections.

A. Array Architecture

A mainstream planar NAND Flash array consists in the regular arrangement of floating-gate transistors schematically depicted in Fig. 1. Such transistors, representing the memory elements of the array storing information in terms of charge in their floating gate, are connected in series along strings and are driven by shared wordlines (WLs) running in the orthogonal direction. WLs are made of either highly doped polysilicon or metal and act as the control gate of the transistors, as shown with the schematic layout of the array in Fig. 2. The memory elements belonging to the same string of the NAND array are integrated on a silicon stripe [Fig. 2(b)] patterned at the wafer surface by means of shallow trench isolations (STIs) [Fig. 2(a)] [8], [9].

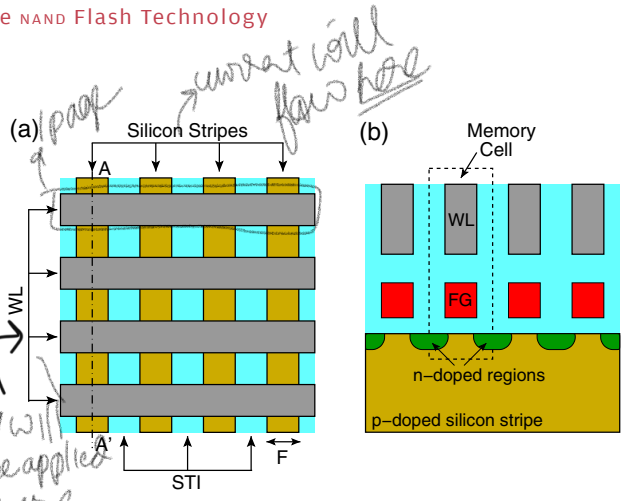


Fig. 2. ***(a) Schematic planar layout of a NAND Flash array. (b) Schematic vertical cross section of the array along a memory string [direction A-A' in part (a) of the figure].***

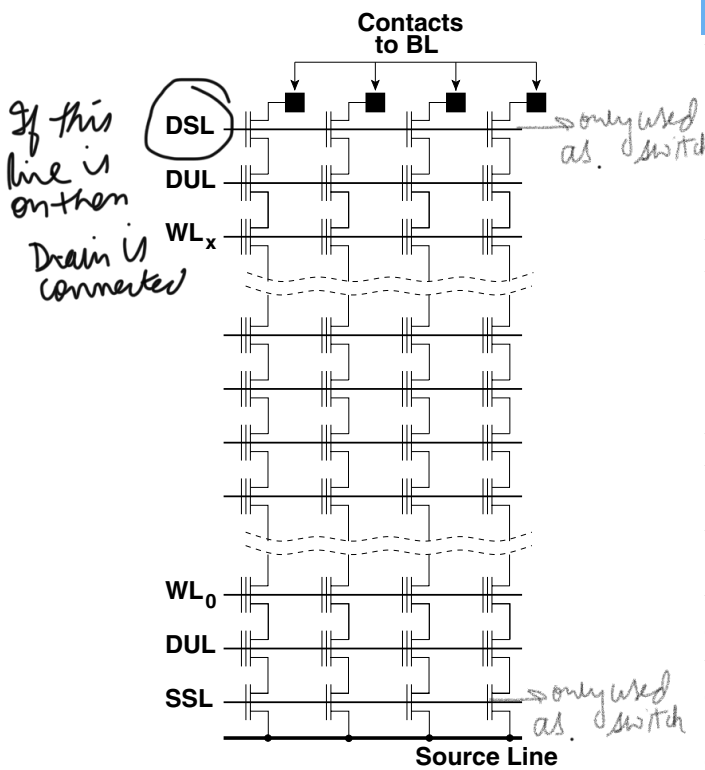


Fig. 1. ***Schematic description of a NAND Flash array, highlighting the strings of floating-gate transistors driven by shared WLs running in the orthogonal direction. Strings are selected by means of a drain select transistor, connecting the string to a BL and driven by a DSL, and a source select transistor, connecting the string to an implanted source line and driven by an SSL. The upper and lower cells in the strings are dummy cells in deca-nanometer technologies [7], driven by dummy lines (DULs).***

The width of the silicon stripes, corresponding to the width W of the floating-gate transistors, is commonly referred to as the feature size of the technology node (F). Typically, the area of a memory cell is nearly equal to $4F^2$, meaning that the pitch of both the WLs and the silicon stripes in Fig. 2(a) is about $2F$, i.e., the minimum value allowed by the litographic and etching processes [9], [10]. This makes the planar net bit storage density (NBSD) of NAND Flash arrays, corresponding to $1/4F^2$ in the case of single bit storage per cell, extremely high, representing one of the reasons of the great success of this technology. Note, in fact, that a high NBSD is a fundamental starting point for any memory technology whose cost and performance are related to the area occupancy of the memory array.

In order to store and access data safely, any memory technology requires additional service elements other than the memory devices. In the case of the NAND Flash technology, the number of the service elements and their burden on the chip area is relatively low, representing another key advantage of this storage solution. In particular, all of the cells in a NAND string are addressed by a single drain select transistor and a single source select transistor driven, respectively, by a drain select line (DSL) and a source select line (SSL), as schematically shown in Fig. 1. A source line is, then, connected to the source select transistors and a contact is needed to connect each drain select transistor to one of the bitlines (BLs) of the memory array. Although the area of these service elements may be significant if compared to that of each memory transistor individually (the gate length of the select transistors, for instance, is still close to 100 nm in state-of-the-art technologies with $F \simeq 15$ nm), the sharing of this area among all the cells in the NAND string allows to strongly reduce its relative impact on the total area of the memory array. To this aim, the number of cells in the NAND string has also been increased from 8 in the early stages of technology development [12], [13] up to 128 in state-of-the-art planar technologies [14], [15]. With a weak burden on the array area traded off for strong benefits on array operation, moreover, this increase has also allowed to consider the two

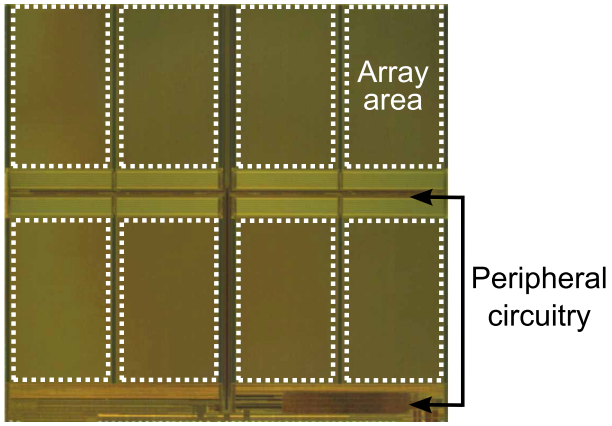


Fig. 3. ***Micrograph of a 16-nm NAND Flash memory die [11], highlighting the array area and the peripheral circuitry area.***

memory cells adjacent to the two select transistors as dummy cells [7], i.e., cells not to be used to store data. This has been required by the fact that the operation of these cells can be markedly different from that of the other cells in the string, being they not close to another memory cell but, precisely, to a select transistor [16]. Besides, as typically done in any storage system, part of the memory array has been devoted to spare strings for management information, error correction codes (ECCs), and redundancy, aiming at achieving the most reliable array operation. As will be discussed in more detail in Section III-A, all the previous service elements needed inside the memory array introduce an area overhead close, roughly, to the 20% of the total area of the NAND chip. In addition to that, another overhead term comes from the peripheral circuitry needed to drive the memory array and to interface the NAND chip with outside controllers. The peripheral circuitry, highlighted in the die micrograph of a 16-nm planar NAND Flash chip in Fig. 3, manages the voltage waveforms needed to store and read data in selected memory cells, increases the internally available operating voltage through charge pumps, and offers input/output pads for chip connection. The area overhead of the peripheral circuitry depends on a large variety of design solutions. For instance, it significantly changes between a one-sided circuitry and a two-sided circuitry [17]. Besides, it largely depends on array segmentation [18]. As a consequence, this overhead may span a relatively wide interval of values, being typically in-between 20% and 35% of the total chip area [19]–[24].

Accounting for all the service elements required by a memory technology is extremely important to correctly assess its performance and compare it with other storage solutions. In the case of integrated memory technologies, this can be done by defining a gross bit storage density (GBSD) as the ratio between the chip storage capacity (C_{chip}) and area (A_{chip}). Although the previous discussion highlighted that many factors contribute to the reduction of the GBSD of NAND Flash chips with respect to their ideal NBSD, this reduction can be considered relatively low, and, more importantly, did not prevent the NAND technology to outperform not only any other integrated nonvolatile memory solution targeting mass storage applications but also magnetic data recording on HDDs [2], [6].

B. Main Array Operations

Storing data in a NAND Flash array involves changing the floating-gate charge of the memory cells to set their threshold voltage (V_T) close to one of 2^{BPC} well-defined levels, where BPC is the number of bits stored in each cell. Fig. 4 shows a schematic view of this cell V_T discretization in a NAND technology with BPC = 1, conventionally called a single-level cell (SLC) or 1b/cell technology [Fig. 4(a)], BPC = 2, called a multilevel cell (MLC) or 2b/cell technology [Fig. 4(b)], and BPC = 3, called a triple-level cell (TLC) or 3b/cell technology [Fig. 4(c)]. Each V_T level corresponds to a different charge state of the cell floating gate and is associated to a specific string of the BPC bits. In particular, the floating-gate charge becomes more negative moving from the lowest V_T level, commonly referred to as the erased (E) cell state, to the highest V_T level. The E state represents the starting point of all of the array cells before data storage, consisting in selective program operations increasing cell V_T through the addition of electrons to their floating gate. All of the cell states different from the E state are therefore called programmed (P) states. Moving the memory cells from a P back to the E state is obtained through an erase operation. Finally, sensing cell state is called a read operation and involves applying read voltages (V_{RX} in Fig. 4) in-between the V_T levels to the control gate of the cell under test and checking the current flowing through it.

Before discussing more in detail the read, program, and erase operations in the NAND array, it is worth pointing out

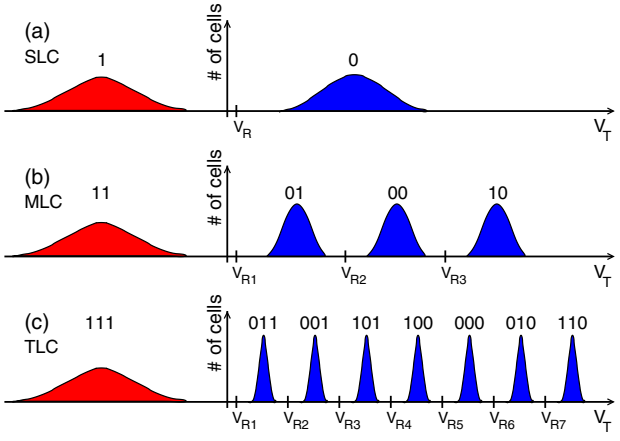


Fig. 4. ***Schematic picture for the V_T distribution of cells in the erased state (red) and in the programmed states (blue) in (a) a single-level cell (SLC); (b) a multilevel cell (MLC); and (c) a triple-level cell (TLC) NAND Flash technology. V_{RX} are read voltages. The total separation between the erased and the highest programmed V_T level in TLC technologies can be as large as 7–8 V.***

that the placement of the V_T levels along the voltage axis in Fig. 4 comes, historically, from the operation of NAND Flash chips with positive voltages only, allowing just one V_T level below 0 V. To overcome this limitation, however, virtual negative read algorithms have been recently proposed [25], [26] and are frequently used in state-of-the-art technologies. These algorithms are mainly based on changing the electrostatic potential of cell channel during read to achieve a horizontal shift of the cell V_T levels between the read and data retention conditions. In so doing, while the cell V_T levels reported in Fig. 4 are representative of cell states during read, some of the lowest programmed levels actually fall below 0 V under data retention conditions.

1) Read Operation: In order to sense the current flowing through a selected cell when a read voltage V_{RX} is applied to its control gate, all of the other cells and the two selectors in the same string must act as pass transistors [1], [12], [13]. This is obtained with the voltage scheme schematically depicted in Fig. 5(a): the voltage of the WL of the selected and unselected cells is raised, respectively, to V_{RX} and V^R_{pass}, in the presence of positively biased DSL and SSL allowing string connection to the BL (biased at $V_{\text{BL,R}} > 0$ V) and to the grounded source line. V^R_{pass} must be higher than the maximum V_T level in Fig. 4 to allow all of the unselected cells in the string to be highly conductive irrespective of their memory state. These operating conditions result in a current flow through the string and the BL that is mainly limited by the selected cell. From the integration of this current, a sense amplifier [17] allows then to assess if the cell V_T level is lower or higher than V_{RX} and, from that, the bits stored in the cell are derived (in the case of MLC and TLC technologies, bit decoding requires multiple sense operations with different V_{RX}).

The read time, i.e., the time required to accomplish a read operation on the array, has always been of few tens of microseconds over the last 15 years for NAND Flash chips [11], [19], [21], [27], [28], with, of course, a nonnegligible increase when moving from SLC to MLC to TLC technologies. The possibility to reduce this time, which can be considered rather long if compared to that of other memory technologies (NOR Flash chips allow read times well below 100 ns [29]–[32], for instance), is mainly limited by the large parasitic resistances and capacitances of the WLs and BLs of the NAND array, which are very long in order to offer the highest storage capacity without compromising the GBSD of the memory chip. The parasitic elements of the WLs and BLs constrain the array read time by introducing, first, RC delays in the microsecond timescale. In addition to that, they increase the sensing time of the small string current (of few tens of nanoamperes in modern technologies, owing to the large string resistance in series to the selected cell) when read is performed by directly integrating this current on the BL capacitance [17], [33].

As a result of the rather long read time, random access to single cells is unfavorable for NAND Flash chips and sequential access to large amounts of data gathered in parallel during the read operation is necessary to make the most of them. Read is, therefore, performed on a page basis, with typical size of 16 KB today [11], [15], [34]. Cells storing data belonging to the same page are all placed along the same WL and are sensed in parallel with the voltage scheme of Fig. 5(a) by raising their BLs to the read voltage $V_{\text{BL,R}}$. More specifically, two main page architectures are typically used, namely, the even/odd BL (EOBL) architecture [14], [17], [33], [35] and the all BL (ABL) architecture [17], [36]–[39]. In the former, cells contributing to the same page are half of those along the same WL and a single sense amplifier is shared between two adjacent BLs. As a result, cells storing data in the same page are all those at even or odd positions along the WL. In the latter, instead, a sense amplifier is used independently for each BL and cells contributing to the same page are all those along the same WL.

Thanks to the high degree of parallelism allowed by page operation and exploiting array segmentation and optimization [18] to further boost access performance, read throughputs well above 100 MB/s, i.e., comparable to or even higher than those of NOR Flash chips, have been recently reported for the latest NAND technologies [14], [22], [40]. Such throughput values make the NAND technology highly competitive with all the other nonvolatile memory solutions targeting mass storage applications.

2) Program Operation: Uniform Fowler–Nordheim tunneling of electrons from the channel to the floating gate

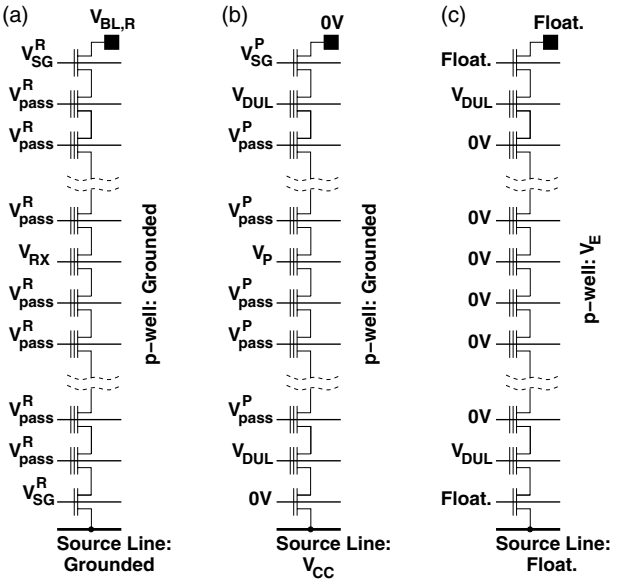


Fig. 5. ***Schematic description of the voltages applied to the NAND string to (a) read and (b) program a selected memory cell in the string (the selected cell is that whose WL voltage is V_{RX} and V_P during read and program, respectively). (c) Voltage scheme to erase the entire NAND block. V^R_{SG} and V^P_{SG} are the voltages applied to the gates of the biased select transistors during read and program, respectively.***

through the thin SiO_2 layer in-between them (called, therefore, the tunnel oxide of the device) has rapidly emerged as the elective physical mechanism to selectively program the memory cells in a NAND Flash array [41]–[44]. Fowler–Nordheim tunneling, in fact, outperforms all the other physical mechanisms allowing electron injection to the cell floating gate, e.g., channel hot-electron injection exploited in NOR Flash technologies [45], [46], in terms of the electron storage efficiency. Electron storage efficiency is the ratio between the number of electrons stored in the cell floating gate and the number of carriers flowing through device contacts while programming the memory cell and, in the case of Fowler–Nordheim tunneling, is almost equal to 1. This allows to exploit the main benefit of the NAND architecture, i.e., parallelism, not only during read but also during program operations. Thanks to the low current flowing through the selected cell and the resulting low power dissipation, the program operations can, in fact, be performed on the same page basis of the read operations [33], [43], allowing high program throughputs above 20 MB/s today [15], [34]. Moreover, uniform Fowler–Nordheim tunneling also allows good array reliability, better, in particular, than that achievable with channel-hot electron programming [44], [47].

Fig. 5(b) shows the voltage scheme required for Fowler–Nordheim tunneling to take place in a selected cell in the string. A high positive program voltage V_P is applied to the WL of the selected cell, with all of the other WLs at a pass voltage V^P_{pass} (higher than V^R_{pass}). To avoid programming cells in the unselected strings by self-boosting their channel [48], the SSL is grounded, then keeping off all the source select transistors. Only the drain select transistor is, therefore, turned on by applying a positive voltage V^P_{SG} to its gate, connecting the string to the grounded BL. In order to reduce the electric field in the dielectric material isolating the DSL/SSL and the DUL, these latter lines are biased at a voltage V_{DUL} lower than V^P_{pass}. As a result of this voltage scheme and of the high electric field created in the tunnel oxide of the selected cell by V_P, electrons uniformly tunnel from the grounded channel of this cell to its floating gate, where they are stored. Electron storage in the floating gate reduces the electrostatic potential of this region and, in turn, increases cell V_T by an amount

$$\Delta V_T = \frac{q\Delta n}{C_{pp}} \tag{1}$$

where C_{pp} is the control-gate-to-floating-gate capacitance, q is the elementary charge and Δn is the number of stored electrons.

The program operation on a selected memory cell comes to a conclusion when device V_T overcomes a target program-verify level V_{PV} [43], [50]. To check this condition, cell programming typically consists in a program-and-verify algorithm, with a sequence of program pulses followed by verify operations. Verify operations are just read operations comparing cell V_T with V_{PV}, to determine whether ($V_T > V_{PV}$) or not ($V_T < V_{PV}$) programming is complete (in the latter case, more program pulses are applied). For better program performance, the amplitude V_P of the program pulses applied to the selected cell is typically increased by a constant step V_S, resulting in the so-called incremental step pulse programming (ISPP) [48], [49], [51]. Fig. 6(a) schematically describes the sequence of program and verify operations performed within the program algorithm by highlighting the WL voltage applied to the selected cell and assuming ISPP. The regular increase of V_P from one program pulse to the next allows the cell to enter a stationary programming regime in which its average ΔV_T ($\langle \Delta V_T \rangle$) per step nearly equals V_S [48], [49], [51], [52], as shown in Fig. 6(b). This allows not only to make variability of cell parameters negligible for the final programmed V_T distribution (cell-to-cell variability impacts only the number of ISPP pulses needed to reach V_{PV}) but also to narrow the programmed V_T distribution just by adopting smaller V_S (resulting in a tradeoff between program time and accuracy), as shown in Fig. 7. ISPP also allows to keep a nearly constant electric field in the cell tunnel oxide (about 11–12 MV/cm) [48], [49], [51], [52], offering the best solution for the reliability (requiring low electric fields) versus program speed (requiring high electric fields) tradeoff [49]. Note, finally, that, although the tunnel-oxide thickness has been slowly

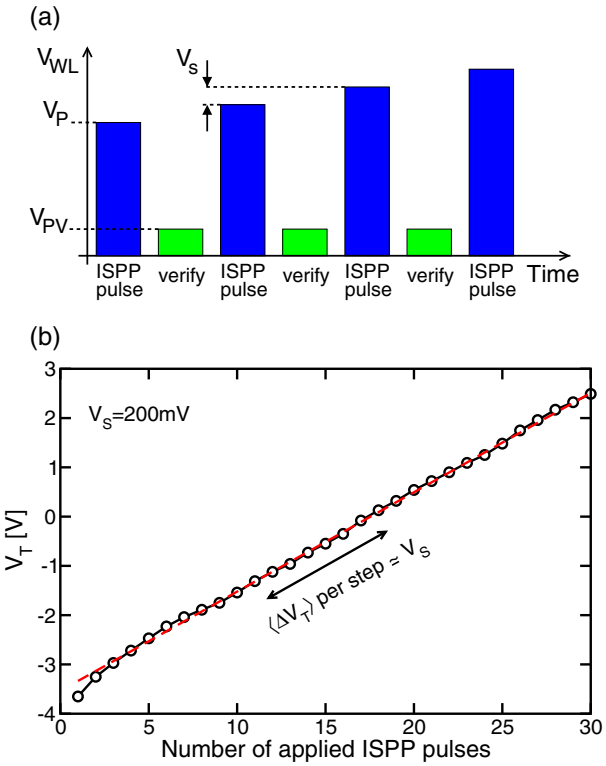


Fig. 6. ***(a) Schematic description of the WL voltage applied to the selected cell during a program-and-verify algorithm adopting ISPP for better program performance. (b) Example of the resulting V_T transient measured on a deca-nanometer NAND cell belonging to a test array structure in the case of $V_S = 200$ mV. Note that the test array allows to monitor negative V_T values.***

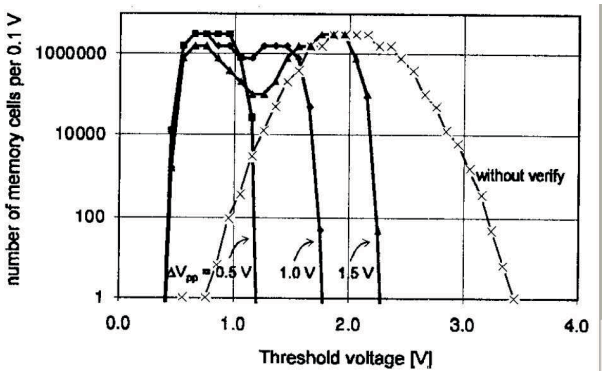


Fig. 7. *V_T distribution resulting from a program-and-verify algorithm making use of ISPP with different V_S (indicated as ΔV_{pp} in the figure), measured on a 16-Mb test NAND Flash array. Reprinted from [49].*

reduced over the years and is only of 6–7 nm in state-of-the-art planar technologies [53], V_P close to or even higher than 20 V are typically reached during ISPP.

To conclude this discussion on the program operation, it is important to point out that the program time in NAND Flash chips is commonly of few hundreds of microseconds in the case of SLC technologies [19], [21], [54]–[56], of about 1 ms for MLC technologies [11], [20], [22], [28], [39], [57]–[60] and of a few milliseconds in the case of TLC technologies [14], [61], owing to the higher program accuracy required in the latter cases (see Fig. 4). In particular, in the case of MLC and TLC technologies, program accuracy typically requires the reduction of the ISPP step amplitude V_S or the application of more complex program-and-verify algorithms [62]–[65]. Besides, a variety of multiple-round programming schemes have been proposed and are typically used for these technologies, aiming at improving the program accuracy by reaching the target number of stored BPC through subsequent program operations [59], [64], [66]–[68]. Note, anyway, that, owing to the page-based operation of NAND Flash chips, the most meaningful parameter to assess their program performance is not the program time but the program throughput, which, thanks to the increase of the page size, has been largely increased with technology scaling, as will be shown in Section III-B.

3) Erase Operation: Flash memory chips are electrically erasable programmable read-only memory (EEPROM) components allowing electrical cell erase with block granularity [1], [45], [46], [69]. In the case of a NAND array, a memory block is made of all the strings sharing the same WLs, resulting in a typical block size of 4 MB for state-of-the-art planar MLC technologies [15], [34]. Block erase is accomplished within few milliseconds [19], [22], [23], [56]–[58], [60], [70] by the voltage scheme depicted in Fig. 5(c): a high positive voltage V_E is applied to the *p*-well of the array with all of the WLs grounded, leading to a uniform Fowler–Nordheim tunneling of electrons from the floating gate to the channel area of the memory cells [41], [42], [71]. To limit the electrical stress on the other parts of the block, including the string selectors, all the BLs, the source line, the DSL, and the SSL are left floating. Moreover, similarly to the program case, a positive voltage V_{DUL} is applied to the DUL to reduce the electric field in the insulators between these lines and the DSL/SSL. After the erase pulse, an erase verify operation checks whether V_T of all the cells in the block is below 0 V (this is accomplished by a read operation on the strings with all the WLs grounded) and, in the case this condition is not met, an additional erase pulse is applied with higher V_E. Note, however, that, differently from the program-and-verify algorithms previously discussed, this erase-and-verify algorithm allows only the control of the upper value of the cell V_T distribution, but does not result in distribution tightening.

III. HISTORICAL SCALING TRENDS

In this section, the historical evolution of the most important physical and electrical parameters of the planar NAND Flash technology will be analyzed. Since it was only with the beginning of the new millennium and the widespread diffusion of portable electronic devices that the NAND technology started its explosive market growth, attracting the interest and the research efforts of many of the top semiconductor manufacturers, attention will be focused on the scaling trends between 2001 and 2015. Over this stretch of time, the NAND technology not only became the most important integrated nonvolatile memory solution, overcoming the NOR Flash technology in terms of revenues in 2005 [72], but also eroded the market share of HDDs with the development of SSDs. Since 2015, then, the research and development efforts have been mainly turned to 3-D array architectures, as will be discussed in detail in Sections V and VI.

All of the data that will be presented here were gathered from the papers published by technology manufacturers at the IEEE International Solid-State Circuits Conference (IEEE ISSCC) [7], [11], [14], [15], [19]–[24], [27], [28], [34], [36], [38]–[40], [56]–[61], [66], [68], [70], [73]–[75]. Although what was reported in these publications may not match exactly final products and although the publication time may slightly differ from when each technology entered the market, ISSCC publications can be considered a good reference to investigate the pace of technology evolution, providing details on fully working memory chips.

In the next sections, the miniaturization pace kept by NAND manufacturers over the years will be discussed, starting from the trends of the technology node feature size F and of the planar bit storage density in terms of GBSD. The latter is likely the most important parameter for a memory technology, determining its cost and competitiveness with other storage solutions. Then, the evolution of other array parameters will be reviewed, paying special attention to the program throughput and the page size.

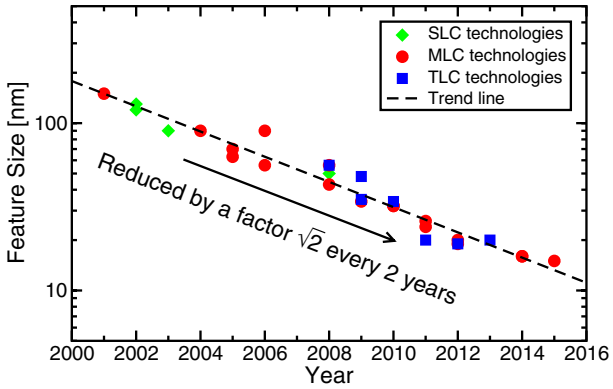


Fig. 8. ***Trend of the feature size F of planar NAND Flash technologies between 2001 and 2015.***

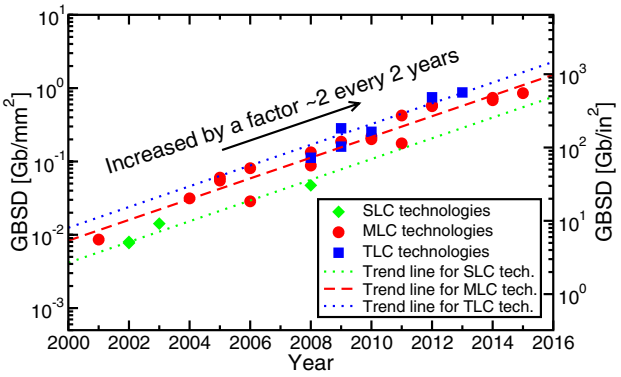


Fig. 9. ***Trend of the GBSD of planar NAND Flash technologies between 2001 and 2015.***

A. Technology Node Feature Size and Bit Storage Density

Fig. 8 shows the feature size F of the planar SLC, MLC, and TLC NAND Flash technologies developed over the last 15 years. The steady miniaturization pace followed by technology manufacturers clearly appears from the extremely good merging of the data points with their average trend line, resulting in a reduction by a factor $\sqrt{2}$ of F every two years. This represents a clear proof that the huge efforts and investments in this field allowed the planar NAND Flash technology to follow Moore's law until the latest node with $F \simeq 15$ nm, totalizing a reduction of F nearly equal to a factor 11 from 2001 to 2015.

The reduction of F allowed the increase of the planar bit storage density highlighted in Fig. 9 in terms of GBSD. To investigate this increase quantitatively, the GBSD trend line was extracted considering only the data points referring to MLC technologies, since these points outnumber those available for SLC and TLC technologies and allow the most accurate extraction procedure. The resulting dashed red line in the figure nicely reveals an increase by a factor 1.92, i.e., very close to 2, of the GBSD every two years, representing another proof of the successful application of Moore's law to the scaling of planar NAND Flash technologies. With this relentless increase of the bit storage density, GBSDs close to 1 Gb/mm^2 were achieved in memory chips developed using the latest technology nodes, making NAND Flash storage highly competitive with magnetic storage on the platters of HDDs [2].

The trend line for the GBSD of MLC technologies can be used to calculate the average gross area occupancy of a memory cell in the array part of the NAND Flash chip, i.e., the area per cell when accounting for all the overhead terms present in the memory array and discussed in Section II-A. Denoting this area as $g \cdot F^2$ and assuming that a fraction X_M of the chip area A_{chip} is occupied by the memory array (this fraction is typically called the array efficiency of the memory chip), it is, in fact, possible to say that

$$A_{\text{chip}} \cdot X_M = g \cdot F^2 \cdot \frac{C_{\text{chip}}}{\text{BPC}} \tag{2}$$

where C_{chip} is the chip storage capacity. From this relation, the GBSD can be calculated as

$$\text{GBSD} = \frac{C_{\text{chip}}}{A_{\text{chip}}} = \frac{X_M \cdot \text{BPC}}{g} \cdot \frac{1}{F^2}. \tag{3}$$

By comparing (3) with the trend line of Fig. 9 for MLC technologies (BPC = 2) and assuming an average $X_M = 0.7$ [20], [22], a factor $g \simeq 6$ can be extracted. Since $4F^2$ corresponds to the real net area of the NAND cell, this means that all the overhead terms coming from the string select transistors, dummy cells, source lines, BL contacts, and spare strings add, on average, a contribution nearly equal to $2F^2$ to the gross area of each useful memory cell in the array. Besides, this result allows to quantify in a factor $g/4X_M \simeq 2.1$ the average reduction of the GBSD from the NBSD = BPC/$4F^2$. This means that, with 30% of the chip area taken by the peripheral circuitry, roughly 20% of the chip area is for service elements in the memory array and the remaining 50% of the chip area is for useful memory cells. The 20% of the chip area taken by the array service elements can be further split almost uniformly between the string overhead terms (select transistors, dummy cells, BL contacts, and source lines) and the spare parts of the array for management information, ECC, and redundancy.

Owing to the limited number of points available in Fig. 9 for SLC and TLC technologies, the trend lines for these technologies were not calculated by data fitting but, instead, by multiplying the extracted trend line for MLC chips by a factor 0.5 and 1.5, respectively. The resulting dotted lines in the figure appear, anyway, in reasonable agreement with the available data points and give the idea of the storage improvements coming from the increase of the BPC. Moreover, this agreement further proves the validity of the previous results for g and the area occupancy of the service elements in the NAND array.

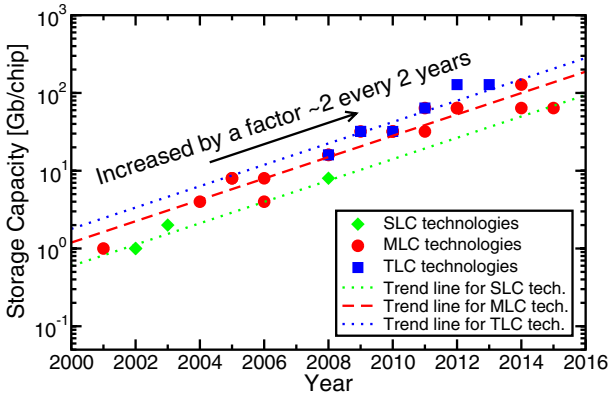


Fig. 10. *Trend of the chip storage capacity C_{chip} of planar NAND Flash technologies between 2001 and 2015.*

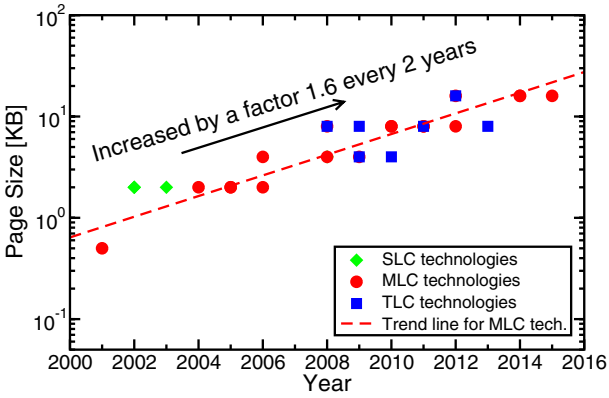


Fig. 12. *Trend of the page size of planar NAND Flash technologies between 2001 and 2015.*

As a final remark, note that the GBSD of NAND Flash chips reported in Fig. 9 are typically a factor 3 to 4 higher than those achievable by NOR Flash chips with comparable F [29], [31], [32], [76], making nonvolatile storage in the latter components more expensive in terms of cost per GB. This is mainly due to a larger net area of the single memory cells in the NOR array, with X_M being only slightly less than in the NAND case. The origin of the larger physical area of memory cells in the NOR array is twofold [77], [78]. First, the NOR architecture offers a direct access to the source and drain terminals of all of the floating-gate transistors, at the expense of the introduction of a BL contact and a source line every two adjacent memory devices. Second, the channel length of the floating-gate transistors cannot be reduced to the minimum dimension allowed by the technology, i.e., F, owing to the need to limit the impact of drain-induced barrier lowering (DIBL), source-to-drain punch through, and drain turn-on effects on device current in the presence of the relatively high BL voltages applied during array operation [77].

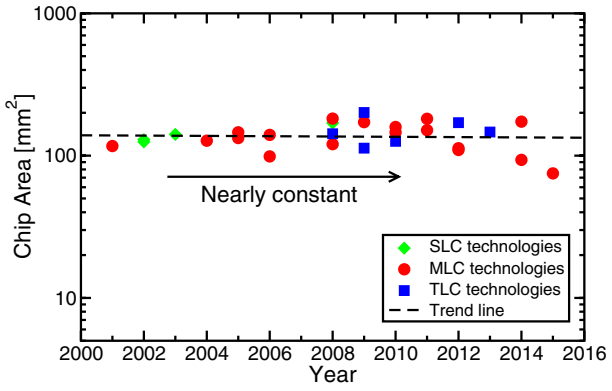


Fig. 11. *Trend of the chip area A_{chip} of planar NAND Flash technologies between 2001 and 2015.*

B. Chip Capacity and Area, Page Size, and Program Throughput

Fig. 10 shows that the increase of the GBSD of NAND Flash memory chips has been pursued along with the increase of the chip storage capacity C_{chip}, which has grown with the same average pace of a factor close to 2 every two years. This means, first, that the chip area A_{chip} has remained nearly constant over the years. This is proved by the data points in Fig. 11, from which an average value $\langle A_{chip} \rangle \simeq 136$ mm^2 can be extracted. This area appears as the best solution in terms of cost per GB of the tradeoff between the increase of C_{chip} and the reduction of the production yield. Note, moreover, that the growth of C_{chip} has led the typical storage capacity of NAND Flash chips to rise from 1 Gb in 2001 to 128 Gb in 2015. This confirms NAND Flash chips as the electronic components with the largest number of integrated devices, far above DRAM and microprocessor units [79].

Fig. 12 shows that the evolution of NAND Flash chips toward higher storage capacities has also resulted in the average increase of the page size of the array. By fitting the data points for MLC technologies, this increase can be quantified in a factor of 1.6 every two years, resulting in a typical page size of 16 KB in state-of-the-art planar arrays. In this regard, it is worth mentioning that the possibility to increase the page size for a given technology node has been mainly limited by the RC delays constraining the maximum length of the array WLs. The RC constraints to the WL length result, in fact, in a limitation to the total number of cells along the WLs, equal to the ratio between the WL length and the cell pitch $2F$. Since cells contributing to the same page of the memory chip are either all those (ABL architecture) or half of those (EOBL architecture) along the same WL, this latter limitation corresponds, in turn, to a limitation to the maximum page size. From this standpoint, the increase of the page size by a factor of 1.6 every two years coming from Fig. 12 appears not far from what was expected in the case of constant WL length and

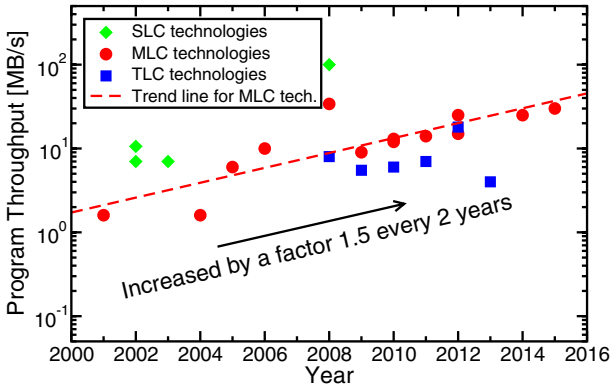


Fig. 13. *Trend of the program throughput of planar NAND Flash technologies between 2001 and 2015.*

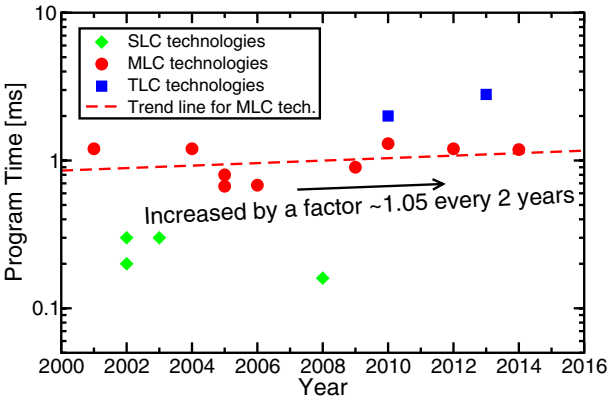


Fig. 14. *Trend of the program time of planar NAND Flash technologies between 2001 and 2015.*

scaling of the cell pitch $2F$ in the WL direction by a factor of $\sqrt{2}$ every two years.

Thanks to the increase of the page size, the throughput performance of NAND Flash memory chips has been significantly improved over the years, as shown in Fig. 13 for the program operation. Referring again to MLC technologies, the trend line for the program throughput reveals an average increase by a factor of 1.5 every two years, i.e., quite close to the average scaling pace of the page size in Fig. 12. This confirms that the time required to perform a program operation on the array has been only weakly increased with technology scaling, as directly appearing in Fig. 14. Keeping the program time basically unaltered over the years has been mainly achieved through constant improvements and innovations in the memory chip, aiming at reducing the impact of new physical issues making the tightening of the V_T distribution during programming more and more difficult, as will be discussed in Section V. As a final remark, it is worth mentioning that the program throughputs shown in Fig. 13 for NAND Flash chips are about a factor 2 to 3 higher than those typically achieved by NOR Flash chips developed with process technologies with comparable F [29], [31], [32]. From the program throughput standpoint, in fact, NOR Flash chips suffer from a quite limited program parallelism (only a few bytes of a page buffer are programmed in parallel), owing to the high current and power requirements of the channel hot-electron injection mechanism used to modify the floating-gate charge in this case. These requirements result in area expensive charge pumps and, in turn, in a tradeoff between the program throughput and the GBSD of NOR Flash chips [80]. Note, instead, that the high efficiency and low current needed by uniform Fowler–Nordheim tunneling prevent this injection mechanism from setting any relevant constraint to the program parallelism of NAND arrays. As a consequence, the program throughput of NAND Flash chips can take full advantage of the increase of the page size following BL pitch reduction, as highlighted by the similar growth rate of the page size and the program throughput in Figs. 12 and 13.

IV. SCALING FROM THE PERSPECTIVE OF APPLICATIONS

The market demand for NAND Flash memories has constantly grown over the last 15 years [81], [82]. This has been the result of a broader and broader range of applications choosing the NAND technology as their elective nonvolatile storage solution, thanks to its possibility to meet some important and specific requirements better than any other memory technology. Moreover, the constant scaling pace followed by NAND Flash chips in terms of F and GBSD has made them more and more convenient from the economic standpoint.

Portable electronic devices were the first to foster the success of the NAND technology. Key requirements for these devices are a high mechanical shock immunity, a low power consumption, and a high volumetric bit storage density (corresponding to the ratio of the planar bit storage density and the thickness of the storage medium). All of these requirements can be met by NAND Flash chips better than, for instance, HDDs. Focusing on the volumetric storage density, note in fact that, although it was only recently that the planar GBSD of NAND Flash chips overcame the bit storage density on the platters of HDDs [2], the thickness of the storage media has always represented a strong point in favor of the NAND Flash technology. While the thickness of a double-sided HDD platter is in the millimeter scale, the thickness of a NAND chip can be thinned down to less than 100 μm [83], making the volumetric bit storage density of NAND chips about a factor 5 to 10 higher than that of HDD platters. This allows to reduce the volume required to provide a given storage capacity in the NAND case, which is of fundamental importance to reduce the area and the thickness of portable devices.

Starting from the low storage capacity segment of portable devices, the NAND Flash technology then broadened its market, addressing medium storage capacity

Fig. 15. ***Picture of the exterior of SSDs with different form factors.***

applications with SSDs and, more recently, the high storage capacity field of big data. This expansion found in the cost gap still existing with respect to HDDs the most relevant restraint, with a consequent attempt to loosen it by increasing the performance gap between the two storage solutions, favorable to the NAND technology. In the next sections, SSDs will be considered as a case study to understand how the NAND technology tackled a market traditionally dominated by HDDs, highlighting the performance and the key parameters of SSDs that are currently triggering a change in the nonvolatile storage scenario.

A. Cost and Performance of SSDs

In the last decade, SSDs made of a multitude of individual NAND Flash chips configured for parallel operation (see Fig. 15) have proved themselves as a compelling storage solution for medium and high capacity applications traditionally dominated by HDDs. This has not come only from the fact that SSDs can provide all of the benefits that allowed the NAND Flash technology to become the elective solution in the low storage capacity market of portable devices, i.e., higher mechanical shock immunity thanks to an all-electrical system, lower power consumptions, and higher volumetric storage densities enabling more streamlined form factors than HDDs. Since the cost gap on a per GB basis still existing between NAND Flash and HDD storage results in less affordable cost differences between the two solutions as the required system capacity increases, SSDs had to offer something more valuable for these systems. This was, of course, operating performance, i.e., speed.

Before moving to performance, however, it is worth further commenting on the cost comparison between SSDs and HDDs. First, the scaling trends of Figs. 10 and 11 revealed that NAND Flash chips could offer over the years a higher and higher storage capacity on the same silicon area, resulting in a constant decrease of the cost per GB of the technology. This has enabled SSDs not only to constantly decrease their cost, but also to become cheaper than HDDs at useful capacity points. This concept can be understood by means of Fig. 16, showing the evolution of the average selling price of 128-, 256-, and 512-GB SSDs in comparison to that of 500-GB HDDs. For the considered storage capacity, the price of HDDs has been almost flat over the years, meaning that 500 GB is today a capacity making the fixed costs of the storage system, mainly associated to its mechanical parts, dominant over those of the storage medium. These fixed costs introduce, therefore, a floor at which the cost of all the HDDs with capacity lower than or equal to 500 GB saturates. SSDs also have a floor cost given by circuit board assemblies and controller application-specific integrated circuits (ASICs), but this is much lower than that of HDDs. This allows to take full advantage of the scaling trends of the NAND Flash technology investigated in Section III, with the possibility to constantly reduce the cost of the SSD system in parallel to the reduction of the cost per GB of the technology. This, eventually, allows, for a given capacity, the price of SSDs to fall below that of HDDs, as shown in Fig. 16. The fact that this has been occurring today for storage capacities in the range between 128 and 256 GB is noteworthy. This capacity range, in fact, is one of the most relevant for the client storage market, which is that of laptops and personal computers. Moreover, the previous cost dynamics will make HDDs economically convenient only for higher and higher capacity applications, where, however, performance plays a more important role.

From the standpoint of performance, it is important to point out that SSD architectures exploit the parallelism of the array of attached NAND Flash chips to address system workloads that are dominated either by long sequential access patterns or by short random read and write commands. The possibility to implement this parallelism

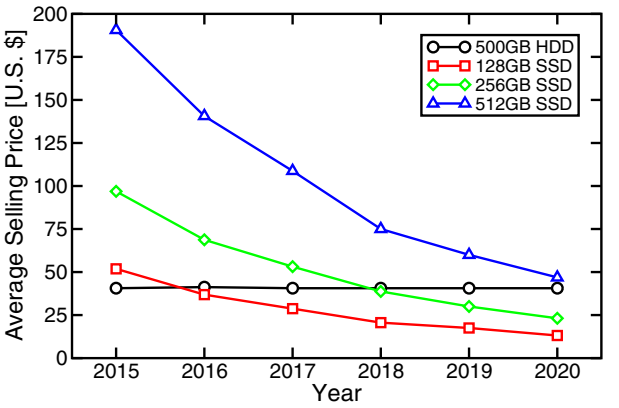


Fig. 16. ***Evolution of the average selling price of 128-, 256-, and 512-GB SSDs in comparison to that of 500-GB HDDs. Source: IDC, Worldwide Solid State Drive Forecast Update, 2016-2020: December 2016, Doc.#US40808816, Dec. 2016.***

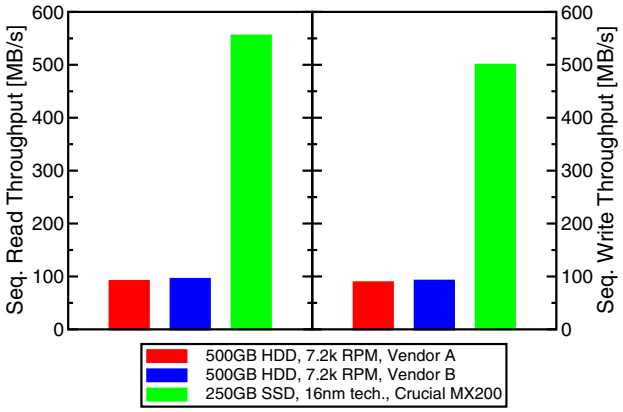


Fig. 17. ***Sequential read and write throughputs of two representative 500-GB HDDs and of a representative 250-GB SSD. Fresh-out-of-the-box values are reported for the SSD.***

represents a strong point in favor of SSDs, which can be only partially achieved by HDDs owing to the inherent serialism of their head-media interface. In the case of long sequential patterns, the SSD breaks the data transfer into segments that fit within the page architecture of each single NAND chip, allowing them to operate independently and in parallel on a portion of the transfer. The benefit of this can be seen in Fig. 17, where the sequential read and write throughputs of representative high-capacity HDDs and SSDs are compared. When, instead, the SSD is required to handle short random commands, parallelism is exploited by simultaneously working on many different host requests. In this case, the common metric used to measure the performance of the storage system becomes the number of random read and write IOPS it can manage. Thanks to parallel chip operation and the ability to immediately address any element in the array of chips, SSDs can provide very high random IOPS, far beyond the typical values of HDDs, as shown in Fig. 18. HDDs are,

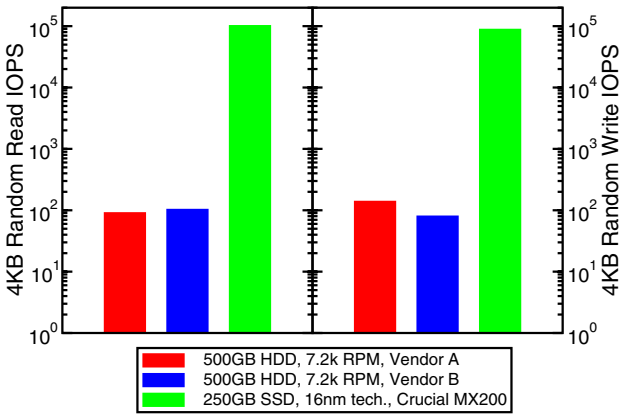


Fig. 18. ***Number of 4-KB random read and write IOPS handled by two representative 500-GB HDDs and by a representative 250-GB SSD. Note the use of logarithmic y-scales. Fresh-out-of-the-box values are reported for the SSD.***

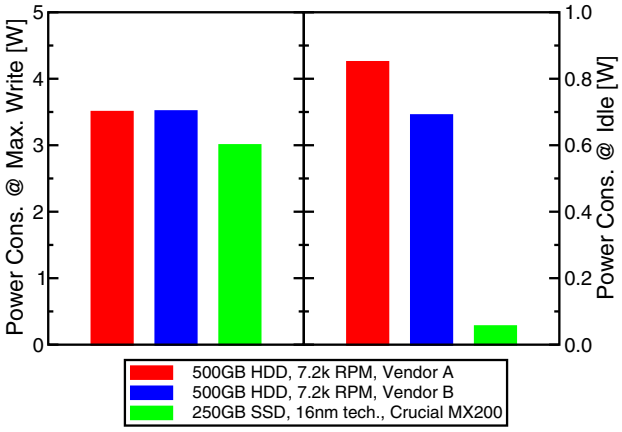


Fig. 19. ***Power consumption during write at maximum speed and when idle of two representative 500-GB HDDs and of a representative 250-GB SSD.***

in fact, limited in parallelism and constrained in the time needed to address any arbitrary location on the storage medium by mechanical delays related to platter rotation and head movement. As a final remark, note that the sequential throughput and the number of random IOPS of an SSD are intimately related both to the performance and the number of NAND Flash chips used in the system. The SSD results of Figs. 17and 18 are, therefore, the outcome of the constant improvement of the storage capacity and of the program throughput of the individual NAND Flash chips discussed in Figs. 10–13.

To complete this short overview on the advantages offered by SSDs, Fig. 19 shows some typical power consumption results, highlighting another strong point in favor of this storage solution with respect to HDDs. Data for power consumption are reported under two extremely different working conditions, namely, when the system writes data at its maximum speed and when the system is idle. Although in both cases the comparison reveals that SSDs outperform HDDs, the benefits offered by SSDs are extraordinary especially under idle conditions.

B. Challenges and Architectural Options for SSDs

Even if the performance benefits are unquestionable, data storage in SSDs creates a unique set of challenges relative to HDDs. First, NAND Flash programming is performed on a page basis while the granularity of the erase operation is the block. This mismatch means that, unlike HDDs, an SSD is not a rewrite in place storage device. When the host rewrites previously written data, the new data are placed on an empty page while the data at the old location are flagged as obsolete. Managing this process introduces a level of abstraction between the physical address space of the NAND array and the logical address space used by the host, commonly referred to as the Flash translation layer (FTL) [84]. It is typical to have several data structures in the FTL,

including a map from the logical address space used by the host and the physical address space of the NAND array along with block statistics to help manage system reliability. As the host writes data into the SSD, the logical to physical map is filled. Then, when the host erases and rewrites some of the previously written data, old links become obsolete and are replaced with new ones. This leads to a situation where some pages in a block have current user data, while others store obsolete information. To handle this situation, the SSD periodically starts a garbage collection process by copying the valid user data in a block into a new one, then erasing the old block.

Although, on the one hand, this management of stored data may seem a burden for the operation of SSDs, on the other hand, it can be exploited to make the most of them from the reliability standpoint. The policy used to identify target blocks during garbage collection, in fact, may create a uniform distribution of program/erase cycles among the NAND Flash chips in the SSD (wear leveling [85], [86]). It is also used to select blocks with infrequently written data and refresh them. All of these precautions allow to handle the endurance and the data retention characteristics of NAND Flash chips at the system level, maximizing the reliability of the storage device. However, this does not come for free. Over their lifetime into an SSD, in fact, user data may be copied many times. The average number of times data are rewritten in the system is the so-called write amplification (WA) [87], which can be considered as a metric for SSD efficiency. In order to successfully manage data reallocation among the NAND Flash chips, in fact, the SSD raw capacity is increased with respect to the storage capacity offered to the user. The ratio of raw capacity to user capacity defines the amount of overprovisioning in a system. As the overprovisioning goes up, the amount of obsolete data that is uniformly distributed in the NAND Flash chips goes down, with a consequent reduction of the WA associated with garbage collection. Within this tradeoff, of course, the cost of the SSD plays a relevant role, since overprovisioning has a direct impact on the final cost of the device. Moreover, even the scaling trends of the page size (Fig. 12) and of the block size of NAND Flash chips come into play in WA and overprovisioning. These scaling trends, in fact, do not fundamentally alter the management of data in the SSD but do create further challenges that need to be overcome. Although the increase of the page size is an important part of the steady increase of single chip and SSD performance, in fact, the consequent decrease of the storage granularity affects the mapping of host data in the system, owing to a higher mismatch between the page size and the size of the basic data unit from the host. Moreover, the page size affects the capacity of the temporary buffers used to aggregate host write commands until full pages can be written. SSD protocols typically have command queues ranging in depth from 32 to 256 [88], [89]. Deep queues help to maintain performance as data are aggregated in volatile buffers and then written to the NAND Flash array. In addition to command queues, SSDs may employ SLC caches to stage the host data in MLC/TLC products until aggregation and write complete successfully [90]. These techniques effectively mitigate the impact on system performance of the growth of the NAND Flash page.

In the assessment of the architectural options for an SSD, the target market segment of the storage device should be carefully considered. Although there are many possible ways to define market segmentation, the one based on target customers is likely the most important. Within this segmentation, client and data center/enterprise SSDs can be defined, with the former dominated by personal computers and laptops. Client storage has a few broad characteristics that can be leveraged to create optimized SSD solutions. A couple of these are duty cycle of activity and average percent full of the storage. Client computing generally has short periods of high traffic levels separated by long periods of relatively low traffic. The user experience is dominated by early life use cases where the storage device is not full. These features are in contrast to those of data center/enterprise storage. SSDs for data center/enterprise storage operate at a much higher duty cycle than client SSDs and with performance specified during the steady state condition of a full storage device. Optimizing the SSD operation according to these features has a fundamental impact on the system performance and reliability. For instance, client SSDs can effectively use nonvolatile write caches to absorb bursts of write activity, relying then upon periods of low activity to organize and destage the cache. A common way to do this today is by exploiting the better performance of SLC operation in chips where MLC/TLC storage has been introduced for GBSD improvement [91]. For example, an SLC NAND Flash cache is typically used in front of an MLC/TLC main storage array. Different manufactures have chosen different strategies for this. Some allocate a fixed number of blocks in the NAND Flash chips for SLC operation and the remainder for MLC/TLC storage. Other dynamically allocate blocks as either SLC or MLC/TLC [90]: when the SSD is operating at less than full capacity, many more blocks are used in SLC mode and only when the SSD becomes full blocks are reallocated as MLC/TLC to assure system capacity. Either caching mechanism allows user experience to be set by the SLC performance with, at the same time, the cost benefits of MLC/TLC storage. This is just an example of how SSD performance can be further boosted at the system level when target market applications are defined, representing a huge benefit of the technology.

V. SCALING ISSUES FOR THE PLANAR NAND FLASH TECHNOLOGY

Although the miniaturization of the planar NAND Flash technology went on for more than two decades, with the successful trends discussed in Section III over the last 15 years,

semiconductor manufacturers are now changing their integration paradigm, moving to 3-D memory arrays. This historic choice was not dictated by the impossibility of developing new planar nodes with smaller F, but by the general agreement on that the GBSD trend of Fig. 9 can be more favorably followed in the future by moving to 3-D memory solutions. In order to better understand the reasons for that, before moving to 3-D array integrations, in this section, the main burdens to planar deca-nanometer NAND Flash technologies will be discussed, starting, first, from process-related issues and moving, then, to the physical issues affecting array performance and reliability.

A. Process-Related Issues

The most important process-related issue making the pursuing of new planar nodes unfavorable with respect to the transition to 3-D integrations of the NAND memory array is the increase of the complexity and cost of the photolithographic process [92]–[94]. In this regard, note that the widely adopted argon/fluorine (ArF) immersion litography allows a minimum F on the silicon wafer of about 38–40 nm [94]. To overcome this limitation, double-patterning techniques were devised and exploited to develop NAND Flash technologies [42] with F between 40 and 20 nm [95]–[97], resulting in a first severe increase of the production complexity and cost. Moving below the 20-nm node involved, more recently, a further relevant increase of the fabrication issues, coming from the need of quadruple-patterning techniques [98]. These techniques, in principle, push the ultimate F limit of planar NAND technologies to 10 nm, allowing the chance for a last planar node in the next few years. Note, however, that, even if this additional node were developed, it would be late with respect to the scaling trend of F shown in Fig. 8, which would have required it already in 2016. This proves that the research efforts and investments of NAND makers have already been moved to 3-D technologies, offering the possibility to increase the chip GBSD by the vertical stacking of many memory layers with less demanding photolitographic steps.

B. Physical Issues

Although the reduction of F allowed to strongly increase the GBSD of NAND Flash chips, the consequent cell miniaturization made also more difficult keeping a correct array operation, owing to more and more relevant issues coming from some physical effects negatively impacting data storage and retention. Among them, the following are worth mentioning.

1) Program Noise: The decrease of the cell area coming from the reduction of F has led, as a side effect, to the decrease of the cell capacitances coupling the floating gate to the channel and to the control gate (C_{pp}). The reduction of C_{pp}, in particular, has largely increased the impact that single electrons stored in the floating gate have on cell V_T, as resulting from (1) [99]–[101]. This, in turn, has decreased the number of electrons to be transferred to/from the floating gate to achieve a selected ΔV_T, as shown in Fig. 20 for $\Delta V_T = 100$ mV [4]. From this figure, it is clear that a variation of the floating-gate charge of a few tens of electrons results in a change of cell V_T that is relevant for the operation of modern arrays, especially when referring to MLC and TLC technologies, on account of their reduced noise margins (see Fig. 4).

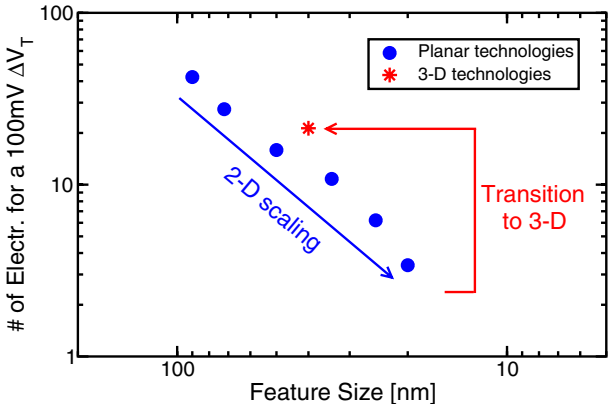


Fig. 20. *Calculated number of electrons to be transferred to/from the cell floating gate to give rise to a 100-mV ΔV_T, as a function of the feature size F of planar NAND Flash technologies. Results for a 3-D NAND technology are also shown [4].*

The most important issue for array operation coming from the reduction of the number of electrons needed to achieve a selected ΔV_T is, by far, the so-called program noise [52], [65], [92], [99], [100], [102]–[104]. This is a direct consequence of the more relevant role played by the fundamental statistical fluctuation of the number of electrons Δn injected into the cell floating gate during a program pulse when the average value $\langle \Delta n \rangle$ corresponding to the target ΔV_T decreases. In this regard, it is important to recall that the program operation of a NAND cell makes use of ISPP to achieve a well-controlled increase of V_T and, in turn, tight programmed V_T distributions in the presence of verify operations in-between the program pulses. Each ISPP pulse gives rise, in particular, to an average $\langle \Delta V_T \rangle \simeq V_S$, corresponding to the injection of an average number $\langle \Delta n \rangle \simeq V_S \cdot C_{pp}/q$ of electrons into the cell floating gate. The decrease of C_{pp} following technology scaling has then led to the reduction of $\langle \Delta n \rangle$ and to accuracy limitations coming from the statistical fluctuation of the actual number Δn of electrons injected during the applied program pulses. The Δn statistics, displaying a sub-Poissonian behavior [51], [52], [101], gives rise, in fact, to a fluctuation of the cell ΔV_T coming from a program pulse, i.e. the program noise, which results in a relevant enlargement of the programmed V_T distribution of the NAND page beyond $V_{PV} + V_S$, which is the theoretical upper boundary for the V_T distribution in the case of an ideal noiseless program-and-verify algorithm. This is clearly

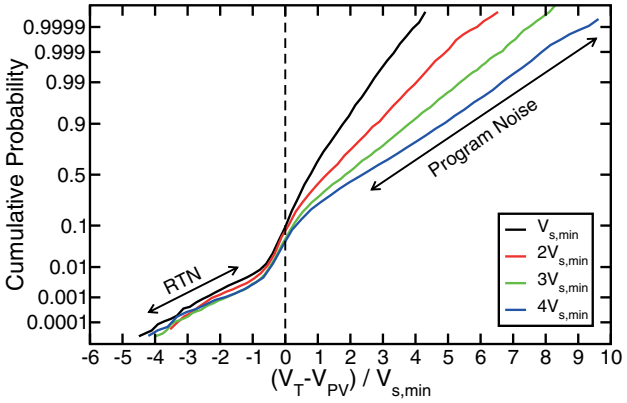


Fig. 21. ***Measured cumulative V_T distribution of a page of a deca-nanometer NAND Flash array as resulting from a program-and-verify algorithm making use of ISPP [65]. Results for different V_S multiple of a minimum value $V_{S,min}$ are shown.***

evident from Fig. 21, showing the programmed V_T distribution of a deca-nanometer NAND Flash array resulting from ISPP with different V_S multiple of a minimum value $V_{S,min}$. To preserve program accuracy in the presence of program noise, a reduction of V_S or more complex program-and-verify algorithms have been used [62]–[65].

As a final remark, note that electron emission from the floating gate during data retention and, in turn, its statistical fluctuation is not typically an issue for NAND Flash arrays. The use of a thick enough tunnel oxide, in fact, has always prevented electron emission from the floating gate from acting as a constraint to the array data retention time, which is limited, instead, by V_T instabilities after heavy cycling, as will be discussed in the next section. It is worth mentioning, moreover, that fluctuations in the electron emission would, anyway, play a minor role with respect to the statistical variability of the physical parameters of the array cells, considering also the sub-Poissonian behavior of the fluctuations at high $\langle \Delta V_T \rangle$ [105]. During a program operation, instead, the impact of cell-to-cell variability is minimized by the self-converging properties of ISPP, making C_{pp} the only parameter whose statistical spread might be relevant for the tightening of the V_T distribution. However, as demonstrated in [106], this is not the case for typical values of the C_{pp} spread.

2) Time-Dependent V_T Instabilities: Along with the impact of single electrons stored in the cell floating gate on V_T, technology scaling has also increased the role played on array operation by the parasitic capture and release of single charge carriers in the memory cell dielectrics [99]. The capture/release of electrons and holes in the cell tunnel oxide, in particular, has become one of the major constraints to array reliability, introducing time-dependent V_T instabilities displacing cell state from that set by the program operation, thus compromising data retention [107].

The origin of the increased concerns from the parasitic capture/release of charge carriers in tunnel-oxide defects is twofold. First, technology scaling has increased the impact on V_T of these charge carriers. As in the case of electrons in the floating gate, in fact, the reduction of the cell area and of intrinsic cell capacitances has resulted in the increase of the average impact on V_T of electrons and holes in the cell tunnel oxide [99]. In addition to that, and differently from the case of electrons in the cell floating gate now, scaling has also resulted in the increase of the standard deviation of the V_T shift coming from single electrons/holes in tunnel-oxide defects, introducing the possibility for some carriers to give rise to very large V_T shifts, even if with a low probability. This has been the direct consequence of the interplay between the localized nature of the trapped charge in tunnel-oxide defects and the percolative nature of the source-to-drain conduction resulting from atomistic doping and 3-D electrostatics in deca-nanometer MOS devices [108]–[117]. The random position of tunnel-oxide defects and of percolation paths in the channel allows, in fact, a single electron to induce a large V_T shift when trapped above a channel percolation path, as shown by the simulation results of Fig. 22. On the other hand, a negligible change of V_T appears when the same electron is trapped far from the major source-to-drain conduction paths in the channel. This results, typically, in an exponential or in a gamma distribution for the V_T shifts

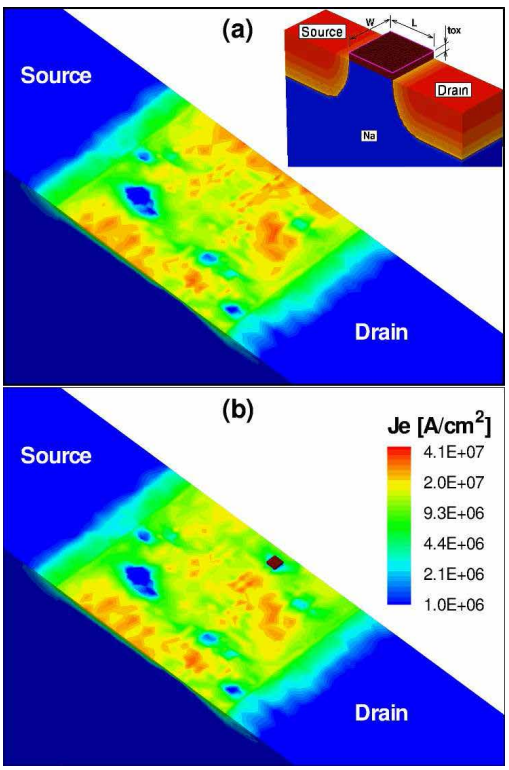


Fig. 22. ***(a) Simulated current density profile in the channel of a deca-nanometer MOS transistor (substrate doping has been implemented as atomistic). (b) Single trapped electron was introduced at the silicon/oxide interface at the position indicated by the square, resulting in a strong change of channel conduction. Reprinted from [108].***

coming from single-carrier trapping in oxide defects [108], [118]–[122].

The second reason that has led to increasing issues from the parasitic capture/release of charge carriers in the cell tunnel oxide has been the statistical variability of the number of tunnel-oxide defects. Note, in fact, that the decrease of cell dimensions has given rise to a reduction of the average number of defects in the cell dielectrics, making the fundamental statistical fluctuation of this number more and more important for array operation [123]–[127]. In particular, this fluctuation, along with the larger statistical dispersion of the V_T shift coming from the capture/release of a single carrier, has strongly increased the statistical dispersion of the V_T instabilities coming from the capture/release phenomena, making more challenging guaranteeing a correct data retention for the entire array.

Fig. 23 shows an example of the V_T instabilities coming from the capture/release of single charge carriers in the tunnel-oxide defects of deca-nanometer NAND Flash memory cells. The capture/release events correspond to the vertical transitions of V_T, whose amplitude and time occurrence are affected by statistical variability and are not predictable. From the time dynamics of the instabilities, two main phenomena are usually identified. The first is random telegraph noise (RTN) occurring when tunnel-oxide defects periodically capture and release single charge carriers from/to the channel of the memory cell, leading to a two-state fluctuation of V_T [108], [118]–[120], [128]–[141]. Owing to the large statistical dispersion of the capture and release time constants [107], [142]–[146], with the possibility of fluctuations in the microsecond timescale, RTN starts affecting the stability of cell V_T immediately after the program operation, introducing a tail of cells with V_T lower than the program-verify level V_{PV} already at the first read operation after program [65], [147], as shown in Fig. 21. Along data retention, moreover, RTN can further contribute to the enlargement of the page V_T distribution toward both higher and lower V_T values, owing to defects with longer and longer time constants coming into play [120]. The second form of time-dependent V_T instabilities in NAND Flash memories results from the so-called postcycling charge detrapping [123]–[126], [148]–[156]. This phenomenon arises from the capture of negative charge in the cell tunnel oxide during the program and erase operations and from the subsequent neutralization of this charge during the idle periods of data retention. The neutralization of the negative charge in the cell tunnel oxide results in the relaxation of cell V_T toward lower values as time elapses, as appearing for the cells shown in Fig. 23. This relaxation occurs over a logarithmic timescale and proceeds by a sequence of discrete events [124]–[126], [156]–[159]. This process enlarges the page V_T distribution toward lower V_T values, as shown in Fig. 24, by an amount increasing with the cell V_T level [154], [157].

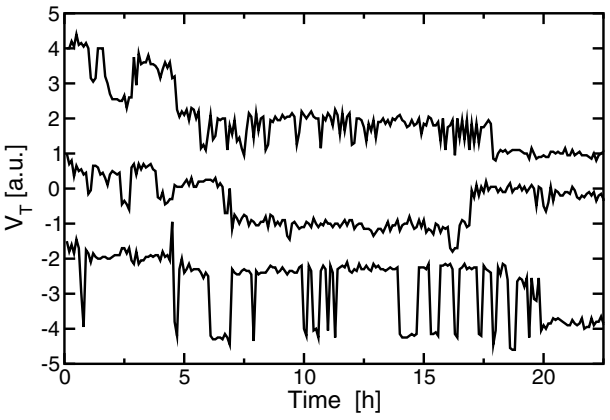


Fig. 23. ***Measured V_T of three programmed deca-nanometer NAND Flash cells for increasing data retention time. Cells have been heavily cycled prior to the experiment.***

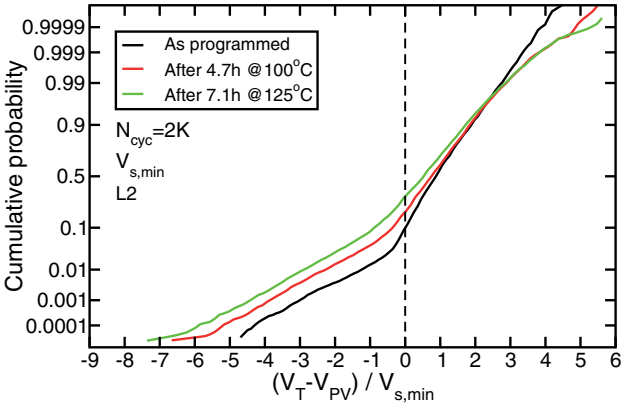


Fig. 24. ***Measured cumulative V_T distribution of a page of a deca-nanometer NAND Flash array as resulting from a program-and-verify operation making use of ISPP and after two subsequent data retention periods at high temperature [65]. N_{cyc} = 2 k program/erase cycles have been performed on the array prior to the experiment. The ISPP step amplitude is what has been called $V_{s,min}$ in Fig. 21.***

As a final remark, note that both the V_T instabilities coming from RTN and those coming from postcycling charge detrapping worsen with the number of program/erase cycles performed on the array [123], [126], [152], [160], constraining array endurance to a few 10^3 program/erase cycles in state-of-the-art technologies. Fig. 25 shows a comparison of the considered enlargement contributions to the width of the programmed V_T distribution in a NAND array after heavy cycling [9], [65].

3) Cell-to-Cell Electrostatic Interference: In order to keep the $4F^2$ unit cell area, the decrease of the feature size F of planar NAND Flash technologies has resulted not only in the reduction of the width W and length L of the floating-gate transistors, but also in the decrease of their distance in the memory array. Although this has allowed to achieve an extremely compact array and high GBSD, it has made the electrostatic interference between cells more and more relevant for array operation [9], [161]–[163]. The most

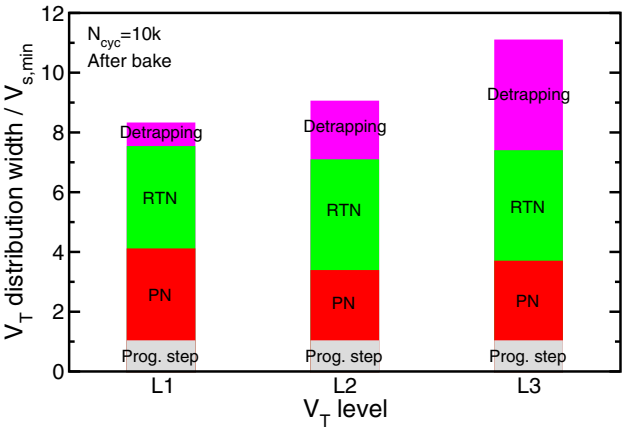


Fig. 25. *Comparison of the relative contributions to the width of the V_T distribution of a page of a deca-nanometer NAND Flash array after a program-and-verify operation making use of ISPP, for the three programmed levels of an MLC technology [65]. The ISPP step amplitude is what has been called $V_{s,min}$ in Fig. 21. PN is program noise.*

detrimental outcome of these electrostatic interference effects is that changes in the charge stored or trapped in a memory cell significantly affect V_T of its adjacent cells, compromising their data retention.

Notwithstanding that electrostatic interference between neighboring cells may show up any time during array operation, its most critical impact occurs when programming adjacent cells. Note, in fact, that the program operation gives rise to large changes of the floating-gate charge of the cells under program (aggressor cells) and this results in relevant parasitic changes of V_T of their adjacent cells (victim cells) [9]. In particular, the program operation on aggressor cells increases V_T of victim cells, adding more burdens on the design of the noise margins of MLC and TLC technologies. Fig. 26 shows an example of the cell-to-cell interference during program in a deca-nanometer NAND Flash chip implementing the EOBL architecture, highlighting the upward shift of the V_T distribution of a page of victim cells, initially on the same V_T level, when aggressor cells on their top or on their left/right are programmed. Although this shift might already be considered an issue for the correct operation of the memory array, things are further complicated by the unpredictable nature of the pattern of data to be stored in aggressor cells. Since the program operation may target a different V_T level for each of the aggressor cells, the parasitic V_T shift of the victim cells cannot be predicted, becoming statistically distributed among the cells. As a consequence, the V_T distribution of the victim cells not only shifts upwards, but also enlarges, increasing its spread. This enlargement partially vanishes the efforts made to tighten the V_T distribution during the previous program operation [9].

In order to reduce the parasitic electrostatic interference among the cells in NAND arrays, many technological solutions have been explored and implemented over the years. Among them, the most successful one has been surely the introduction of air gaps in-between cells belonging to adjacent WLs [53], [99], [165]–[167]. An example of these air gaps is presented in the transmission electron microscopy (TEM) image of a 25-nm NAND Flash array in Fig. 27 (the array structure on a vertical cross section along the string direction is shown). The main aim of the air gaps is to reduce the dielectric constant of the insulator between adjacent cells, reducing, in turn, their mutual parasitic capacitance. A second important solution has been the reduction of the height of the floating gates [11], [14], [53], resulting in a decrease of their lateral coupling, representing an important source of parasitic interference between neighboring cells [161]. Finally, smart programming sequences have also been developed to minimize the V_T shifts of neighboring cells as data are physically stored in the array [17], [59], [64], [66]–[68], [168], [169].

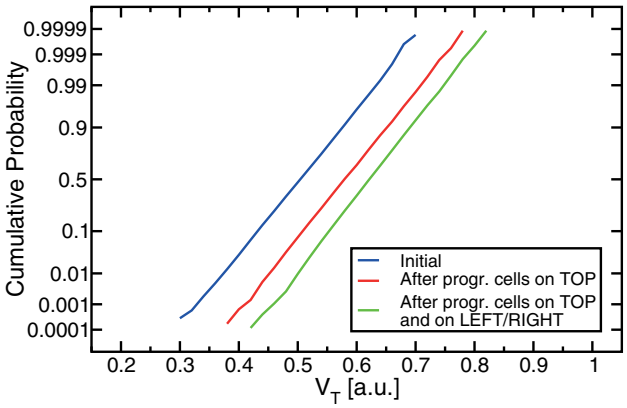


Fig. 26. *Measured cumulative V_T distribution of a page of a deca-nanometer NAND Flash chip implementing the EOBL architecture, before and after programming the cells on their top and on their left/right [164].*

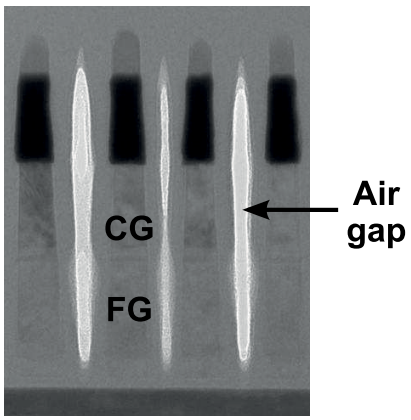


Fig. 27. *TEM image of a vertical cross section of a 25-nm NAND Flash array [99] along the string direction, highlighting the air gaps introduced to reduce parasitic cell-to-cell interference.*

As a final remark, note that, along with cell-to-cell electrostatic interference, disturbs on unselected cells during program operations have a nonnegligible impact on the V_T distribution of a page. Among the possible sources of disturbs, it is worth mentioning those coming from hot-carrier effects close to the source select transistors [16], [92], which have required the introduction of dummy cells in the NAND strings.

VI. FROM PLANAR TO 3-D ARRAY ARCHITECTURES

In this section, the exploitation of the vertical direction to further extend the historical growth of the GBSD of NAND Flash chips will be addressed, presenting the vertical-channel structure as the most important array solution to pursue an equivalent scaling of the technology. The benefits and drawbacks of this structure will be highlighted, considering process complexity and cost, its possibility to relieve the physical issues debated in Section V, the electrical performance of the memory array, and the possible limitations to the number of memory layers which can be vertically stacked. Finally, the impact of the transition from planar to 3-D array architectures on the major NAND Flash applications will be discussed.

A. The Vertical-Channel Cell Array

The first step into the 3-D era of the NAND Flash technology was taken in 2001, when the proof of concept of a vertical NAND string with two memory cells and two select transistors was reported [170]. The first 3-D integration of a memory chip was then presented in 2006 [171], [172] and consisted in vertically stacking two identical planar NAND arrays with horizontal strings and WLs. Notwithstanding the improvements in the integration density coming from this solution, the reduced benefits in terms of cost and process complexity of the simple piling up of planar arrays was soon recognized. Many different and more cost-effective solutions were then proposed, starting from the main idea that a reduction of the critical litographic steps can be achieved if either the silicon channels or the WLs of the NAND strings are integrated orthogonally to the wafer surface. This explains why the proposed solutions can be roughly divided in two main categories, namely, vertical-channel [4], [93], [173]–[180] and vertical-gate [181]–[186] 3-D technologies. Of these two categories, the vertical-channel solution has rapidly become the mainstream integration scheme for 3-D NAND Flash arrays, thanks to some relevant advantages that will be highlighted in the following discussion.

Fig. 28 shows the conceptual schematic structure of a vertical-channel 3-D NAND Flash array. The NAND strings are now orthogonal to the wafer surface, with a vertical cylindrical silicon channel running from the substrate to the BL and crossing a set of WLs acting as the control gates of the memory cell transistors. Similarly to the planar case, two select transistors are present at the top and the bottom of the NAND string and the two memory cells close to them are considered as dummy cells. Both the memory and the select transistors in the array have a gate-all-around structure, which allows the best electrostatic coupling between the (control-)gates and the silicon channels.

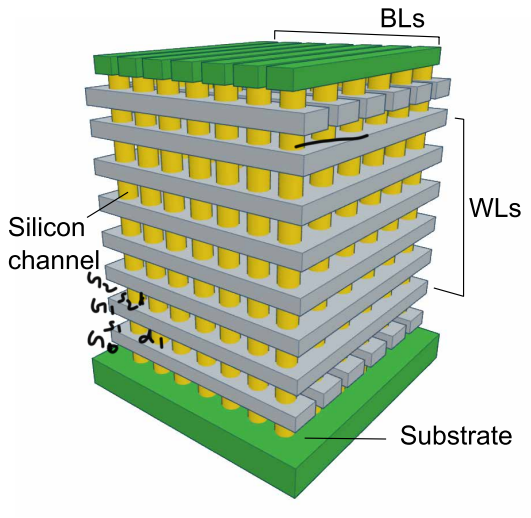


Fig. 28. *Conceptual schematic picture of a vertical-channel 3-D NAND Flash memory array.*

As allowed by any 3-D integration of the memory array, the vertical-channel architecture of Fig. 28 trades off the vertical stacking of many layers of memory cells for less demanding photolitographic process steps. In particular, as discussed in the next section, only 24 memory layers were needed to achieve equivalent planar GBSD comparable to those of state-of-the-art 15-nm planar NAND Flash chips when getting back to $F = 40$ nm, representing the minimum feature size of the single-patterning ArF immersion litography [187]. The possibility to avoid quadruple-patterning photolitographic techniques and to limit the use of double patterning only for very few process steps, typically BL definition, represents a huge benefit of this 3-D array integration, resulting in a strong reduction of its production costs. In this regard, moreover, note that the vertical-channel array structure of Fig. 28 is extremely cost effective [173], [188]. Array manufacturing, in fact, does not require patterning each memory layer individually, but involves the deposition of the vertical stack of materials needed to create the WL planes and their isolations and only a low number of photolitographic and etching steps. The fabrication of the vertical channels of the array, for instance, requires only etching cylindrical holes in the multilayer stack and filling them with silicon (after, of course, gate stack formation). A drawback of this integration scheme is that only polysilicon channels can be obtained, with consequent issues related to the lower electron/hole mobility with respect to mono-crystalline silicon,

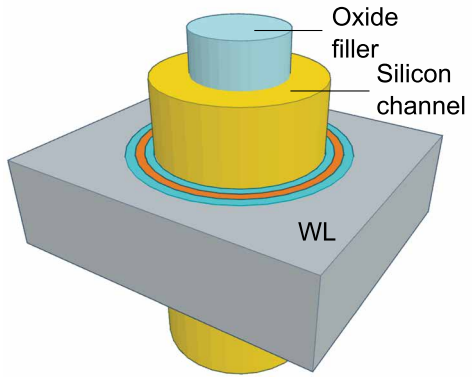


Fig. 29. *Schematic structure of a Macaroni-like vertical channel.*

to the charge trapping/detrapping at the grain boundaries and to the random position of the grain boundaries along the string [4], [93], [189], [190]. To partially relieve these issues, a Macaroni-like channel geometry has been proposed and adopted (see Fig. 29), reducing the polysilicon thickness to about 15 nm and filling the center of the Macaroni with a dielectric material [174], [175], [188].

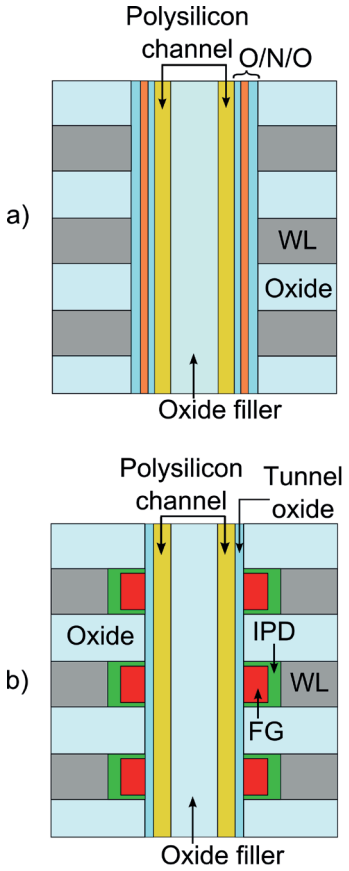


Fig. 30. *Schematic vertical cross section of a 3-D NAND Flash memory array based on (a) charge-trap storage; and (b) floating-gate storage. IPD is inter-poly dielectric and O/N/O stands for oxide/nitride/oxide.*

Despite sharing all of the major features discussed so far, different vertical-channel 3-D NAND Flash array technologies have been developed. The first and, likely, most important difference among them is in the material adopted to store charge in the memory cells, which can be either conventional polysilicon [4], [180] or a dielectric layer with a high density of trap centers, typically silicon nitride, implementing the so-called charge-trap storage [93], [173], [176], [177], [179]. Fig. 30 shows the schematic cross section of the NAND string in the two cases, offering benefits and drawbacks in terms of program/erase performance, reliability, and integration complexity. Another important difference among the vertical-channel 3-D NAND technologies proposed so far is in the WL material, which is highly doped polysilicon in gate-first [4], [173], [176], [180] and metal in gate-last [93], [177], [179] approaches.

B. Performance and Issues of 3-D NAND Flash Technologies

Referring again to the papers presented by technology manufacturers at the IEEE ISSCC [3], [180], [191]–[194], Fig. 31 shows how the first 3-D NAND Flash chips entered the GBSD plot. First, the figure reveals that all of the 3-D technologies proposed so far allow an equivalent planar GBSD higher than that of state-of-the-art planar NAND Flash chips, even when using only 24 memory layers [191]. Moreover, the figure proves the possibility for TLC 3-D technologies to follow and prolong in the years to come the historical scaling trend of planar TLC chips (blue dotted line). Following this trend will require an increasing number of memory layers (N_L) in the vertical-channel NAND string. In this regard, note that in the case of 3-D technologies (3) should be modified into

$$\mathrm{GBSD} = \frac{C_{\mathrm{chip}}}{A_{\mathrm{chip}}} = \frac{X_M \cdot \mathrm{BPC}}{gF^2} \cdot N_L. \tag{4}$$

By comparing (4) with the data points for 3-D technologies in Fig. 31 and taking $F = 40$ nm and $X_M = 0.7$, an average

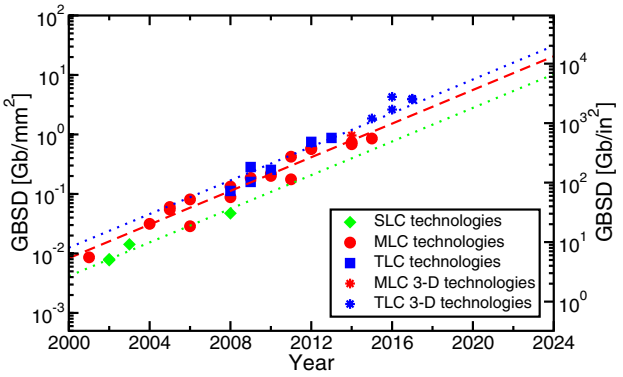


Fig. 31. *Trend of the GBSD of NAND Flash chips, including all the planar nodes between 2001 and 2015 and the 3-D array technologies presented up to 2017.*

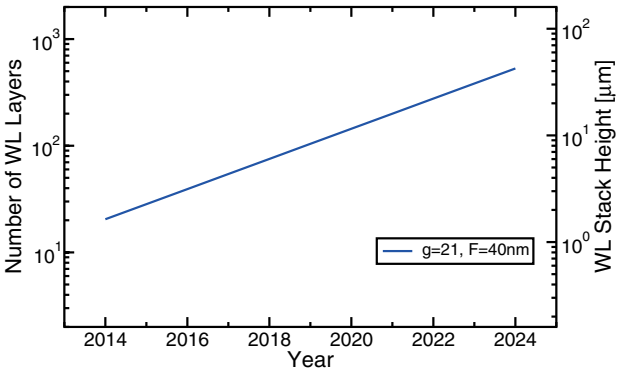


Fig. 32. *Calculated number of memory layers N_L needed in vertical-channel 3-D NAND Flash arrays to keep the historical GBSD trends of Fig. 31 in the next years. The right y-axis shows the total height of the WL stack assuming a WL pitch of 80 nm.*

gross cell area nearly equal to $21F^2$ can be derived. If this area is maintained in the next technology generations, keeping the historical GBSD scaling trends of Fig. 31 will require increasing N_L in the next years according to Fig. 32, predicting the integration of about 512 memory layers by 2024. The increase of N_L will result in the increase of the WL stack height, as shown on the right y-axis of Fig. 32, where a vertical WL pitch equal to 80 nm was assumed. Within the approximations involved in this projection, the WL stack height will reach about 40 μm by 2024, which is not far from today's chip height after wafer thinning.

The increase of N_L needed to prolong the historical GBSD trends in the next years will require to successfully address some critical challenges. First, the increase of the WL stack height and of the length of the vertical polysilicon channels will increase the string resistance, raising issues related to the consequent reduction of the reference read current. In light of these issues, it seems unlikely that the diameter of the silicon channel, about 70 nm today, may be reduced in the future, unless different channel materials are adopted [179], [195]. Besides, the reduction of the diameter of the silicon channel will be constrained by the tapered profile of the hole dug into the stack of materials needed to create the WL planes and their isolations. As schematically shown in Fig. 33, in fact, the etching process typically results in a reduction of the hole diameter moving from top to bottom, coming from an etching angle θ greater than 0. In order to avoid both too large cell-to-cell variability along the string and the increase of g with the increase of N_L, θ will have to be reduced with the development of the new technology nodes [93], [196]–[199]. Note, finally, that the constraints related to the string resistance and the etching angle could be relieved by the reduction of the vertical WL pitch. However, the possibility to reduce this pitch will be limited by the WL resistance, the WL-to-WL parasitic capacitance, the parasitic electrostatic interference between adjacent cells and, for charge-trap-based technologies, the lateral spreading of the stored charge during data retention [93], [200].

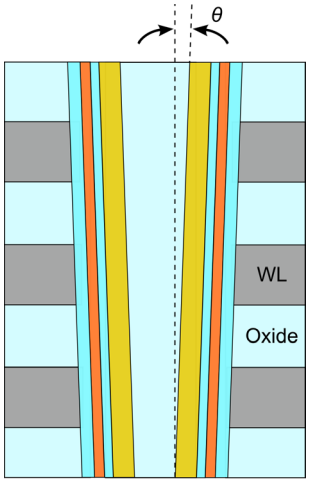


Fig. 33. *Schematic vertical cross section of a 3-D NAND Flash memory array based on charge-trap storage, highlighting the nonuniform cell geometry in the vertical direction coming from the nonideal etching angle θ.*

In terms of array performance, the 3-D NAND Flash technologies presented so far proved themselves able to overcome the major physical issues discussed in Section V-B. First, the larger cell area in the vertical-channel 3-D arrays allowed a significant reduction of the impact of single electrons/holes on cell V_T. This led, in turn, to the increase of the number of electrons needed for a given ΔV_T, as shown in Fig. 20, and to the reduction of the program noise issues [4], [82], [180]. In addition to that, the lower impact of single electrons/holes on V_T allowed also a strong reduction of V_T instabilities coming from RTN and charge trapping/detrapping in the tunnel oxide of the memory cells [4], [82], [107], [180], in spite of the issues arising from grain boundaries in the polysilcon channel [4], [93], [189], [190]. Besides, the gate-all-around cell geometry along with the increase of the WL pitch allowed to strongly reduce cell-to-cell electrostatic interference [4], [93], [179], [180], [191], as shown in Fig. 34.

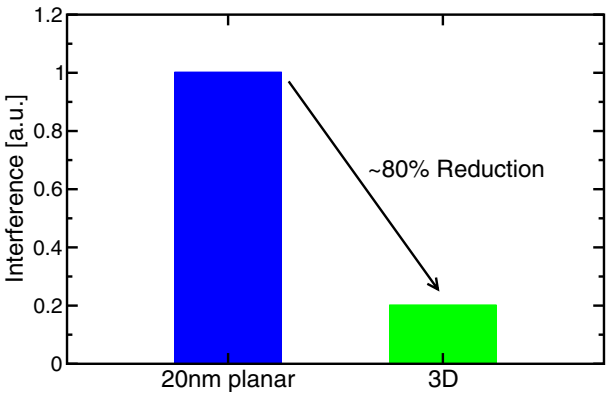


Fig. 34. *Relative comparison of the parasitic cell-to-cell interference effect between a planar 20-nm technology and a vertical-channel 3-D NAND Flash array [4].*

Note, in this regard, that in the 3-D array of Fig. 28 parasitic electrostatic interference is limited only to adjacent cells in the vertical direction.

The reduction of program noise, of time-dependent V_T instabilities, and of parasitic cell-to-cell interference allowed strong improvements in the program performance of the reported 3-D NAND Flash chips [3], [5], [6], [180], [191]–[194], thanks also to the favorable cell electrostatics arising from the cylindrical cell geometry [176], [196], [201], [202]. These improvements, in turn, will strongly impact the prospects of the NAND Flash technology in its key application fields, as will be discussed in the next section.

C. A Final Glance at Applications

Although the picture of the future scaling trends of 3-D NAND Flash chips and, in turn, of their main applications is still far from clear, the first results presented so far appear quite promising and able to keep the historical trends of the technology. The results of Fig. 31, in particular, suggest the possibility to maintain the constant increase of the chip GBSD over the next years, with a consequent increase of the chip capacity and reduction of the chip cost. Since this is the trend that allowed the NAND Flash technology to address new market applications, and, specifically, to SSDs to start displacing HDDs (see Fig. 16), there is the promise that this process will continue in the future. In terms of performance, the first 3-D chips show a step improvement relative to the latest planar nodes. This is obviously welcomed, for instance, by SSD applications, since these will, in turn, reproduce a similar step performance increase. The longer term performance trends of 3-D chips, however, are still to emerge and only some hypothetical projections can be drawn. Referring to SSDs, for instance, if the page size and the program throughput of a single chip keep increasing with the same scaling law of Figs. 12 and 13, for a given number of chips in the system the read and write bandwidth and the IOPS will improve nearly by a factor of 1.5 every two years, along with a doubling of the system capacity. The trends for a constant capacity SSD with technology scaling are less clear. One SSD architectural option would be to decrease the number of chips in the system by 50% over two years. This would enable a cost reduction but comes with a write performance reduction of up to 25% according to the scaling trends. Other architectural options exist to maintain performance but those would deviate from one or more trends and ultimately represent a cost versus performance tradeoff at the system level.

VII. CONCLUSION

In this paper, the recent evolution of the NAND Flash technology has been reviewed, starting from the historical scaling trends of planar arrays and coming to the recently introduced vertical-channel 3-D architectures. The review pointed out the constant pace of improvements in technology performance over the years, in terms of GBSD, C_{chip}, and data throughputs, which allowed the NAND Flash technology to successfully address broader and broader market applications and to become today the most important nonvolatile integrated memory solution. The possibility for the technology to keep this supremacy in the next years has been highlighted by the analysis of the first 3-D NAND Flash chips, which, despite introducing a paradigm shift in the scaling process, display all the potentialities to prolong the history of success of the technology in the next decade. ■

Acknowledgments

The authors would like to thank C. Miccoli, G. M. Paolucci, M. Bertuccio, S. Beltrami, P. Tessariol, and E. Camerlenghi from Micron Technology Inc. and D. Resnati and G. Nicosia from Politecnico di Milano for their valuable collaboration over the years. A special thanks goes also to all those people who contributed with their skills, expertise, and knowledge to the success of the NAND Flash technology and to the understanding of its basic physics, operation, and reliability.

REFERENCES

[1] F. Masuoka, M. Momodomi, Y. Iwata, and R. Shirota, "New ultra high density EPROM and Flash EEPROM with NAND structure cell," in *IEDM Tech. Dig.*, Dec. 1987, pp. 552–555.

[2] R. E. Fontana, Jr., G. M. Decad, and S. R. Hetzler, "Volumetric density trends (TB/in.^{3}) for storage components: TAPE, hard disk drives, NAND, and Blu-ray," *J. Appl. Phys.*, vol. 117, no. 17, pp. 17E301-1–17E301-4, Apr. 2015.

[3] J.-W. Im *et al.*, "7.2 A 128Gb 3b/cell V-NAND Flash memory with 1Gb/s I/O rate," in *Proc. ISSCC*, Feb. 2015, pp. 130–131.

[4] K. Parat and C. Dennison, "A floating gate based 3D NAND technology with CMOS under array," in *IEDM Tech. Dig.*, Dec. 2015, pp. 48–51.

[5] W. Jeong *et al.*, "A 128 Gb 3b/cell V-NAND Flash memory with 1 Gb/s I/O rate," *IEEE J. Solid-State Circuits*, vol. 51, no. 1, pp. 204–212, Jan. 2016.

[6] D. Kang *et al.*, "256 Gb 3 b/cell V-NAND Flash memory with 48 stacked WL layers," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 210–217, Jan. 2017.

[7] K. Kanda *et al.*, "A 120 mm^2 16Gb 4-MLC NAND Flash memory with 43 nm CMOS technology," in *Proc. ISSCC*, Feb. 2008, pp. 430–431.

[8] S. Aritome *et al.*, "A 0.67 μm^2 self-aligned shallow trench isolation cell (SA-STI cell) for 3 V-only 256 Mbit NAND EEPROMs," in *IEDM Tech. Dig.*, Dec. 1994, pp. 61–64.

[9] S. Aritome and T. Kikkawa, "Scaling challenge of self-aligned STI cell (SA-STI cell) for NAND Flash memories," *Solid-State Electron.*, vol. 82, pp. 54–62, Apr. 2013.

[10] S. Aritome, "Advanced Flash memory technology and trends for file storage application," in *IEDM Tech. Dig.*, Dec. 2000, pp. 763–766.

[11] M. Helm *et al.*, "A 128Gb MLC NAND-Flash device using 16 nm planar cell," in *Proc. ISSCC*, Feb. 2014, pp. 326–327.

[12] M. Momodomi *et al.*, "New device technologies for 5 V-only 4 Mb EEPROM with NAND structure cell," in *IEDM Tech. Dig.*, Dec. 1988, pp. 412–415.

[13] M. Momodomi *et al.*, "An experimental 4-Mbit CMOS EEPROM with a NAND-structured cell," *IEEE J. Solid-State Circuits*, vol. 24, no. 5, pp. 1238–1243, Oct. 1989.

[14] G. Naso *et al.*, "A 128Gb 3b/cell NAND Flash design using 20 nm planar-cell technology," in *Proc. ISSCC*, Feb. 2013, pp. 218–219.

[15] S. Choi *et al.*, "A 93.4 mm^2 64Gb MLC NAND-Flash memory with 16 nm CMOS technology," in *Proc. ISSCC*, Feb. 2014, pp. 328–329.

[16] J.-D. Lee, C.-K. Lee, M.-W. Lee, H.-S. Kim, K.-C. Park, and W.-S. Lee, "A new programming disturbance phenomenon in NAND Flash memory by source/drain hot-electrons generated by GIDL current," in *Proc. Non-Volatile Semiconductor Memory Workshop*, Feb. 2006, pp. 31–33.

[17] R. Micheloni and L. Crippa, "Multi-bit NAND Flash memories for ultra-high density storage devices," in *Advances in Non-Volatile Memory and Storage Technology*, Y. Nishi, Ed. Birmingham, U.K.: Woodhead Publishing, 2014, ch. 3, pp. 75–119.

[18] D. Richter, *Flash Memories: Economic Principles of Performance, Cost and Reliability Optimization*. New York, NY, USA: Springer-Verlag, 2013.

[19] H. Nakamura *et al.*, "A 125 mm^2 1Gb NAND Flash memory with 10 MB/s program throughput," in *Proc. ISSCC*, Feb. 2002, pp. 1–2.

[20] K. Takeuchi *et al.*, "A 56 nm CMOS 99 mm^2 8 Gb multi-level NAND Flash memory with 10 MB/s program throughput," in *Proc. ISSCC*, 2006, pp. 1–2.

[21] D. Nobunaga *et al.*, "A 50 nm 8Gb NAND Flash memory with 100 MB/s program throughput and 200 MB/s DDR interface," in *Proc. ISSCC*, Feb. 2008, pp. 426–427.

[22] H. Kim *et al.*, "A 159 mm^2 32 nm 32 Gb MLC NAND-Flash memory with 200 MB/s asynchronous DDR interface," in *Proc. ISSCC*, Feb. 2010, pp. 442–443.

[23] K.-T. Park *et al.*, "A 7MB/s 64 Gb 3-bit/cell DDR NAND Flash memory in 20 nm-node technology," in *Proc. ISSCC*, Feb. 2011, pp. 212–213.

[24] K. Fukuda *et al.*, "A 151 mm^2 64 Gb MLC NAND Flash memory in 24 nm CMOS technology," in *Proc. ISSCC*, 2011, pp. 198–199.

[25] M. Goldman, K. Pangal, G. Naso, and A. Goda, "25 nm 64 Gb 130 mm^2 3 bpc NAND Flash memory," in *Proc. IMW*, May 2011, pp. 1–4.

[26] K.-S. Shim *et al.*, "Inherent issues and challenges of program disturbance of 3D NAND Flash cell," in *Proc. IMW*, May 2012, pp. 1–4.

[27] R. Micheloni *et al.*, "A 4Gb 2b/cell NAND Flash memory with embedded 5b BCH ECC for 36 MB/s system read throughput," in *Proc. ISSCC*, Feb. 2006, pp. 1–2.

[28] R. Zeng *et al.*, "A 172 mm^2 32 Gb MLC NAND Flash memory in 34 nm CMOS," in *Proc. ISSCC*, Feb. 2009, pp. 236–237.

[29] M. Taub *et al.*, "A 90 nm 512 Mb 166 MHz multilevel cell Flash memory with 1.5 MByte/s programming," in *Proc. ISSCC*, Feb. 2005, pp. 54–55.

[30] C. Villa *et al.*, "A 125 MHz burst-mode flexible read-while-write 256 Mbit 2b/c 1.8V NOR Flash memory," in *Proc. ISSCC*, Feb. 2005, pp. 52–53.

[31] C. Villa *et al.*, "A 65 nm 1 Gb 2 b/cell NOR Flash with 2.25 MB/s program throughput and 400 MB/s DDR interface," in *Proc. ISSCC*, 2007, pp. 476–477.

[32] J. Javanifard *et al.*, "A 45 nm self-aligned-contact process 1 Gb NOR Flash with 5 MB/s program speed," in *Proc. ISSCC*, Feb. 2008, pp. 424–425.

[33] T. Tanaka *et al.*, "A quick intelligent page-programming architecture and a shielded bitline sensing method for 3 V-only NAND Flash memory," *IEEE J. Solid-State Circuits*, vol. 29, no. 11, pp. 1366–1373, Nov. 1994.

[34] M. Sako *et al.*, "A low-power 64Gb MLC NAND-Flash memory in 15 nm CMOS technology," in *Proc. ISSCC*, Feb. 2015, pp. 128–129.

[35] T.-S. Jung *et al.*, "A 117-mm^2 3.3-V only 128-Mb multilevel NAND Flash memory for mass storage applications," *IEEE J. Solid-State Circuits*, vol. 31, no. 11, pp. 1575–1583, Nov. 1996.

[36] R. Cernea *et al.*, "A 34 MB/s-program-throughput 16 Gb MLC NAND with all-bitline architecture in 56 nm," in *Proc. ISSCC*, Feb. 2008, pp. 420–421.

[37] R.-A. Cernea *et al.*, "A 34 MB/s MLC write throughput 16 Gb NAND with all bit line architecture on 56 nm technology," *IEEE J. Solid-State Circuits*, vol. 44, no. 1, pp. 186–194, Jan. 2009.

[38] Y. Li *et al.*, "128 Gb 3b/cell NAND Flash memory in 19 nm technology with 18 MB/s write rate and 400 Mb/s toggle mode," in *Proc. ISSCC*, Feb. 2012, pp. 436–437.

[39] N. Shibata *et al.*, "A 19 nm 112.8 mm^2 64 Gb multi-level Flash memory with 400 Mb/s/pin 1.8 V toggle mode interface," in *Proc. ISSCC*, Feb. 2012, pp. 422–423.

[40] D. Lee *et al.*, "A 64 Gb 533 Mb/s DDR interface MLC NAND Flash in sub-20 nm technology," in *Proc. ISSCC*, Feb. 2012, pp. 430–431.

[41] R. Kirisawa, S. Aritome, R. Nakayama, T. Endoh, R. Shirota, and F. Masuoka, "A NAND structured cell with a new programming technology for highly reliable 5 V-only Flash EEPROM," in *Symp. VLSI Tech. Dig.*, Jun. 1990, pp. 129–130.

[42] R. Shirota *et al.*, "A 2.3 μm^2 memory cell structure for 16 Mb NAND EEPROMs," in *IEDM Tech. Dig.*, Dec. 1990, pp. 103–106.

[43] T. Tanaka *et al.*, "A 4-Mbit NAND-EEPROM with tight programmed V_t distribution," in *Symp. VLSI Circuits Tech. Dig.*, Jun. 1990, pp. 105–106.

[44] S. Aritome, "NAND Flash innovations," *IEEE Solid-State Circuits Mag.*, vol. 5, no. 4, pp. 21–29, Apr. 2013.

[45] F. Masuoka, M. Asano, H. Iwahashi, T. Komuro, and S. Tanaka, "A new Flash E^2 PROM cell using triple polysilicon technology," in *IEDM Tech. Dig.*, Dec. 1984, pp. 464–467.

[46] G. Verma and N. Mielke, "Reliability performance of ETOX based Flash memories," in *Proc. IRPS*, Apr. 1988, pp. 158–166.

[47] S. Aritome *et al.*, "A reliable bi-polarity write/erase technology in Flash EEPROMs," in *IEDM Tech. Dig.*, Dec. 1990, pp. 111–114.

[48] K.-D. Suh *et al.*, "A 3.3 V 32 Mb NAND Flash memory with incremental step pulse programming scheme," *IEEE J. Solid-State Circuits*, vol. 30, no. 11, pp. 1149–1156, Nov. 1995.

[49] G. J. Hemink, T. Tanaka, T. Endoh, S. Aritome, and R. Shirota, "Fast and accurate programming method for multi-level NAND EEPROMs," in *Symp. VLSI Tech. Dig.*, Jun. 1995, pp. 129–130.

[50] T. Tanaka *et al.*, "A quick intelligent program architecture for 3 V-only NAND-EEPROMs," in *Symp. VLSI Circuits Dig. Tech. Papers*, Jun. 1992, pp. 20–21.

[51] C. Monzio Compagnoni, R. Gusmeroli, A. S. Spinelli, and A. Visconti, "Analytical model for the electron-injection statistics during programming of nanoscale NAND Flash memories," *IEEE Trans. Electron Devices*, vol. 55, no. 11, pp. 3192–3199, Nov. 2008.

[52] C. Monzio Compagnoni, A. S. Spinelli, R. Gusmeroli, S. Beltrami, A. Ghetti, and A. Visconti, "Ultimate accuracy for the NAND Flash program algorithm due to the electron injection statistics," *IEEE Trans. Electron Devices*, vol. 55, no. 10, pp. 2695–2702, Oct. 2008.

[53] D. James, "Recent advances in memory technology," in *Proc. ASMC*, May 2013, pp. 386–395.

[54] K.-D. Suh *et al.*, "A 3.3 V 32 Mb NAND Flash memory with incremental step pulse programming scheme," in *Proc. ISSCC*, Nov. 1995, pp. 128–129.

[55] K. Imamiya *et al.*, "A 130 mm^2 256 Mb NAND Flash with shallow trench isolation technology," in *Proc. ISSCC*, Sep. 1999, pp. 112–113.

[56] J. Lee *et al.*, "A 1.8 V 1 Gb NAND Flash memory with 0.12 μm STI process technology," in *Proc. ISSCC*, Oct. 2002, pp. 1–2.

[57] T. Cho *et al.*, "A 3.3 V 1 Gb multi-level NAND Flash memory with non-uniform threshold voltage distribution," in *Proc. ISSCC*, Sep. 2001, pp. 1–2.

[58] S. Lee *et al.*, "A 3.3 V 4 Gb four-level NAND Flash memory with 90 nm CMOS technology," in *Proc. ISSCC*, Feb. 2004, pp. 1–2.

[59] T. Hara *et al.*, "A 146 mm^2 8 Gb NAND Flash memory with 70 nm CMOS technology," in *Proc. ISSCC*, Jun. 2005, pp. 44–45.

[60] D.-S. Byeon *et al.*, "An 8 Gb multi-level NAND Flash memory with 63 nm STI CMOS process technology," in *Proc. ISSCC*, 2005, pp. 46–47.

[61] G. G. Marotta *et al.*, "A 3bit/cell 32Gb NAND Flash memory at 34nm with 6MB/s program throughput and with dynamic 2b/cell blocks configuration mode for a program throughput increase up to 13MB/s," in *Proc. ISSCC*, 2010, pp. 444–445.

[62] V. Moschiano, G. Santin, T. Vali, and M. Rossini, "Non-volatile multilevel memory cell programming," U.S. Patent 7692971 B2, Apr. 6 2010.

[63] C. Miccoli, C. Monzio Compagnoni, A. S. Spinelli, and A. L. Lacaita, "Investigation of the programming accuracy of a double-verify ISPP algorithm for nanoscale NAND Flash memories," in *Proc. IRPS*, 2011, pp. 833–838.

[64] S.-H. Shin *et al.*, "A new 3-bit programming algorithm using SLC-to-TLC migration for 8MB/s high performance TLC NAND Flash memory," in *Symp. VLSI Circuits Tech. Dig.*, 2012, pp. 132–133.

[65] G. M. Paolucci, C. Monzio Compagnoni, A. S. Spinelli, A. L. Lacaita, and A. Goda, "Fitting cells into a narrow V_T interval: Physical constraints along the lifetime of an extremely scaled NAND Flash memory array," *IEEE Trans. Electron Devices*, vol. 62, no. 5, pp. 1491–1497, May 2015.

[66] Y. Li *et al.*, "A 16Gb 3b/ Cell NAND Flash memory in 56nm with 8MB/s write rate," in *Proc. ISSCC*, 2008, pp. 506–507.

[67] C. Trinh *et al.*, "A 5.6MB/s 64Gb 4b/Cell NAND Flash memory in 43nm CMOS," in *Proc. ISSCC*, 2009, pp. 246–247.

[68] C. Lee *et al.*, "A 32-Gb MLC NAND Flash memory with Vth endurance enhancing schemes in 32 nm CMOS," in *Proc. ISSCC*, 2010, pp. 446–447.

[69] V. N. Kynett *et al.*, "An in-system reprogrammable 256K CMOS Flash memory," in *Proc. ISSCC*, 1988, pp. 132–133.

[70] J. Lee *et al.*, "A 1.8 V 2 Gb NAND Flash memory for mass storage applications," in *Proc. ISSCC*, 2003, pp. 1–2.

[71] S. Aritome *et al.*, "Extended data retention characteristics after more than 10^4 write and erase cycles in EEPROMs," in *Proc. IRPS*, 1990, pp. 259–264.

[72] G. Wong, "Market and applications for NAND Flash memories," in *Inside NAND Flash Memories*, R. Micheloni, L. Crippa, and A. Marelli, Eds., New York, NY, USA: Springer-Verlag, 2010, ch. 1, pp. 1–18.

[73] S.-H. Chang *et al.*, "A 48nm 32Gb 8-level NAND Flash memory with 5.5MB/s program throughput," in *Proc. ISSCC*, 2009, pp. 240–241.

[74] T. Futatsuyama *et al.*, "A 113mm^2 32Gb 3b/cell NAND Flash memory," in *Proc. ISSCC*, 2009, pp. 242–243.

[75] T. Y. Kim *et al.*, "A 32Gb MLC NAND Flash memory with Vth margin-expanding schemes in 26nm CMOS," in *Proc. ISSCC*, 2011, pp. 202–203.

[76] G. Campardo *et al.*, "40 mm^2 3-V-only 50-MHz 64-Mb 2-b/cell CHE NOR Flash memory," *IEEE J. Solid-State Circuits*, vol. 35, no. 11, pp. 1655–1667, Nov. 2000.

[77] P. Pavan, R. Bez, P. Olivo, and E. Zanoni, "Flash memory cells-an overview," *Proc. IEEE*, vol. 85, no. 8, pp. 1248–1271, Aug. 1997.

[78] R. Bez, E. Camerlenghi, A. Modelli, and A. Visconti, "Introduction to Flash memory," *Proc. IEEE*, vol. 91, no. 4, pp. 489–502, Apr. 2003.

[79] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[80] R. Sundaram *et al.*, "A 128 Mb NOR Flash memory with 3 MB/s program time and low-power write performance by using in-package inductor charge-pump," in *Proc. ISSCC*, 2005, pp. 50–51.

[81] G. Atwood, S. De Boer, K. Prall, and L. Somerville, "A semiconductor memory development and manufacturing perspective," in *Proc. ESSCIRC*, 2014, pp. 1–6.

[82] P. Cappelletti, "Non volatile memory evolution and revolution," in *IEDM Tech. Dig.*, Dec. 2015, pp. 241–244.

[83] R. E. Fontana, G. M. Decad, and S. R. Hetzler, "The impact of areal density and millions of square inches (MSI) of produced memory on petabyte shipments of TAPE, NAND Flash, and HDD storage class memories," in *Proc. MSST*, 2013, pp. 1–8.

[84] J. Kim, J. M. Kim, S. H. Noh, S. L. Min, and Y. Cho, "A space-efficient Flash translation layer for CompactFlash systems," *IEEE Trans. Consum. Electron.*, vol. 48, no. 2, pp. 366–375, May 2002.

[85] R. Micheloni, A. Marelli, and S. Commodaro, "NAND overview: From memory to systems," in *Inside NAND Flash Memories*, R. Micheloni, L. Crippa, and A. Marelli, Eds. New York, NY, USA: Springer-Verlag, 2010, ch. 2, pp. 19–53.

[86] Y.-H. Chang and L.-P. Chang, "Efficient wear leveling in NAND Flash memory," in *Inside Solid State Drives (SSDs)*, R. Micheloni, A. Marelli, and K. Eshghi, Eds.New York, NY, USA: Springer-Verlag, 2013, ch. 9, pp. 233–257.

[87] A. Jagmohan, M. Franceschini, and L. Lastras, "Write amplification reduction in NAND Flash through multi-write coding," in *Proc. MSST*, 2010, pp. 1–6.

[88] *P420m 2.5-Inch PCIe NAND Flash SSD, Product Data Sheet*, Micron Technol. Inc., 2012.

[89] *M510DC 2.5-Inch TGC Enterprise SATA NAND Flash SSD, Product Data Sheet*, Micron Technol. Inc., 2015.

[90] D. Glen, "Optimized client computing with dynamic write acceleration," Micron Tech. Marketing Brief, Micron Technology Inc., Boise, ID, USA, Tech. Rep., 2014.

[91] S. Im and D. Shin, "ComboFTL: Improving performance and lifespan of MLC Flash memory using SLC Flash buffer," *J. Syst. Architect.*, vol. 56, no. 12, pp. 641–653, Dec. 2010.

[92] Y. Park, J. Lee, S. S. Cho, G. Jin, and E. Jung, "Scaling and reliability of NAND Flash devices," in *Proc. IRPS*, 2014, pp. 2E.1.1–2E.1.4.

[93] S.-K. Park, "Technology scaling challenge and future prospects of DRAM and NAND Flash memory," in *Proc. IMW*, 2015, pp. 1–4.

[94] S. Aritome, *NAND Flash Memory Technologies*. Hoboken, NJ, USA: Wiley, 2015.

[95] D. Kwak *et al.*, "Integration technology of 30nm generation multi-level NAND Flash for 64Gb NAND Flash memory," in *Symp. VLSI Tech. Dig.*, Sep. 2007, pp. 12–13.

[96] B. Hwang *et al.*, "Smallest bit-line contact of 76nm pitch on NAND Flash cell by using reversal PR (photo resist) and SADP (self-align double patterning) process," in *Proc. SEMI*, 2007, pp. 356–358.

[97] B. T. Park *et al.*, "32nm 3-bit 32Gb NAND Flash memory with DPT (double patterning technology) process for mass production," in *Symp. VLSI Tech. Dig.*, 2010, pp. 125–126.

[98] J. Hwang *et al.*, "A middle-1X nm NAND Flash memory cell (M1X-NAND) with highly manufacturable integration technologies," in *IEDM Tech. Dig.*, 2011, pp. 199–202.

[99] K. Prall and K. Parat, "25nm 64Gb MLC NAND technology and scaling challenges," in *IEDM Tech. Dig.*, Dec. 2010, pp. 102–105.

[100] C. Monzio Compagnoni *et al.*, "First detection of single-electron charging of the floating gate of NAND Flash memory cells," *IEEE Electron Device Lett.*, vol. 36, no. 2, pp. 132–134, Feb. 2015.

[101] G. Nicosia *et al.*, "A single-electron analysis of NAND Flash memory programming," in *IEDM Tech. Dig.*, Dec. 2015, pp. 378–381.

[102] C. Monzio Compagnoni *et al.*, "First evidence for injection statistics accuracy limitations in NAND Flash constant-current Fowler-Nordheim programming," in *IEDM Tech. Dig.*, Dec. 2007, pp. 165–168.

[103] C. Friederich *et al.*, "Novel model for cell–system interaction (MCSI) in NAND Flash," in *IEDM Tech. Dig.*, Dec. 2008, pp. 831–834.

[104] C. Monzio Compagnoni, C. Miccoli, A. L. Lacaita, A. Marmiroli, A. S. Spinelli, and A. Visconti, "Impact of control-gate and floating-gate design on the electron-injection spread of decananometer NAND Flash memories," *IEEE Electron Device Lett.*, vol. 31, no. 11, pp. 1196–1198, Nov. 2010.

[105] C. Miccoli *et al.*, "Impact of neutral threshold-voltage spread and electron-emission statistics on data retention of nanoscale NAND Flash," *IEEE Electron Device Lett.*, vol. 31, no. 11, pp. 1202–1204, Nov. 2010.

[106] G. Nicosia *et al.*, "Investigation of the program operation of NAND Flash cells with a single-electron resolution," *IEEE Trans. Electron Devices*, vol. 63, no. 6, pp. 2360–2366, Jun. 2016.

[107] A. Goda, C. Miccoli, and C. Monzio Compagnoni, "Time dependent threshold-voltage fluctuations in NAND Flash memories: From basic physics to impact on array operation," in *IEDM Tech. Dig.*, Dec. 2015, pp. 374–377.

[108] A. Ghetti, C. Monzio Compagnoni, A. S. Spinelli, and A. Visconti, "Comprehensive analysis of random telegraph noise instability and its scaling in Deca–nanometer Flash memories," *IEEE Trans. Electron Devices*, vol. 56, no. 8, pp. 1746–1752, Aug. 2009.

[109] H.-S. Wong and Y. Taur, "Three-dimensional 'atomistic' simulation of discrete random dopant distribution effects in sub-0.1 μm MOSFET's," in *IEDM Tech. Dig.*, Dec. 1993, pp. 705–708.

[110] A. Asenov, A. R. Brown, J. H. Davies, and S. Saini, "Hierarchical approach to 'atomistic' 3-D MOSFET simulation," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 18, no. 11, pp. 1558–1565, Nov. 1999.

[111] N. Sano, K. Matsuzawa, M. Mukai, and N. Nakayama, "Cover image On discrete random dopant modeling in drift-diffusion simulations: Physical meaning of 'atomistic' dopants," *Microelectron. Rel.*, vol. 42, no. 2, pp. 189–199, 2002.

[112] A. Asenov, A. R. Brown, J. H. Davies, S. Kaya, and G. Slavcheva, "Simulation of intrinsic parameter fluctuations in decananometer and nanometer-scale MOSFETs," *IEEE Trans. Electron Devices*, vol. 50, no. 9, pp. 1837–1852, Sep. 2003.

[113] G. Roy, A. R. Brown, F. Adamu-Lema, S. Roy, and A. Asenov, "Simulation study of individual and combined sources of intrinsic parameter fluctuations in conventional Nano-MOSFETs," *IEEE Trans. Electron Devices*, vol. 53, no. 12, pp. 3063–3070, Dec. 2006.

[114] A. Ghetti, M. Bonanomi, C. Monzio Compagnoni, A. S. Spinelli, A. L. Lacaita, and A. Visconti, "Physical modeling of single-trap RTS statistical distribution in Flash memories," in *Proc. IRPS*, May 2008, pp. 610–615.

[115] A. Ghetti *et al.*, "Scaling trends for random telegraph noise in deca-nanometer Flash memories," in *IEDM Tech. Dig.*, Dec. 2008, pp. 835–838.

[116] A. Ghetti, S. M. Amoroso, A. Mauri, and C. Monzio Compagnoni, "Doping engineering for random telegraph noise suppression in deca-nanometer Flash memories," in *Proc. IMW*, May 2011, pp. 91–94.

[117] A. Ghetti, S. M. Amoroso, A. Mauri, and C. Monzio Compagnoni, "Impact of nonuniform doping on random telegraph noise in Flash memory devices," *IEEE Trans. Electron Devices*, vol. 59, no. 2, pp. 309–315, Feb. 2012.

[118] R. Gusmeroli *et al.*, "Defects spectroscopy in SiO_2 by statistical random telegraph noise analysis," in *IEDM Tech. Dig.*, Dec. 2006, pp. 483–486.

[119] K. Fukuda, Y. Shimizu, K. Amemiya, M. Kamoshida, and C. Hu, "Random telegraph noise in Flash memories—model and technology scaling," in *IEDM Tech. Dig.*, Dec. 2007, pp. 169–172.

[120] C. Monzio Compagnoni, R. Gusmeroli, A. S. Spinelli, A. L. Lacaita, M. Bonanomi, and A. Visconti, "Statistical model for random telegraph noise in Flash memories," *IEEE Trans. Electron Devices*, vol. 55, no. 1, pp. 388–395, Jan. 2008.

[121] J. Franco *et al.*, "Impact of single charged gate oxide defects on the performance and scaling of nanoscaled FETs," in *Proc. IRPS*, Apr. 2012, pp. 5A.4.1–5A.4.6.

[122] S. M. Amoroso *et al.*, "Investigation of the RTN distribution of nanoscale MOS devices from subthreshold to on-state," *IEEE Electron Device Lett.*, vol. 34, no. 5, pp. 683–685, May 2013.

[123] N. Mielke *et al.*, "Flash EEPROM threshold instabilities due to charge trapping during program/erase cycling," *IEEE Trans. Device Mater. Reliab.*, vol. 4, no. 3, pp. 335–344, Sep. 2004.

[124] C. Miccoli *et al.*, "Resolving discrete emission events: A new perspective for detrapping investigation in NAND Flash memories," in *Proc. IRPS*, Apr. 2013, pp. 3B.1.1–3B.1.6.

[125] G. M. Paolucci, C. Monzio Compagnoni, C. Miccoli, A. S. Spinelli, A. L. Lacaita, and A. Visconti, "Revisiting charge trapping/detrapping in Flash memories from a discrete and statistical standpoint—Part I: V_T instabilities," *IEEE Trans. Electron Devices*, vol. 61, no. 8, pp. 2802–2810, Aug. 2014.

[126] G. M. Paolucci, C. Monzio Compagnoni, C. Miccoli, A. S. Spinelli, A. L. Lacaita, and A. Visconti, "Revisiting charge trapping/detrapping in Flash memories from a discrete and statistical standpoint—Part II: On-field operation and distributed-cycling effects," *IEEE Trans. Electron Devices*, vol. 61, no. 8, pp. 2811–2819, Aug. 2014.

[127] B. Kaczer *et al.*, "The defect-centric perspective of device and circuit reliability—From individual defects to circuits," in *Proc. ESSDERC*, Sep. 2015, pp. 218–225.

[128] S. Machlup, "Noise in semiconductors: Spectrum of a two-parameter random signal," *J. Appl. Phys.*, vol. 25, pp. 341–343, May 1954.

[129] K. S. Ralls *et al.*, "Discrete resistance switching in submicrometer silicon inversion layers: Individual interface traps and low-frequency 1/*f*? noise," *Phys. Rev. Lett.*, vol. 52, pp. 228–231, Jan. 1984.

[130] M. J. Kirton and M. J. Uren, "Noise in solid-state microstructures: A new perspective on individual defects, interface states and low-frequency 1/*f* noise," *Adv. Phys.*, vol. 38, no. 4, pp. 367–468, 1989.

[131] K. K. Hung, P. K. Ko, C. Hu, and Y. C. Cheng, "Random telegraph noise of deep-submicrometer MOSFET's," *IEEE Electron Device Lett.*, vol. 11, pp. 90–92, Feb. 1990.

[132] A. Ohata, A. Toriumi, M. Iwase, and K. Natori, "Observation of random telegraph signals: Anomalous nature of defects at the Si/SiO_2 interface," *J. Appl. Phys.*, vol. 68, pp. 200–204, Feb. 1990.

[133] Z. Shi, J.-P. Mieville, and M. Dutoit, "Random telegraph signals in deep submicron n-MOSFET's," *IEEE Trans. Electron Devices*, vol. 41, no. 7, pp. 1161–1168, Jul. 1994.

[134] M.-H. Tsai and T.-P. Ma, "The impact of device scaling on the current fluctuations in MOSFET's," *IEEE Trans. Electron Devices*, vol. 41, no. 11, pp. 2061–2068, Nov. 1994.

[135] H. Kurata *et al.*, "The impact of random telegraph signals on the scaling of multilevel Flash memories," in *Symp. VLSI Circuit Dig.*, Jun. 2006, pp. 112–113.

[136] N. Tega *et al.*, "Anomalously large threshold voltage fluctuation by complex random telegraph signal in floating gate Flash memory," in *IEDM Tech. Dig.*, Dec. 2006, pp. 491–494.

[137] H. Miki *et al.*, "Quantitative analysis of random telegraph signals as fluctuations of threshold voltages in scaled Flash memory cells," in *Proc. IRPS*, Apr. 2007, pp. 29–35.

[138] A. S. Spinelli, C. Monzio Compagnoni, R. Gusmeroli, M. Ghidotti, and A. Visconti, "Investigation of the random telegraph noise instability in scaled Flash memory arrays," *Jpn. J. Appl. Phys.*, vol. 47, pp. 2598–2601, Apr. 2008.

[139] J. P. Campbell *et al.*, "Random telegraph noise in highly scaled nMOSFETs," in *Proc. IRPS*, Apr. 2009, pp. 382–388.

[140] J. P. Campbell *et al.*, "Large random telegraph noise in sub-threshold operation of nano-scale nMOSFETs," in *Proc. ICICDT*, May 2009, pp. 17–20.

[141] T. Nagumo, K. Takeuchi, T. Hase, and Y. Hayashi, "Statistical characterization of trap position, energy, amplitude and time constants by RTN measurement of multiple individual traps," in *IEDM Tech. Dig.*, Dec. 2010, pp. 628–631.

[142] A. Mauri *et al.*, "Impact of atomistic doping and 3D electrostatics on the variability of RTN time constants in Flash memories," in *IEDM Tech. Dig.*, Dec. 2011, pp. 405–408.

[143] N. Castellani, C. Monzio Compagnoni, A. Mauri, A. S. Spinelli, and A. L. Lacaita, "Three-dimensional electrostatics- and atomistic doping-induced variability of RTN time constants in nanoscale MOS devices—Part I: Physical investigation," *IEEE Trans. Electron Devices*, vol. 59, no. 9, pp. 2488–2494, Sep. 2012.

[144] C. Monzio Compagnoni, N. Castellani, A. Mauri, A. S. Spinelli, and A. L. Lacaita, "Three-dimensional electrostatics- and atomistic doping-induced variability of RTN time constants in nanoscale MOS devices—Part II: Spectroscopic implications," *IEEE Trans. Electron Devices*, vol. 59, no. 9, pp. 2495–2500, Sep. 2012.

[145] T. Grasser *et al.*, "The paradigm shift in understanding the bias temperature instability: From reaction-diffusion to switching oxide traps," *IEEE Trans. Electron Devices*, vol. 58, no. 11, pp. 3652–3666, Nov. 2011.

[146] T. Grasser, "Stochastic charge trapping in oxides: From random telegraph noise to bias temperature instabilities," *Microelectron. Reliab.*, vol. 52, no. 1, pp. 39–70, Jan. 2012.

[147] C. Monzio Compagnoni, M. Ghidotti, A. L. Lacaita, A. S. Spinelli, and A. Visconti, "Random telegraph noise effect on the programmed threshold-voltage distribution of Flash memories," *IEEE Electron Device Lett.*, vol. 30, no. 9, pp. 984–986, Sep. 2009.

[148] R. Yamada, Y. Mori, Y. Okuyama, J. Yugami, T. Nishimoto, and H. Kume, "Analysis of detrap current due to oxide traps to improve Flash memory retention," in *Proc. IRPS*, Apr. 2000, pp. 200–204.

[149] R. Yamada, T. Sekiguchi, Y. Okuyama, J. Yugami, and H. Kume, "A novel analysis method of threshold voltage shift due to detrap in a multi-level Flash memory," in *Symp. VLSI Tech. Dig.*, Jun. 2001, pp. 115–116.

[150] J.-D. Lee, J.-H. Choi, D. Park, and K. Kim, "Degradation of tunnel oxide by FN current stress and its effects on data retention characteristics of 90 nm NAND Flash memory cells," in *Proc. IRPS*, Apr. 2003, pp. 497–501.

[151] J.-D. Lee, J.-H. Choi, D. Park, and K. Kim, "Effects of interface trap generation and annihilation on the data retention characteristics of Flash memory cells," *IEEE Trans. Device Mater. Reliab.*, vol. 4, no. 1, pp. 110–117, Mar. 2004.

[152] N. Mielke, H. P. Belgal, A. Fazio, Q. Meng, and N. Righos, "Recovery effects in the distributed cycling of Flash memories," in *Proc. IRPS*, 2006, pp. 29–35.

[153] C. Monzio Compagnoni *et al.*, "Investigation of the threshold voltage instability after distributed cycling in nanoscale NAND Flash memory arrays," in *Proc. IRPS*, 2010, pp. 604–610.

[154] C. Miccoli, C. Monzio Compagnoni, S. Beltrami, A. S. Spinelli, and A. Visconti, "Threshold-voltage instability due to damage recovery in nanoscale NAND Flash memories," *IEEE Trans. Electron Devices*, vol. 58, no. 8, pp. 2406–2414, Aug. 2011.

[155] C. Miccoli *et al.*, "Assessment of distributed-cycling schemes on 45nm NOR Flash memory arrays," in *Proc. IRPS*, 2012, pp. 2A.1.1–2A.1.7.

[156] G. M. Paolucci *et al.*, "A new spectral approach to modeling charge trapping/detrapping in NAND Flash memories," in *Proc. IRPS*, 2014, pp. 2E.2.1–2E.2.6.

[157] D. Resnati *et al.*, "A step ahead toward a new microscopic picture for charge trapping/detrapping in Flash memories," in *Proc. IRPS*, 2016, pp. 6C-3-1–6C-3-7.

[158] D. Resnati, G. Nicosia, G. M. Paolucci, A. Visconti, and C. Monzio Compagnoni, "Cycling-induced charge trapping/detrapping in Flash memories—Part I: Experimental evidence," *IEEE Trans. Electron Devices*, vol. 63, no. 12, pp. 4753–4760, Dec. 2016.

[159] D. Resnati, G. Nicosia, G. M. Paolucci, A. Visconti, and C. Monzio Compagnoni, "Cycling-induced charge trapping/detrapping in Flash memories—Part II: Modeling," *IEEE Trans. Electron Devices*, vol. 63, no. 12, pp. 4761–4768, Dec. 2016.

[160] C. Monzio Compagnoni, A. S. Spinelli, S. Beltrami, M. Bonanomi, and A. Visconti, "Cycling effect on the random telegraph noise instabilities of nor and NAND Flash arrays," *IEEE Electron Device Lett.*, vol. 29, no. 8, pp. 941–943, Aug. 2008.

[161] J.-D. Lee, S.-H. Hur, and J.-D. Choi, "Effects of floating-gate interference on NAND Flash memory cell operation," *IEEE Electron Device Lett.*, vol. 23, no. 5, pp. 264–266, May 2002.

[162] A. Ghetti, L. Bortesi, and L. Vendrame, "3D simulation study of gate coupling and gate cross-interference in advanced floating gate non-volatile memories," *Solid-State Electron.*, vol. 49, no. 11, pp. 1805–1812, 2005.

[163] M. Park, K. Kim, J.-H. Park, and J.-H. Choi, "Direct field effect of neighboring cell transistor on cell-to-cell interference of NAND Flash cell arrays," *IEEE Electron Device Lett.*, vol. 30, no. 2, pp. 174–177, Feb. 2009.

[164] A. Spessot *et al.*, "Variability effects on the V_T distribution of nanoscale NAND Flash memories," in *Proc. IRPS*, 2010, pp. 970–974.

[165] D. Kang *et al.*, "The air spacer technology for improving the cell distribution in 1 giga bit NAND Flash memory," in *Proc. Non-Volatile Semiconductor Memory Workshop*, 2006, pp. 36–37.

[166] S. Kim, W. Cho, J. Kim, B. Lee, and S. Park, "Air-gap application and simulation results for low capacitance in 60nm NAND Flash memory," in *Proc. Non-Volatile Semiconductor Memory Workshop*, 2007, pp. 54–55.

[167] S. Lee, "Scaling challenges in NAND Flash device toward 10 nm technology," in *Proc. IMW*, 2012, pp. 1–4.

[168] N. Shibata *et al.*, "A 70 nm 16 Gb 16-level-cell NAND Flash memory," *IEEE J. Solid-State Circuits*, vol. 43, no. 4, pp. 929–937, Apr. 2008.

[169] D.-H. Lee and W. Sung, "Least squares based coupling cancelation for MLC NAND Flash memory with a small number of voltage sensing operations," *J. Signal Process. Syst.*, vol. 71, no. 3, pp. 189–200, 2013.

[170] T. Endoh *et al.*, "Novel ultrahigh-density Flash memory with a stacked-surrounding gate transistor (S-SGT) structured cell," in *IEDM Tech. Dig.*, 2001, pp. 33–36.

[171] S.-M. Jung *et al.*, "Three dimensionally stacked NAND Flash memory technology using stacking single crystal Si layers on ILD and TANOS structure for beyond 30nm node," in *IEDM Tech. Dig.*, Dec. 2006, pp. 37–40.

[172] E.-K. Lai *et al.*, "A multi-layer stackable thin-film transistor (TFT) NAND-type Flash memory," in *IEDM Tech. Dig.*, Dec. 2006, pp. 41–44.

[173] H. Tanaka *et al.*, "Bit cost scalable technology with punch and plug process for ultra high density Flash memory," in *VLSI Symp. Tech. Dig.*, 2007, pp. 14–15.

[174] Y. Fukuzumi *et al.*, "Optimal integration and characteristics of vertical array devices for ultra-high density, bit-cost scalable Flash memory," in *IEDM Tech. Dig.*, Dec. 2007, pp. 449–452.

[175] 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 Tech. Dig.*, Dec. 2009, pp. 27.3.1–27.3.4.

[176] 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 *Symp. VLSI Technol. Dig.*, 2009, pp. 136–137.

[177] J. Jang *et al.*, "Vertical cell array using TCAT (terabit cell array transistor) technology for ultra high density NAND Flash memory," in *Symp. VLSI Technol. Dig.*, 2009, pp. 192–193.

[178] S. Whang *et al.*, "Novel 3-dimensional dual control-gate with surrounding floating-gate (DC-SF) NAND Flash cell for 1Tb file storage application," in *IEDM Tech. Dig.*, Dec. 2010, pp. 668–671.

[179] E.-S. Choi and S.-K. Park, "Device considerations for high density and highly reliable 3D NAND Flash cell in near future," in *IEDM Tech. Dig.*, Dec. 2012, pp. 211–214.

[180] T. Tanaka *et al.*, "A 768 Gb 3b/cell 3D-floating-gate NAND Flash memory," in *Proc. ISSCC*, 2016, pp. 142–143.

[181] W. Kim *et al.*, "Multi-layered vertical gate NAND Flash overcoming stacking limit for terabit density storage," in *Symp. VLSI Technol. Dig.*, 2009, pp. 188–189.

[182] H.-T. Lue *et al.*, "A highly scalable 8-layer 3D vertical-gate (VG) TFT NAND Flash using junction-free buried channel BE-SONOS device," in *Symp. VLSI Technol. Dig.*, 2010, pp. 131–132.

[183] C.-H. Hung *et al.*, "A highly scalable vertical gate (VG) 3D NAND Flash with robust program disturb immunity using a novel PN diode decoding structure," in *Symp. VLSI Technol. Dig.*, 2011, pp. 68–69.

[184] S.-H. Chen *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 *IEDM Tech. Dig.*, Dec. 2012, pp. 21–24.

[185] H.-T. Lue, S.-H. Chen, Y.-H. Shih, K.-Y. Hsieh, and C.-Y. Lu, "Overview of 3D NAND Flash and progress of vertical gate (VG) architecture," in *Proc. IMW*, 2012, pp. 1–4.

[186] C.-P. Chen *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 *Symp. VLSI Technol. Dig.*, 2012, pp. 91–92.

[187] K.-T. Park *et al.*, "Three-dimensional 128 Gb MLC vertical NAND Flash memory with 24-WL stacked layers and 50 MB/s high-speed programming," *IEEE J. Solid-State Circuits*, vol. 50, no. 1, pp. 204–213, Jan. 2015.

[188] R. Shirota, "Developments in 3D-NAND Flash technology," in *Advances in Non-Volatile Memory and Storage Technology*. Amsterdam, The Netherlands: Elsevier, 2014, ch. 2, pp. 27–74.

[189] Y.-H. Hsiao *et al.*, "Modeling the impact of random grain boundary traps on the electrical behavior of vertical gate 3-D NAND Flash memory devices," *IEEE Trans. Electron Devices*, vol. 61, no. 6, pp. 2064–2070, Jun. 2014.

[190] R. Degraeve *et al.*, "Statistical poly-Si grain boundary model with discrete charging defects and its 2D and 3D implementation for vertical 3D NAND channels," in *IEDM Tech. Dig.*, Dec. 2015, pp. 121–124.

[191] 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 *Proc. ISSCC*, 2014, pp. 334–335.

[192] D. Kang *et al.*, "256Gb 3b/cell V-NAND Flash memory with 48 stacked WL layers," in *Proc. ISSCC*, 2016, pp. 130–131.

[193] R. Yamashita *et al.*, "A 512 Gb 3 b/cell Flash memory on 64-word-line-layer BICS technology," in *Proc. ISSCC*, 2017, pp. 196–197.

[194] C. Kim *et al.*, "A 512 Gb 3 b/cell 64-stacked WL 3D V-NAND Flash memory," in *Proc. ISSCC*, 2017, pp. 202–203.

[195] E. Capogreco *et al.*, "MOVPE $In_{1-x}Ga_xAs$ high mobility channel for 3-D NAND memory," in *IEDM Tech. Dig.*, Dec. 2015, pp. 40–43.

[196] A. Goda and K. Parat, "Scaling directions for 2D and 3D NAND cells," in *IEDM Tech. Dig.*, Dec. 2012, pp. 2.1.1–2.1.4.

[197] A. Arreghini, G. Van den Bosch, G. S. Kar, and J. Van Houdt, "Ultimate scaling projection of cylindrical 3D SONOS devices," in *Proc. IMW*, 2012, pp. 168–169.

[198] Y. Yanagihara, K. Miyaji, and K. Takeuchi, "Control gate length, spacing and stacked layer number design for 3D-stackable NAND Flash memory," in *Proc. IMW*, 2012, pp. 1–4.

[199] A. J. Walker, "A rigorous 3-D NAND Flash cost analysis," *IEEE Trans. Semicond. Manuf.*, vol. 26, no. 4, pp. 619–625, Nov. 2013.

[200] A. Maconi *et al.*, "Comprehensive investigation of the impact of lateral charge migration on retention performance of planar and 3D SONOS devices," *Solid-State Electron.*, vol. 74, pp. 64–70, Aug. 2012.

[201] E. Nowak *et al.*, "In-depth analysis of 3D silicon nanowire SONOS memory characteristics by TCAD simulations," in *IMW Tech. Dig.*, 2010, pp. 116–119.

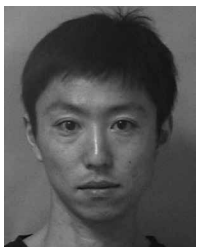
[202] S. M. Amoroso, C. Monzio Compagnoni, A. Mauri, A. Maconi, A. S. Spinelli, and A. L. Lacaita, "Semi-analytical model for the transient operation of gate-all-around charge-trap memories," *IEEE Trans. Electron Devices*, vol. 58, no. 9, pp. 3116–3123, Sep. 2011.

ABOUT THE AUTHORS

Christian Monzio Compagnoni (Senior Member, IEEE) received the Laurea (*cum laude*) degree in electronic engineering and the Ph.D. degree in information technology from the Politecnico di Milano, Milan, Italy, in 2001 and 2005, respectively.

Since 2006, he has been with the Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, first in the capacity of Assistant Professor (from April 2006 to March 2015) and then of Professor of Electronic Engineering (since March 2015). His research activities have been devoted to the basic physics involved in the operation and in the reliability of solid-state technologies for data storage, with emphasis on deca-nanometer NOR and NAND Flash memories and on emerging memories based on discrete traps. On these topics, he authored more than 100 papers published in international journals and conference proceedings and he holds two U.S. patents.

Dr. Monzio Compagnoni is a recipient of five awards at the IEEE-International Reliability Physics Symposium ("Outstanding paper Award" in 2008, "Best Student Paper Award" in 2012, 2013, and 2014, and "Best Poster Award" in 2015) and served in the technical program committee of the IEEE International Reliability Physics Symposium in 2009 ("Memory" committee), 2010 ("Memory" committee), and 2016 ("Memory and Product IC Reliability" committee). Since November 2015, he has been serving as an Associate Editor of the IEEE TRANSACTIONS ON ELECTRON DEVICES for the area "Memory Devices and Technology."



Akira Goda received the B.S. and M.S. degrees in chemistry from the University of Tokyo, Tokyo, Japan, in 1995 and 1997, respectively.

In 1997, he joined Toshiba Corporation Semiconductor Company, Japan, where he worked on the device and process development of NAND Flash memory. In 2004, he joined Micron Technology, Inc., Boise, ID, USA, where he is working on the development and scaling of 2-D and 3-D NAND Flash cells. He is a Fellow at Micron Technology.

Alessandro S. Spinelli (Senior Member, IEEE) was born in Bergamo, Italy, in 1966. He received the Laurea (*cum laude*) and Ph.D. degrees in electronics engineering from the Politecnico di Milano, Milano, Italy.

He was with the Università degli Studi dell'Insubria, Como, Italy, as an Associate Professor of Electronics in 1998 and was a Visiting Professor with the Institute National Polytechnique de Grenoble, Grenoble, France, in 2001. Since 2006, he has been a Full Professor of Electronics with the Politecnico di Milano. He has conducted experimental and theoretical research in electronics instrumentation and microelectronics, coauthoring more than 200 papers published in international journals or presented at international conferences. His current research interests include experimental characterization and modeling of nonvolatile memory cells reliability.

Dr. Spinelli has served on the technical committees of the IEEE International Electron Devices Meeting and the International Reliability Physics Symposium.

Peter Feeley received the B.S. and M.S. degrees in electrical engineering from the University of Idaho, Moscow, ID, USA.

He is the Director of the Media Architecture Group at Micron Technology, Inc., Boise, ID, USA. His responsibilities include error control and media management algorithms for SSDs and other system level products. Prior to joining Micron, he was director of ASIC development at a fabless semiconductor company focusing on storage controllers and communication link-layer error management. He holds over 50 patents.

Andrea L. Lacaita (Fellow, IEEE) received the Laurea degree (*cum laude*) in nuclear engineering from Politecnico di Milano, Milan, Italy, in 1985.

From 1987 to 1992, he was a Researcher of the CNR (Italian National Research Council). In 1992, he was appointed EE Associate Professor at Politecnico di Milano. He has been EE Full Professor since 2000. He was a Visiting Scientist at the AT&T Bell Laboratories, Murray Hill, NJ, USA (1989–1990), IBM T.J. Watson Research Center, Yorktown Heights, NY, USA (1999), and Data Storage Institute, Singapore (2011). He is author or coauthor of more than 350 papers published in international journals or presented at international conferences. His current research interests include reliability and modeling of nonvolatile memory technologies, sensors, and MEMS.

Dr. Lacaita has been serving in conference committees: IEEE International Electron Devices Meeting (IEDM) IEEE Symposium on VLSI Technology, European Solid-State Device Research Conference (ESSDERC), and Non-volatile Memory Technology Symposium (NVMTS) .

Angelo Visconti (Member, IEEE) was born in Italy in 1966. He received the Laurea degree in physics (*cum laude*) from the University of Milano, Milano, Italy, in 1997.

In 1997, he joined the Non-Volatile Memory Process Development Group, STMicroelectronics (later Numonyx and afterward Micron), Agrate Brianza, Italy. Since then, he has been involved in the development of many generations of high density Flash memory process. His current research interest is focused on reliability of high density memory devices. He is author or coauthor of more than 150 scientific publications and two books and holds 22 patents. He served on many technical committees in International Electron Devices Meeting (IEDM), International Reliability Physics Symposium (IRPS), Joint Electron Device Engineering Council (JEDEC), and Automotive Electronics Council (AEC) and co-chaired several memory sessions at IRPS and IEDM. He won several awards from IRPS, JEDEC, Radiation Effects on Components & Systems Conference (RADECS), and Nuclear and Space Radiation Effects Conference (NSREC).