Reliability of NAND Flash Arrays: A Review of What the 2-D–to–3-D Transition Meant

Christian Monzio Compagnoni, *Senior Member, IEEE*,
and Alessandro S. Spinelli, *Senior Member, IEEE*

(Invited Paper)

***Abstract*—This paper reviews what changed in the reliability of NAND Flash memory arrays after the paradigm shift in technology evolution determined by the transition from 2-D to 3-D integration schemes. Starting from a quick glance at the fundamentals of raw array reliability, the reasons for its worsening with the evolution of 2-D technologies will be discussed, focusing on the physical phenomena which contributed more to that outcome. By exploring the dependence of the magnitude of these phenomena on cell and array parameters, the abrupt improvements achieved from the 3-D transition in terms of raw array reliability will then be explained, highlighting also that these improvements were turned into new opportunities for the technology. Finally, the physical issues specific to 3-D arrays will be addressed, providing a glimpse of the challenges that the NAND Flash technology will have to face from the standpoint of array reliability in the near future.**

***Index Terms*—3-D NAND Flash arrays, Flash memories, semiconductor device modeling, semiconductor device reliability.**

I. Introduction

THE NAND Flash technology has become, today, the undisputed leader in the nonvolatile memory market, largely overcoming the hard-disk drive (HDD) technology in terms of revenues [1]. This outcome has been determined by the capability of the NAND Flash solution to address quite a variety of applications better than any other storage technology, thanks to successful tradeoffs among cost, performance, and reliability. At the heart of all that there is, of course, the possibility to steadily increase the integration density of the NAND array and, in turn, the memory capacity per chip. This is clearly proved in Fig. 1 in terms of gross bit storage density (GBSD), i.e., the ratio between the storage capacity and the total chip area, of the NAND Flash chips presented at the IEEE International Solid-State Circuits Conference (IEEE ISSCC) since 2001 (see [2] for further details on the analysis methodology).

Up to ~2015, the GBSD increase was achieved, first of all, through a constant pace miniaturization of the memory cells in 2-D (planar) arrays, whose schematic structure is shown in Fig. 2. This miniaturization trend is evident from Fig. 3, where the feature size of the 2-D technologies used to make the chips considered in Fig. 1 has been reported. A steady reduction of the feature size nearly equal to a factor $\sqrt{2}$ every 2 years can be extracted from the data points [2], leading to the ~15-nm node in the middle of the 2010s decade. The second driving force which was exploited to increase the GBSD of 2-D NAND Flash chips was the increase of the number of bits of information stored per cell. In particular, the transition from 1-bit/cell storage [single-level cell (SLC) technologies] to 2-bit/cell storage [multi-level cell (MLC) technologies] allowed a step

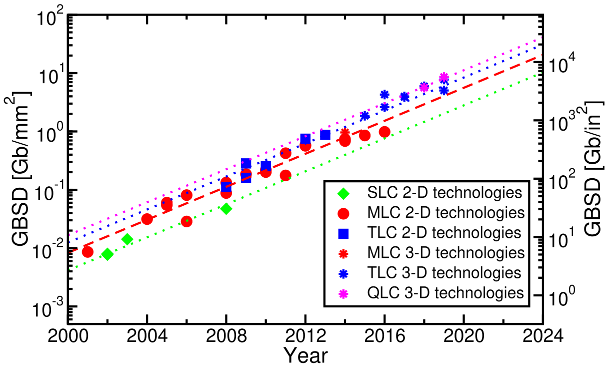

Fig. 1. GBSD of the SLC, MLC, and TLC 2-D and MLC, TLC, and QLC 3-D NAND Flash chips presented at the IEEE ISSCC since 2001 (figure updated with respect to what reported in [2], with the inclusion of [4], [15]–[19]). See [2] for the methodology used to extract the reported trend lines.

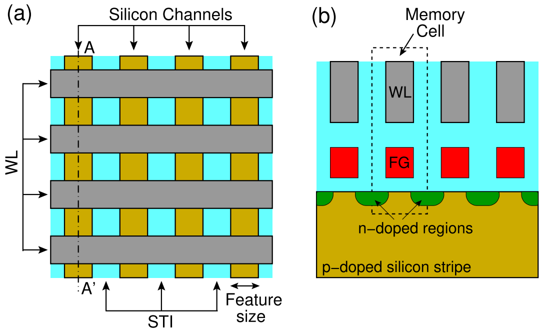

Fig. 2. Schematic for a simple 2-D NAND Flash array, showing (a) array top view and (b) its vertical cross section along one of the string channels (cut A-A′ in (a)). STI stands for shallow trench isolation. The pitch of WLs and of the silicon channels is typically twice the value of the technology feature size. See [2] for further details on the array structure.

Manuscript received March 18, 2019; revised May 7, 2019; accepted May 15, 2019. Date of publication June 11, 2019; date of current version October 29, 2019. The review of this paper was arranged by Editor R. Koval. *(Corresponding author: Alessandro S. Spinelli.)*

The authors are with the Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, 20133 Milan, Italy (e-mail: christian.monzio@polimi.it; alessandro.spinelli@polimi.it).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TED.2019.2917785

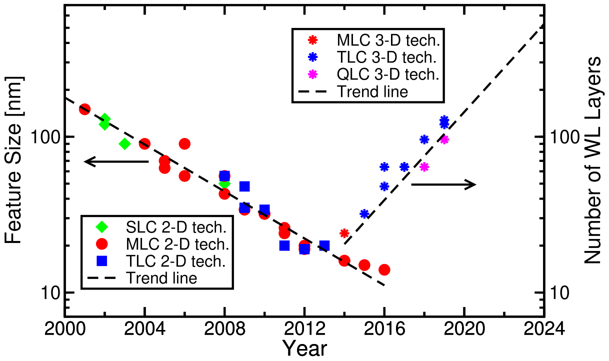

Fig. 3. Feature size and number of WL layers of the technologies used to make, respectively, the 2-D and 3-D NAND Flash chips considered in Fig. 1. See [2] for the methodology used to extract the reported trend lines.

increase of the GBSD by a factor 2 (from the green to the red trend line in Fig. 1). The following exploitation of 3-bit/cell storage [triple-level cell (TLC) technologies] allowed to increase the GBSD by a factor 1.5 with respect to MLC technologies (from the red to the blue trend line in Fig. 1). Thanks to the combined action of a small feature size and a number of bits per cell greater than 1, GBSD close to 1 Gbit/mm^2 were reached by the last 2-D NAND Flash chips [3], [4].

As typically happens with evolutionary approaches, however, the improvements achieved by the 2-D NAND Flash technology over the years came at the expense of an increase in process and system complexity [5]–[7]. This was due to the need not only to reduce the technology feature size but also to tackle the general worsening with cell miniaturization of some physical issues which could have compromised array reliability and, in turn, performance [5], [8]–[10]. As a result of that, around 2015 a paradigm shift in the integration and evolution of the NAND Flash technology was considered more favorable than the miniaturization-based approach. This paradigm shift consisted in moving from 2-D to 3-D arrays, with the idea that high GBSD could be reached even with relatively large memory cells if many of them were stacked along the vertical direction. In particular, the integration of vertical-channel NAND strings through a punch and plug process rapidly became the mainstream solution for 3-D NAND Flash arrays [11]–[14], thanks to its cost-effectiveness and to some relevant benefits in array performance and reliability [14]. Fig. 4 shows a simple schematic structure for such 3-D arrays, highlighting that NAND strings consist here in the series connection of gate-all-around (GAA) memory transistors resulting from the intersection of cylindrical polysilicon channels running orthogonally to the substrate surface with a number of word line (WL) planes.

Vertical-channel 3-D arrays came as a relief to the process and system complexity of the NAND Flash technology. First of all, they allowed to use conventional single-patterning ArF immersion lithography, whose minimum feature size is ~40 nm [20], for most of the process flow, limiting the need for double-patterning techniques [21]. Then, they allowed to reduce the impact on array reliability and performance of the main physical issues that constrained the operation of 2-D arrays [10], [14], [21], [22], thanks also to a cell size much larger than that of the last planar nodes. Their being less miniaturized, anyway, did not preclude 3-D technologies from successfully prolonging the historical GBSD trends of 2-D technologies over the second half of the 2010s decade, as shown in Fig. 1. As previously stated, this was achieved by relying just on the steady increase in the number of cells, i.e., WL layers, stacked along the vertical direction, as shown in Fig. 3. Figs. 1 and 3 reveal, besides, that 3-D technologies made 3-bit/cell storage their elective solution, with the possibility to exploit 4 bit/cell storage [quadruple-level cell (QLC) technologies] to achieve another 33% increase in the chip GBSD [15], [23].

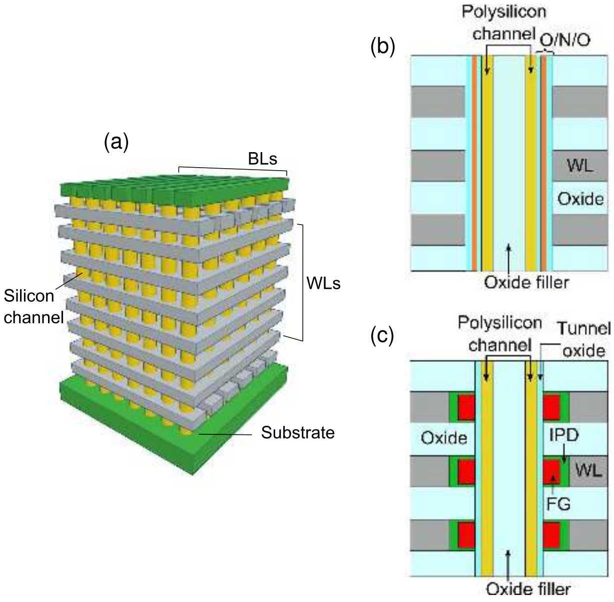

Fig. 4. Schematic for a simple 3-D vertical-channel NAND Flash array, showing (a) array structure in a 3-D perspective and (b) and (c) its vertical cross section along one of the string channels in the case of a charge-trap based and a floating-gate-based technology (O/N/O stands for oxide/nitride/oxide and IPD stands for interpoly dielectric). The oxide layers isolating the WLs and the gate-stack in-between the WLs and the silicon channels are not shown in (a) for better figure readability. The outer diameter of the polysilicon channel and the WL pitch is currently about 70–80 and 50–60 nm, respectively. See [2] for further details on the array structure.

In this paper, the success of the 2-D-to-3-D transition will be discussed from the standpoint of array reliability. The reasons for the worsening of the raw reliability of 2-D arrays with technology evolution will be, first, reviewed. Then, the attention will be drawn on the physical phenomena which played a major role on that worsening, pointing out the dependence of their magnitude on cell and array parameters. In so doing, the abrupt improvements achieved in the raw array reliability thanks to the 3-D transition will be explained. Finally, some of the future challenges that 3-D technologies will have to face to keep their equivalent scaling running at full speed will be examined.

II. Fundamentals of NAND Array Reliability

In general terms, the reliability of a nonvolatile memory array represents its capability to store some data and allow for their correct retrieval after a relatively long stretch of time

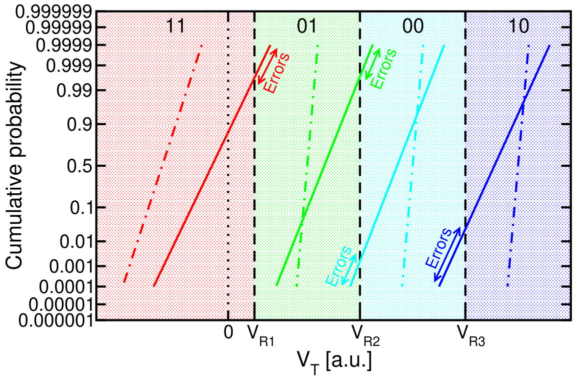

Fig. 5. Schematic for the V_T states (regions of different color) and their associated bits in an MLC NAND Flash array, with the read voltage levels V_{RX} discriminating them highlighted. Dashed-dotted lines represent ideal V_T distributions for the cells in the V_T state of the same color, while solid lines are an example of distributions affected by the physical phenomena giving rise to write, disturb, and data retention errors. Errors occur when the distribution of a V_T state exceeds the read voltage levels bounding it, as shown by the example. Note that, in the Flash memory field, a constant current criterion is typically used to define cell V_T.

(months or years) even though, in the meanwhile, the array is required to perform some other operations or its working conditions, e.g., temperature, are changed. Reliability worsens with the increase in the number of program and erase cycles that cells in the array underwent, since these operations are typically a source of strong electrical stress for the array and its cells. Consequently, a maximum number of program and erase cycles, called the array *endurance*, can be safely performed on each cell of the array. Within that number, some reliability specifications are guaranteed, such as a minimum data retention time, under certain temperature conditions and with a maximum bit error rate in data storage and retrieval. Beyond that number, instead, the array may lose its capability either to store data or to retrieve them according to its reliability specifications.

In the case of NAND Flash arrays, data storage consists in moving the threshold voltage (V_T) of floating-gate (or charge-trap) transistors from a previously set erased (low) state. In particular, cell V_T may be either left to its erased state or increased up to one among $2^n - 1$ possible programmed (high) states, with n being the number of bits stored per cell, as shown in Fig. 5. Program and erase operations are performed by charge exchange between the channel and the storage layer of the transistors, achieved through uniform Fowler–Nordheim tunneling over the channel area of the devices. Data retrieval consists in read operations determining the V_T state of the transistors through comparisons of device V_T with the read voltage levels V_{RX} bounding the V_T states, highlighted in Fig. 5. See [2] for a more detailed review of the read/program/erase schemes in a NAND Flash array.

Errors in the operation of the NAND array occur when the detected V_T state of the memory cells during data retrieval is different from that targeted during data storage (see Fig. 5). These errors may happen due to essentially three reasons: bad placement of cell V_T during programming (*write errors*), poor immunity of cell V_T to other operations performed in the array (*disturb errors*), and time-dependent instability of cell V_T (*data retention errors*). Of course, the classification of errors into write, disturb, and data retention errors would be rigorous if only one of the three possible error sources were present at a time. In fact, errors in memory operation are the outcome of the evolution of cell V_T from the program operation to the data retrieval request, with the placement of cell V_T, its immunity to other operations performed in the array and its time-dependent instabilities concurrently playing a role in determining whether the detected cell V_T state matches in the end the originally aimed state or not. Anyway, the previous classification can be considered as a way to identify at least the dominant reason for the detected error. In the following, the physical issues responsible for errors will be summarized.

A. Write Errors

Write errors are due to the inaccurate increase of cell V_T during programming. Since program-and-verify algorithms made of multiple programming pulses with intermediate read operations are used to stop the increase of cell V_T when this overcomes a selected program-verify level (V_{PV}) [2], inaccuracies in V_T placement typically consist in cell over-programming. This may arise from fundamental fluctuations in the number of electrons tunneling from the channel to the storage layer of the memory cells (typically called *program noise*) [24], [25], from anomalous or erratic tunneling in the presence of tunnel-oxide defects [10], [26], [27], or from abrupt changes in the tunneling rate due to floating-gate depletion effects [10], [28], [29].

B. Disturb Errors

Disturb errors are due to changes in the V_T of a previously programmed (*victim*) cell when read or program operations are performed on other cells in the array. These changes typically consist in a parasitic increase of the victim cell V_T, with two main origins. The first is the undesired injection of electrons from the channel to the storage layer of the victim cell determined by the positive WL bias involved in read and program operations. In this case, the V_T change of the victim cell, which is properly said to come from a *read disturb* [7], [30], [31] or a *program disturb* [7], [30], [32], [33], depends on the cell memory state, on the possible presence in the cell tunnel-oxide of defects enhancing its low-field conduction and on cell position along the NAND string. The second main origin of changes in the V_T of the victim cell is the modification of its electrostatic and conduction environment during read resulting from the change of the V_T state of other cells in the array. When the cells whose memory state is modified are those adjacent to the victim cell, the latter experiences a V_T increase generally attributed to cell-to-cell electrostatic interference [8], [34], [35]. When, instead, the increase of the victim cell V_T comes from the increase of the series resistance of its nonadjacent cells along the same NAND string when these are programmed, it is usually classified as a backpattern effect [7], [8].

C. Data Retention Errors

Data retention errors are due to changes in cell V_T when no operations are performed in the array. A large variety of physical phenomena may contribute to these errors, but the most relevant among them are those arising from tunnel-oxide defects. Some of these defects may act as a source of trap-assisted tunneling (TAT), enhancing the low-field conductivity of the tunnel-oxide and changing the amount of charge in the cell storage layer over a timescale that is precluded to direct tunneling [36]–[39]. Other defects, instead, may largely affect cell V_T just by changing their occupancy and without modifying the amount of charge in the cell storage layer. This is the case of tunnel-oxide defects giving rise to random telegraph noise (RTN) fluctuations of cell V_T by repeatedly capturing and emitting electrons or holes over time [40]–[43]. This is also the case of charged tunnel-oxide defects that, through a relaxation process resulting in the end in their neutralization and healing, introduce discrete V_T shifts along a preferential direction, a phenomenology typically referred to with a rather oversimplified terminology as charge detrapping from the tunnel-oxide [44]–[49].

From the previous discussion, it should be clearly evident that tunnel-oxide defects may impact the reliability of NAND Flash arrays in many different ways. Starting from a non-neglible *native* value, the concentration of these defects grows significantly with the increase of the number of program/erase cycles performed on the cells [note, in this regard, that the contribution to the tunnel-oxide current resulting from the increase of TAT with program/erase cycles is typically called stress-induced leakage current (SILC)] [37], [40], [48], [50]. As discussed at the beginning of this section, this worsens array reliability up to the point that the required specifications can no longer be guaranteed or that data storage is no longer feasible at all. The latter case results from an erase or program failure, meaning that in the maximum time slot allowed for either one or the other operation this cannot be completed [7]. It is worth mentioning that a relevant role on that can also be played by the increase with program/erase cycles of the interface state density at the channel/tunnel-oxide interface [50].

III. Impact of the Evolution of 2-D Technologies and of the 3-D Transition on Array Reliability

The evolution of 2-D NAND Flash technologies according to what is shown in Figs. 1 and 3 had an extremely critical impact on many of the physical phenomena mentioned in the previous section as constraints to array reliability. A proof of that is directly given by the constant strengthening of the error-correction codes (ECCs) [7], [30] adopted to guarantee the array reliability specifications with technology evolution. An example of this strengthening is given in Fig. 6, where the trend of the maximum number of bits correctable per code word by the Bose-Chaudhuri-Hocquenghem (BCH) ECC algorithm implemented in a family of NAND Flash devices is reported [5]. The figure highlights that not only the reduction of the technology node feature size but also the increase in the number of bits stored per cell required a stronger ECC. The increase in the number of bits per cell, in fact, results in the reduction of the width of the intervals associated to the V_T states, making the criterion for an error in memory operation to occur more severe (see Fig. 5). This typically means that worse *raw*, i.e., uncorrected, reliability specifications must be tolerated during array operation (e.g., a higher raw bit error rate for a given data retention time and working temperature) by adopting stronger system-level correction schemes. The reduction of the width of the V_T states in MLC and TLC technologies is due to the constraints set by both the program (maximum voltage, time) and reliability specifications to the highest V_T state. In particular, from the standpoint of array reliability, limitations to the increase of the highest V_T state come from the worsening of the disturb and data retention issues. The increase of the highest V_T state, for instance, results in the need for higher WL voltages during the program and read operations, leading to stronger program and read disturbs. In addition, due to the field acceleration experienced by TAT and charge detrapping [37], [39], [48], the increase in the highest V_T state also results in stronger V_T instabilities, limiting the possibility to guarantee the array data retention specifications.

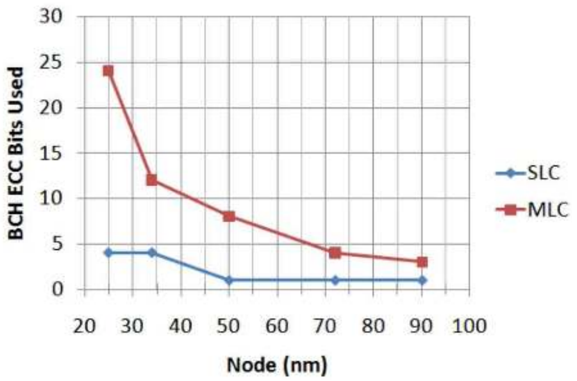

Fig. 6. Evolution of the maximum number of bits correctable per code word by the BCH ECC algorithm implemented in a family of 2-D NAND Flash devices. Reprinted from [5].

Similar to the increase of the number of bits per cell, even the reduction of the feature size of 2-D NAND Flash technologies resulted in the need to tolerate a worse raw array reliability by means of stronger system-level corrections, as highlighted from the ECC trends in Fig. 6. Cell-to-cell electrostatic interference, RTN, charge detrapping, and program noise can be considered as the physical phenomena which have contributed more to this outcome, becoming, in the end, the most critical constraints to array reliability. The reason for that is twofold: all of them are almost intrinsic phenomena and their magnitude largely increases with the decrease of the planar cell and array dimensions. With the reduction of the technology feature size below 100 nm, these phenomena started giving rise to changes in cell V_T that during array operation could become, with high probability, a relevant fraction of the width of the array V_T states in the case of MLC and TLC storage, which, in the meantime, became the mainstream solutions for 2-D NAND Flash arrays.

To better clarify why RTN and program noise can be considered as almost intrinsic phenomena, Fig. 7 shows the cumulative distribution of cell V_T measured on a decananometer 2-D NAND Flash array after a program-and-verify operation making use of incremental step pulse programming (ISPP),

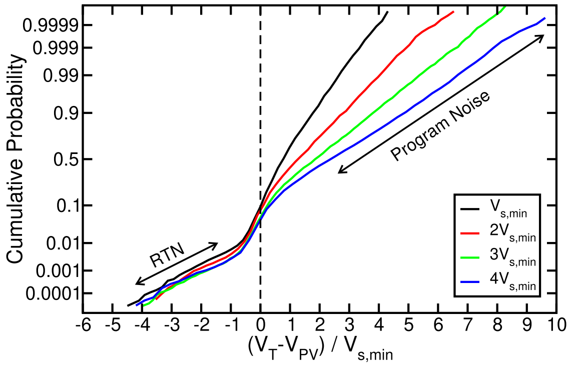

Fig. 7. Measured cumulative V_T distribution of a page of a decananometer NAND Flash array, as resulting from a program-and-verify algorithm making use of ISPP with step multiple of a minimum value $V_{s,\min}$. Data were achieved on a fresh (uncycled) array. Reprinted from [2].

with step amplitude multiple of a minimum value $V_{s,\min}$ [2]. In principle, the algorithm should tighten the V_T distribution in an interval between $V_{\rm PV}$ and $V_{\rm PV}$ plus the step amplitude V_s [32], [52]. Actually, instead, when the V_T distribution is read at the end of the program operation, with high probability cells are found at V_T levels not only above $V_{\rm PV} + V_s$ but also below the program-verify level $V_{\rm PV}$. The former effect is mainly due to program noise, resulting in a more-than-required number of electrons transferred to cell floating gate during the last ISPP pulse [52]. The latter effect, instead, is due to RTN. As time elapses, in fact, RTN may move the V_T of some cells below $V_{\rm PV}$, even though it was above this level at the last verify operation of the program algorithm [53]. In this regard, note that RTN may also contribute to the enlargement of the programmed V_T distribution rightwards in the case of small V_s or reduced program noise [53] and that the overall height of the RTN tails further increases after program/erase cycling, due to the growth of the number of defects in the cell tunnel-oxide. Fig. 8 shows, in a similar way, that charge detrapping during data retention affects the entire V_T distribution, determining a change of its average value and an increase of its spread [44], [46]–[49], [51]. The same holds for cell-to-cell electrostatic interference [8], [54]. Note, instead, that this is not the case for other reliability issues. For instance, TAT introduces a low-voltage tail in the V_T distribution of cells on the highest memory states during data retention and a high-voltage tail in the V_T distribution of cells on the lowest memory states during data retention and read disturb [30], [37], [55], [56]. These tails are due to cells having a configuration of defects in their tunnel-oxide particularly unlucky and determining a large increase of its low-field conductivity. For given data retention and read disturb specifications, the height of the tails in the V_T distributions is strongly affected by the tunnel-oxide thickness [56] and by the position of the highest and lowest V_T states [37]. Differently from the case of RTN, whose tail in the V_T distribution can be only marginally reduced by process optimizations [57], these strong dependences allowed an easy way to curb the impact of TAT on array operation and to maintain it as an *anomalous* phenomenon. In particular, keeping the tunnel-oxide relatively thick has been the key solution to keep TAT under control. As a result, cell miniaturization came along with just a weak reduction of the tunnel-oxide thickness down to 6–7 nm [58]. This reduction, to some extent, was allowed by the strengthening of ECCs and by the consequent increase of the maximum tolerable raw bit error rate that accompanied technology evolution.

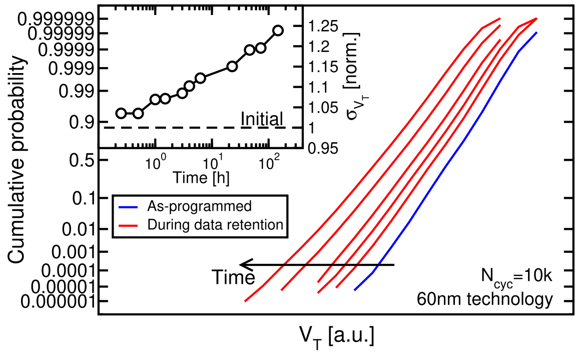

Fig. 8. Measured cumulative V_T distribution of a page of a decananometer NAND Flash array as resulting from the program operation and after increasing time during data retention at room temperature. Changes in the distribution over time are due to charge detrapping from the tunnel-oxide of the memory cells. The array cells underwent $N_{\rm cyc} = 10$ k program/erase cycles prior to the test. The final program operation was performed with a program-and-verify algorithm making use of ISPP with a loose step amplitude [51]. The inset shows the evolution of the standard deviation of the distribution over time.

A. Magnitude of the Most Relevant Issues for Array Reliability

Given the almost intrinsic nature of the phenomena, the strong increase in the magnitude of program noise, RTN, charge detrapping, and cell-to-cell electrostatic interference with array miniaturization resulted in severe limitations to the raw reliability of 2-D NAND Flash technologies. In the following, this increase will be explained through the expression for some representative parameters of the phenomena, discussing the benefits coming from the recent transition to 3-D arrays.

1) Program Noise: Program noise introduces a fundamental limitation to the accuracy of the program-and-verify algorithms of NAND Flash arrays. This limitation stems from the statistical nature of the process ruling electron injection from the channel to the storage layer of the memory cells. Due to this statistical nature, the number of electrons injected into the storage layer per programming pulse during an ISPP operation is affected by variability and so is the resulting V_T shift (ΔV_T). In [25], the standard deviation of ΔV_T, which can be considered as the parameter highlighting the magnitude of the phenomenon, was demonstrated to obey the following simple formula:

$$\sigma_{\Delta V_T} = \sqrt{\frac{q}{\gamma\, C_{\rm pp}}\left(1 - e^{-\gamma\,\langle \Delta V_T \rangle}\right)} \qquad (1)$$

where q is the elementary charge, $C_{\rm pp}$ is the control-gate–to–floating-gate capacitance, γ is related to the slope of the tunneling current versus floating-gate voltage in a semilogarithmic plot, and $\langle \Delta V_T \rangle$ is the average ΔV_T resulting from the programming pulse. The previous equation correctly accounts for the impact on the injection process of the electrostatic feedback following electron storage in the cell gate stack and was

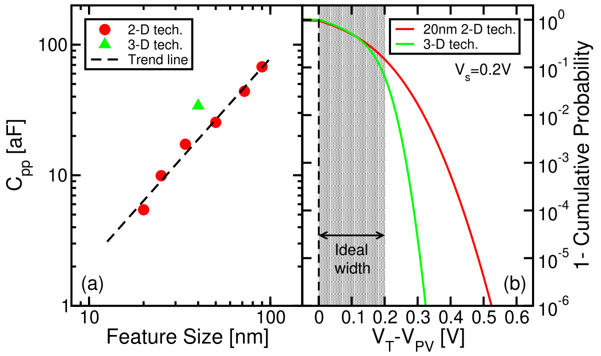

Fig. 9. (a) Cell C_{pp} as a function of the feature size of 2-D and 3-D NAND Flash technologies (data from [14]). (b) Calculated V_T distribution resulting from a program-and-verify operation making use of ISSP with $V_s = 0.2$ V, in the case of a 20-nm 2-D technology and a 3-D technology. Only the effects of program noise have been accounted for in the calculation.

validated against both floating-gate and charge-trap devices (in the latter case, straightforward changes in the definition of C_{pp} and γ are needed) [25], [59]. In the case of small $\langle \Delta V_T \rangle$, i.e., small number of stored electrons per programming pulse, the electrostatic feedback following electron storage plays a negligible role and the equation can be simplified into that resulting for a pure Poissonian injection:

$$\text{Program Noise} : \sigma_{\Delta V_T} \simeq \sqrt{\frac{q\langle \Delta V_T \rangle}{C_{\text{pp}}}} \tag{2}$$

Equation (2) is an approximation typically good enough to handle ISPP operations with V_s of few hundreds of millivolt, which are the most interesting when increasing the number of bits stored per cell. The equation clearly shows that the miniaturization of cell dimensions in 2-D NAND Flash arrays resulted in the increase of the magnitude of program noise via the decrease of C_{pp}. This latter decrease, which followed the reduction of the length (L) and width (W) of the floating-gate transistors, is shown in Fig. 9(a). With C_{pp} reaching the aF scale in the last planar nodes, a severe reduction of the accuracy of the program-and-verify algorithms was unavoidable, as shown in Fig. 9(b). Then, to reach the needed programming accuracy, either a reduction of V_s or more complex program-and-verify algorithms [60] were adopted.

From the standpoint of program noise, the transition to 3-D technologies strongly relieved array reliability. This was a direct consequence of the large increase in cell dimensions that followed this transition [2] and that brought C_{pp} back to a few tens of aF, as shown in Fig. 9(a) (from the figure, an increase of C_{pp} by nearly a factor 6 can be extracted when moving from the 20-nm 2-D node to 3-D technologies). Thanks to the larger C_{pp}, a strong reduction of $\sigma_{\Delta V_T}$ and, in turn, of the impact of program noise on the width of the programmed V_T distribution was achieved in 3-D arrays. This is highlighted in Fig. 9(b), where the programmed V_T distribution in the 3-D case is much narrower and closer to the ideal limit of $V_{\text{PV}} + V_s$ than in the 2-D case.

2) RTN: A very common and simple way to address RTN instabilities in NAND Flash arrays is by considering the statistical distribution of the ΔV_T experienced by cells between two consecutive read operations. An example of that distribution is reported in Fig. 10(a) for a decananometer 2-D NAND Flash array [61]. Exponential tails departing toward the positive and negative ΔV_T direction clearly appear from the figure, representing a marked feature of RTN fluctuations of cell V_T over time [41], [42]. Although the height of these tails depends on the average number of active RTN defects in the cell tunnel-oxide, and then on process quality and the number of program/erase cycles that cells underwent, the slope of the tails is directly related to the statistical distribution of the V_T shift induced by a single RTN trap (ΔV_T^1) [41], [42]. This latter distribution arises from the variability in the impact on V_T of localized tunnel-oxide defects placed at different positions over the cell channel area, in the presence of the typical percolative source-to-drain conduction of nanoscale devices [43]. In the most common case, the ΔV_T^1 distribution has been shown to approximate an exponential statistics [43] and this is the reason for the exponential behavior of the tails in the ΔV_T distribution shown in Fig. 10(a) [42]. The slope λ (unit: [mV/dec]) of these tails can be considered as the most representative design-dependent parameter for the RTN magnitude in NAND Flash arrays. From TCAD analyses [43], the following expression was derived for it:

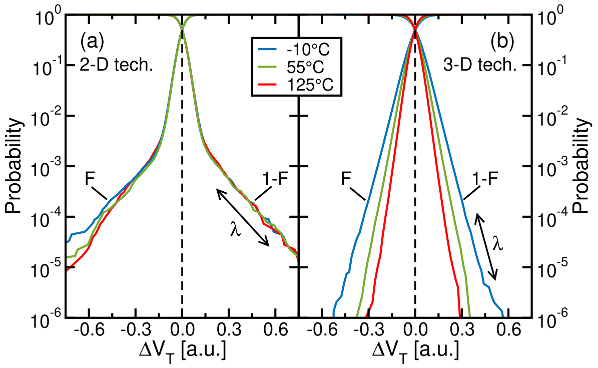

Fig. 10. Measured RTN ΔV_T distribution of a (a) 16-nm 2-D and (b) 3-D NAND Flash array, at different read temperatures. F stands for cumulative probability and $1 - F$ for its complementary. The two voltage axes have been normalized by the same arbitrary constant. Reprinted from [61].

$$\text{RTN} : \lambda = k t_{\text{ox}} \sqrt{N_a} / (\alpha_G W \sqrt{L}) \tag{3}$$

where k is a proportionality constant, t_{ox} is the tunnel-oxide thickness, N_a is the channel doping concentration, and α_G is the control-gate-to-floating-gate capacitive coupling ratio. Although different exponents for W and L have been reported [62], especially when changing the current level for V_T extraction [63], the previous formula clearly highlights that the miniaturization of cell dimensions in 2-D arrays had as a direct consequence the increase of λ. This increase meant, in turn, a stronger contribution from RTN to the widening of the cell V_T distribution and, therefore, to the worsening of the raw array reliability.

Similar to the case of program noise, the transition from 2-D to 3-D arrays strongly relieved the RTN issues. To prove that, Fig. 10(b) shows the ΔV_T distribution for a 3-D array. This distribution appears much narrower than that shown

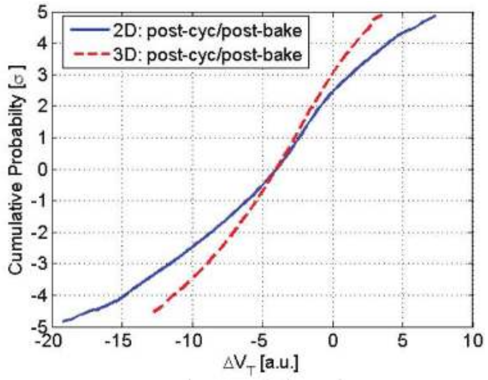

Fig. 11. Measured ΔV_T distribution due to charge detrapping in a decananometer 2-D and in a 3-D NAND Flash array. ΔV_T was evaluated as the cell V_T shift between a read operation performed before and a read operation performed after a high-temperature bake phase, with the latter following a program/erase cycling phase on the array. To achieve the same $\langle \Delta V_T \rangle$, a higher number of program/erase cycles had to be performed on the array in the 3-D case. Reprinted from [22].

in Fig. 10(a) for a decananometer 2-D array, revealing a much more stable cell V_T in the 3-D case. To be more quantitative, a reduction of λ by a factor from 5 to 6 can be extracted for the 3-D with respect to the 2-D array addressed in the figure. Although a formula equivalent to (3) has not been derived yet for 3-D cells, this reduction of λ can be attributed mainly to the large W and L of the GAA transistors in 3-D arrays, reducing the strength of their percolative channel conduction. In this regard, however, it is worth pointing out that the transition to 3-D arrays not only resulted in larger cells and in a change of cell geometry (from planar to hollow cylindrical) but also in different sources of percolation in cell channel. As will be better discussed in the next section, in fact, the process flow for vertical-channel array manufacturing at the present time allows to achieve only a polysilicon cell channel. As a consequence, not only atomistic doping but also the position of grain boundaries and trapping therein significantly contribute to nonuniformities in channel inversion and transport. Since the RTN magnitude is largely affected by these nonuniformities, some relevant changes in the RTN features have already been highlighted. One of them is the strengthening of RTN instabilities when temperature is reduced, clearly evident from the results at different temperatures in Fig. 10(b). This strengthening, which was not observed on planar cells [Fig. 10(a)], has been attributed to more relevant constraints set by polysilicon grain boundaries on channel conduction at lower temperatures, making conduction more percolative [64]. In addition, a much weaker dependence of RTN on the number of program/erase cycles performed on the cells has been reported [65].

3) Charge Detrapping: Charge detrapping is a transient process resulting in the neutralization and healing of some charged tunnel-oxide defects [48], [49]. Typically, defects involved in the phenomenon are negatively charged and their neutralization gives rise to a V_T reduction over the logarithmic timescale during data retention. Due to variability in the number of defects per cell, in the probability for their neutralization within a certain stretch of time and in their impact on cell V_T, the ΔV_T transient experienced by each cell is affected by a relevant statistical dispersion. As discussed in relation to Fig. 8, this statistical dispersion broadens the array V_T distribution as its average value decreases. In this regard, note that a Poissonian statistics is typically assumed for the number of defects per cell involved in the process [49], leading to a simple relation between the variance ($\sigma^2_{\Delta V_T}$) and the average value ($\langle \Delta V_T \rangle$) of cell ΔV_T [46]

$$\text{Charge Detr.}: \sigma^2_{\Delta V_T} = -\langle \Delta V_T \rangle \cdot \left(\langle \Delta V_T^1 \rangle + \frac{\sigma^2_{\Delta V_T^1}}{\langle \Delta V_T^1 \rangle} \right) \quad (4)$$

with $\langle \Delta V_T^1 \rangle$ representing the average V_T shift resulting from the neutralization of one single defect and $\sigma^2_{\Delta V_T^1}$ being the variance of this shift.

From the standpoint of array reliability, both $\langle \Delta V_T \rangle$ and $\sigma^2_{\Delta V_T}$ are important parameters to come to the raw bit error rate caused by charge detrapping for some data retention specifications (e.g., time and temperature). The fact that (4) sets a proportionality between these two parameters allows to focus the attention just on one of them. In the choice, $\sigma^2_{\Delta V_T}$ can be considered as more complete than $\langle \Delta V_T \rangle$, providing a better assessment of the magnitude of charge detrapping in NAND Flash arrays. $\langle \Delta V_T \rangle$, in fact, is the product between the average number of defects neutralized in a certain time slot ($\langle n_d \rangle$) and $\langle \Delta V_T^1 \rangle$ [46]. Consequently, (4) means that $\sigma^2_{\Delta V_T}$ depends not only on $\langle n_d \rangle$ and $\langle \Delta V_T^1 \rangle$ but also on the ΔV_T^1 statistics through $\sigma^2_{\Delta V_T^1}$. This allows $\sigma^2_{\Delta V_T}$ to include more information on the detrapping process. It is worth mentioning, however, that typically the ΔV_T^1 distribution needed to reproduce the detrapping phenomenology is quite close to an exponential statistics [49], in close analogy with RTN. This can be explained by considering that, both in the case of charge detrapping and RTN, the V_T shift arising from a tunnel-oxide defect is affected by the same variability, coming from the localized nature of the defect over the channel area and from the percolative source-to-drain conduction of the cells. By assuming a purely exponential distribution for ΔV_T^1, then the term inside the parentheses in (4) becomes just $2 \cdot \langle \Delta V_T^1 \rangle$. Although this means that only two parameters, namely, $\langle n_d \rangle$ and $\langle \Delta V_T^1 \rangle$ are involved in the expressions for $\langle \Delta V_T \rangle$ and $\sigma^2_{\Delta V_T}$, the latter can still be considered to provide a better assessment of the magnitude of charge detrapping than the former. This is due to the fact that $\sigma^2_{\Delta V_T}$ has a stronger dependence than $\langle \Delta V_T \rangle$ on $\langle \Delta V_T^1 \rangle$, which is the parameter depending on cell design. To better understand this point, it is worth considering that $\langle n_d \rangle$ depends on process quality, on cell immunity to the electrical stress created by the program/erase cycles and on the number of program/erase cycles that cells underwent, so it cannot be considered as specifically related to cell design as $\langle \Delta V_T^1 \rangle$ is. $\langle \Delta V_T \rangle$ and $\sigma^2_{\Delta V_T}$ can be modified through $\langle n_d \rangle$ and its previously discussed dependences, but for a given $\langle \Delta V_T \rangle$, the intrinsic magnitude of the charge detrapping process appears in terms of a larger or smaller $\sigma^2_{\Delta V_T}$ through its unique dependence on $\langle \Delta V_T^1 \rangle$.

Similar to λ in the case of RTN, $\langle \Delta V_T^1 \rangle$ largely grew with the reduction of the feature size of 2-D technologies, mainly due to the reduction of cell W and L. Consequently, for a

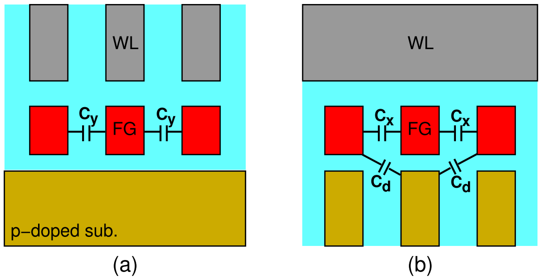

Fig. 12. Schematic for the major capacitive couplings giving rise to cell-to-cell electrostatic interference in 2-D arrays. Reference is made to an array with planar WLs and thin floating-gates [3], whose cross section is shown along (a) one of the string channels and (b) one of the array WLs. C_x and C_y are floating-gate-to-floating-gate capacitances, while C_d accounts for the direct coupling between the floating gate and the channel area of adjacent cells.

given $\langle \Delta V_T \rangle$, $\sigma^2_{\Delta V_T}$ increased and, with that, the impact of charge detrapping on the raw array reliability. With this in mind, it is easy to understand why the 2-D to 3-D transition represented a strong relief also from the standpoint of this phenomenon. Fig. 11 highlights, in fact, that, for the same $\langle \Delta V_T \rangle$, a relevant narrowing of the ΔV_T distribution coming from charge detrapping was achieved when moving from decananometer 2-D arrays to 3-D arrays. As for program noise and RTN, this was the direct consequence of the increase in cell dimensions, which allowed to reduce $\langle \Delta V_T^1 \rangle$ and, in turn, all the reliability issues coming from charge detrapping. In this regard, it is also worth mentioning that a larger number of program/erase cycles was needed for the 3-D array to achieve the same $\langle \Delta V_T \rangle$ of the 2-D array in Fig. 11, meaning that the former array is more robust than the latter against program/erase aging.

4) Cell-to-Cell Electrostatic Interference: While program noise, RTN, and charge detrapping depend mainly on the parameters of the memory transistors, cell-to-cell electrostatic interference strongly depends also on array design. Cell-to-cell electrostatic interference, in fact, arises from the impact of the charge stored in the gate-stack of the cells that are first neighbors to a victim cell on the electrostatic and conduction environment of the latter. Due to this impact, a parasitic change of V_T appears from a read operation on the victim cell when the amount of charge in its first neighbors is modified (first neighbors of the victim cell will be referred to as *aggressor* cells in the following). In 2-D arrays, this effect can be traced back mainly to the capacitive coupling between the floating gate of the aggressor cells and the floating gate or the channel area of the victim cell, as shown in Fig. 12. This makes the phenomenon dependent on some relevant cell and array parameters, such as: 1) the cell floating-gate height [3]; 2) the floating-gate distance along the bit-line (BL) direction; (corresponding to the array WL half-pitch) and along the WL direction (corresponding to the BL half-pitch) [34]; 3) the depth of the WL penetration between the floating-gates of adjacent cells in the WL direction [35]; and 4) the dielectric constant of the material surrounding the floating-gates [5], [67].

The magnitude of cell-to-cell electrostatic interference can be assessed by referring to the shift of the victim cell V_T (ΔV_T^{vic}) resulting from a change in the aggressor cell V_T (ΔV_T^{agg}). Taking the pure floating-gate-to-floating-gate capacitive coupling for reference (similar conclusions can be drawn referring to the floating-gate-to-channel contribution), the two shifts can be easily related by the following formula [35]:

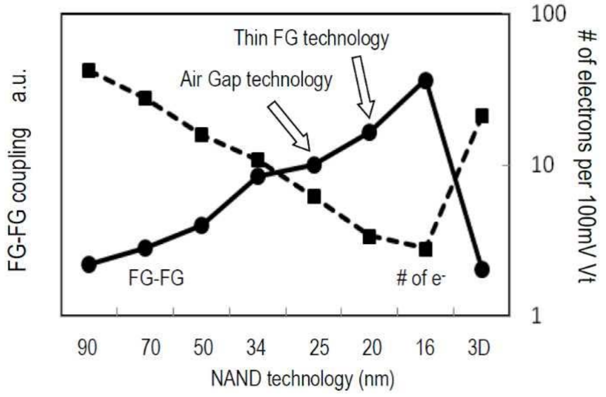

Fig. 13. Evolution of the magnitude of cell-to-cell electrostatic interference with the feature size of 2-D NAND Flash technologies (solid line). The data point for a 3-D array is also shown. Reprinted from [66].

$$\text{C2C Elec. Interf.}: \Delta V_T^{\text{vic}} \simeq \frac{C_{\text{FG}}^{\text{par}}}{C_{\text{FG}}^{\text{tot}}} \cdot \Delta V_T^{\text{agg}} \qquad (5)$$

where $C_{\text{FG}}^{\text{par}}$ is the parasitic capacitance between cell floating-gates either in the WL or in the BL direction and $C_{\text{FG}}^{\text{tot}}$ is the total floating-gate capacitance of each cell. The previous expression highlights that the ratio $C_{\text{FG}}^{\text{par}}/C_{\text{FG}}^{\text{tot}}$ represents the term summarizing how strong the parasitic electrostatic interaction between the victim and the aggressor cell is. Due to the miniaturization of the planar cell and array dimensions (cell W and L, WL and BL pitch) with a rather limited decrease of the vertical cell dimensions (floating-gate height, tunnel-oxide and interpoly dielectric thickness), the previous ratio has constantly increased with the reduction of the feature size of 2-D technologies, leading to the increase of the magnitude of cell-to-cell electrostatic interference shown in Fig. 13. In the figure, the exploitation of two process solutions which allowed to curb the phenomenon with technology scaling is highlighted. These solutions are the introduction of air gaps in-between adjacent floating-gates [5], [67] and the adoption of thin floating-gate cells [3].

In spite of all efforts to limit electrostatic interference, it became probably the most relevant reliability issue for the last 2-D NAND technologies and a strong relief to it was offered only by the transition to 3-D arrays. In this regard, a reduction of cell-to-cell interference by about the 80% is typically ascribed to this transition [14], [21], as shown in Fig. 13. This reduction was the outcome of three main positive features of 3-D arrays. The first is the GAA structure of the memory cells (see Fig. 4). The full overlap of the WL planes over cell channel allows to screen the channel and the storage layer of the victim cell from the charge stored by the aggressor cells on the same WL plane. Consequently, electrostatic interference is limited only to adjacent cells along the same vertical string, as shown in Fig. 14. The second positive feature is the larger WL pitch of 3-D arrays (50–60 nm) with respect to the last 2-D arrays (~30 nm). This corresponds to

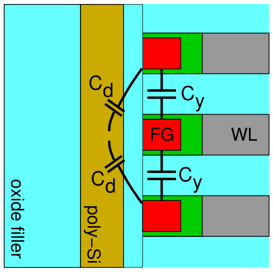

Fig. 14. Schematic for the major capacitive couplings giving rise to cell-to-cell electrostatic interference in 3-D arrays. Reference is made to a floating-gate-based technology.

an increased separation between the regions where a charge is stored along the same string, reducing their electrostatic interaction. Finally, solutions adopting a charge-trap layer for charge storage allowed to avoid a direct capacitive coupling between the storage regions of adjacent cells, making cell-to-cell interference more a channel-related phenomenon.

B. Outcomes of the Reliability Improvements Allowed by the 3-D Transition

From the previous discussions, it should be clear that the 3-D transition allowed to largely mitigate all the major physical issues constraining the raw reliability of 2-D NAND Flash arrays. Thanks to that, some abrupt improvements were achieved by moving to 3-D arrays. First of all, the strong reduction of cell-to-cell electrostatic interference allowed to simplify and speed up the programming schemes adopted by TLC technologies. In this regard, direct single-round programming of the memory cells to the 8 V_T states needed for TLC storage replaced more complex and time-consuming multiple-round programming schemes [23], [68]. Then, the reduction of cell-to-cell electrostatic interference along with the reduction of program noise, RTN, and charge detrapping allowed to achieve and keep narrower cell V_T distributions. This paved the way to the exploitation of QLC storage, further enhancing the chip GBSD (in this latter case, a two-round programming algorithm is typically adopted [23]). Finally, the stronger immunity of RTN and charge detrapping to cell aging contributed to increasing by up to a factor 10 the endurance of 3-D arrays with respect to their last 2-D predecessors. In this regard, note that strong improvements in array endurance have been reported both in the case of 3-D technologies allowing the connection of the string channel to the p substrate [21] and in the case of technologies without such connection [69]. Although, in either case, the program and erase operations are performed via Fowler–Nordheim tunneling over the channel area of the memory cells, in the former case the increase of the string potential needed for cell erase is directly achieved by biasing the p substrate contact, while in the latter it is obtained by exploiting hole generation by band-to-band tunneling at the source (and BL) junctions of the channel, with a consequent accumulation of holes in the more central regions of the string [70].

IV. Reliability Issues Specific to 3-D Arrays

Although the benefits of the 2-D-to-3-D transition are unquestionable, the changes in the process flow and in the cell and array design determined by this transition led to some new physical phenomena affecting array reliability. Among them, the most relevant are likely those coming from the polycrystalline nature of the silicon channel in the vertical NAND strings. In fact, even though possible process flows resulting in monocrystalline silicon have been recently proposed [73], at the present time all the integration schemes for 3-D arrays adopted by major semiconductor manufacturers give rise only to a polysilicon channel for the strings. Many drawbacks arise from that and set future challenges for the technology, mainly related to the presence and the haphazardness in the configuration of the polysilicon grain boundaries. Due to their high defect density, in fact, grain boundaries create energy barriers which limit the string current during the read operations, with a relative impact with respect to the drift-diffusion inside the grains depending on the read current level, temperature, and average grain size [74]. This, in turn, means that grain boundaries represent a relevant source of randomness for current transport, contributing to the variability not only of cell V_T but also of its temperature sensitivity [74]. In addition, as discussed in the previous section, the role of grain boundaries on current transport impacts RTN and introduces its nonnegligible temperature dependence in 3-D arrays. Finally, charge capture/release at the grain boundaries was shown to introduce history dependent instabilities in the BL current sensed during the read operations and, therefore, in cell V_T [71], [75]. These instabilities share the same origin of the typical overshoot and undershoot effects in the current of polysilicon thin-film transistors [76], [77], which is the bias-dependent change of the average occupancy of the trap states at the grain boundaries over extended timescales. In the case of 3-D Flash arrays, these instabilities may show up under different forms. For instance, Fig. 15 shows that the BL current resulting from some read operations performed at increasing time delays from a program pulse tends to increase [71]. This can be explained by considering that large trapping of electrons at the polysilicon grain boundaries occurs during the program pulse and that some of these electrons are then released slowly during the next data retention phase. This gives rise to the increase in the free electron concentration in the conduction band of the polysilicon channel when read operations are performed, with a consequent increase in the sensed BL current over time. In a similar way, instabilities in the BL current are expected whenever the average occupancy of the defects at the polysilicon grain boundaries changes over time or as a result of the change of the string potential induced by the voltages applied to the WLs. In this regard, it is worth mentioning that the change of the string potential in 3-D arrays may be affected by transient phenomena triggered by the fronts of the WL pulses. Fig. 16 shows, for instance, that a large decrease of the string potential may appear after the falling edge of a positive pulse applied to the WLs due to channel cutoff from the contacted p substrate (or, in some integration schemes, from the complete lack of channel connection to the p substrate) [72]. The decrease in the string potential is a function of the V_T state of the programmed cells in the string and is recovered over relatively long timescales.

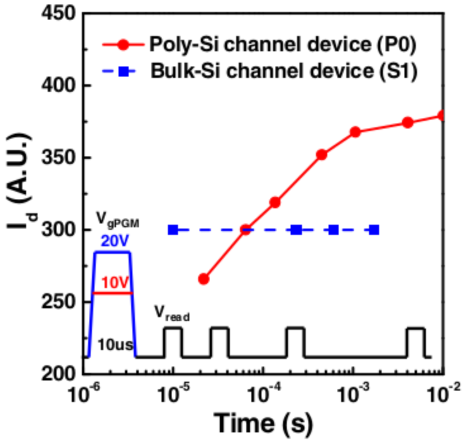

Fig. 15. Drain current of a polysilicon channel 3-D cell (red curve) and of a monocrystalline silicon channel planar cell (blue curve) resulting from a sequence of read operations at increasing time delays from a program pulse. Reprinted from [71].

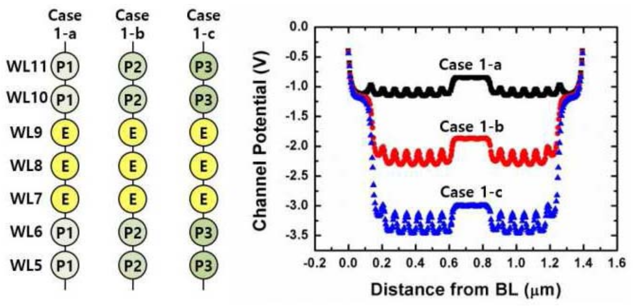

Fig. 16. Simulation results for the electrostatic potential along the channel of a 3-D NAND string after the falling edge of a positive pulse applied to the WLs, for different V_T states of the programmed cells in the string. Reprinted from [72].

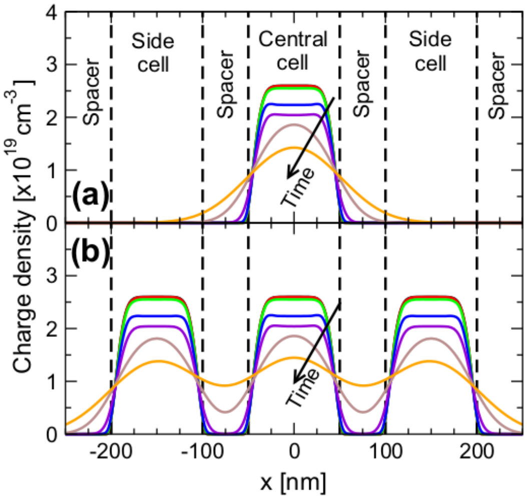

Fig. 17. Simulation results for the lateral migration of electrons along the charge-trap layer of a 3-D NAND string during data retention at 85 °C. (a) Electrons were initially stored under the gate of the central cell only, with concentration calibrated to increase cell V_T by 3 V from the neutral value. (b) Electrons were initially stored under the gate of all the three cells to increase their V_T by 3 V from the neutral value. No holes were assumed in the storage layer and no charge was assumed in the spacing regions at the beginning of data retention. Reprinted from [78], with permission from Elsevier (©2012, Elsevier).

A second relevant reliability issue specific to 3-D arrays is the lateral migration of charge along the storage layer of the memory cells. Although integration schemes relying on floating-gate storage [14] are immune to this effect, solutions based on charge storage in a continuous charge-trap layer running all along the string length [Fig. 4(b)] see in it an additional data retention constraint [79]. In this latter case, in fact, cell operation relies on the possibility to 1) store charge during program and erase just over a length corresponding to the gate (i.e., WL) region of each memory cell and 2) keep it localized over that length, thanks to the discrete and localized nature of the traps in the storage layer. Actually, both the previous conditions are typically met only marginally. In particular, during data retention charge may migrate along the storage layer, giving rise to V_T instabilities for the memory cells. These instabilities depend on the V_T state of the memory cells along the string, on the possible mismatch of the electron and hole concentration profiles resulting from the program and erase operations and on the charge density present due to string history in the spacing regions between the WL planes [79]. Fig. 17 shows some simulation results for the evolution of the electron concentration along the storage layer of a charge-trap-based 3-D NAND string during data retention [78]. In Fig. 17(a), electrons were initially stored under the gate of the central cell only, with a concentration calibrated to achieve a V_T shift from the neutral value (ΔV_T) equal to 3 V. In Fig. 17(b), instead, electrons were initially stored under the gate of all the three adjacent cells to let each of them achieve $\Delta V_T = 3$ V. No holes were assumed in the storage layer. Even though at the beginning of data retention no charge is present in the spacing regions, as data retention time elapses electrons migrate toward these regions in a sort of diffusion process. Focusing on the central cell, this leads to a decrease in both the amount of electrons stored under its gate and its ΔV_T. The resulting ΔV_T transient of the central cell is reported in Fig. 18 as a function of retention time, in the case with the two side cells initially in the neutral state. Due to the possibility for electrons to reach the channel region of the side cells [see the orange curve in Fig. 17(a)], the phenomenon may also give rise in this case to an increase of the V_T of these cells at long times [78]. It is worth pointing out that, in the case with all of the three adjacent cells initially programmed to the same ΔV_T, a lower V_T loss is expected at long times for the central cell with respect to what reported in Fig. 18 [78], [79]. When all of the three cells are initially programmed, in fact, lateral migration of electrons toward the spacing regions proceeds from both their edges and, therefore, a higher and more uniform electron concentration results over these regions with respect to the case of neutral side cells (compare (a) and (b) of Fig. 17). As a result, the lateral migration of electrons from the central cell is mitigated at long times and so is its V_T loss [78], [79]. Finally, note that, as shown in Fig. 18, the reduction of the channel length L of the memory cells makes the phenomenon more and more relevant since the charge lost over the spacing regions represents a higher fraction of the total charge initially stored under cell gate when L is shorter [80]. All of these effects

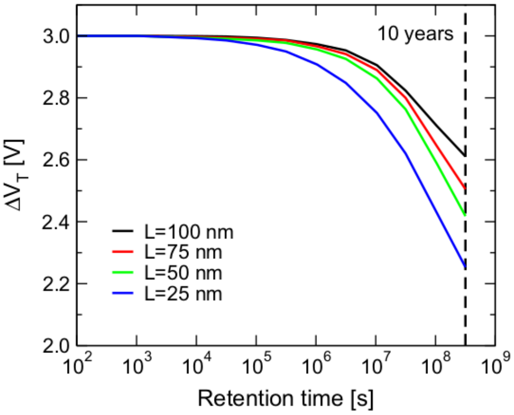

Fig. 18. Simulation results for the evolution of ΔV_T, i.e., the V_T shift from the neutral value, of the central cell in Fig. 17 during data retention at 55 °C. An initial $\Delta V_T = 3$ V was assumed for the central cell, while its adjacent cells were assumed in the neutral state. Results for different gate lengths L of the memory cells are reported. Reprinted from [78], with permission from Elsevier (©2012, Elsevier).

will have to be carefully considered in the future evolution of 3-D NAND Flash technologies.

To complete the discussion, it is worth mentioning that lateral charge migration along the charge-trap storage layer has also been invoked to explain a fast V_T loss observed after cell programming in 3-D arrays [81]. However, additional sources for this fast V_T loss may be electron emission from shallow traps in the charge-trap layer and in the dielectrics inside the gate-stack of the memory cells, as previously reported on charge-trap-based planar technologies [82]. Finally, it is important to recall that, although the same disturbs affecting 2-D arrays were inherited by 3-D arrays, some additional issues from the standpoint of read and program disturbs appeared in the latter case, which have to be faced by careful optimizations in the technology and in the array working conditions [83]–[86].

V. Conclusion

This paper presented an overview of what the 2-D-to-3-D transition meant for the reliability of NAND Flash arrays. After a quick glance at its fundamentals, the raw array reliability was discussed focusing on the impact that the evolution of 2-D technologies had on it. The magnitude of the most relevant physical phenomena constraining array reliability was then investigated as a function of cell and array parameters, highlighting the reasons why the 3-D transition came as a relief to all of them. Finally, some new physical issues specific of 3-D arrays have been discussed, pointing out their constraints to the reliability of future technology nodes.

Acknowledgment

The authors would like to thank A. Goda, E. Camerlenghi, P. Tessariol, and L. Chiavarone from Micron Technology Inc., and A. L. Lacaita from the Politecnico di Milano for discussions and support. The authors would also like to thank all the current and former Ph.D. students from the Politecnico di Milano who contributed with them to the investigation of NAND Flash reliability over the years.

References

[1] R. E. Fontana, Jr., and G. M. Decad, “Moore’s law realities for recording systems and memory storage components: HDD, tape, NAND, and optical,” *AIP Adv.*, vol. 8, no. 5, pp. 056506-1–056506-5, 2018. doi: 10.1063/1.5007621.

[2] C. Monzio Compagnoni, A. Goda, A. S. Spinelli, P. Feeley, A. L. Lacaita, and A. Visconti, “Reviewing the evolution of the NAND Flash technology,” *Proc. IEEE*, vol. 105, no. 9, pp. 1609–1633, Sep. 2017. doi: 10.1109/JPROC.2017.2665781.

[3] G. Naso *et al.*, “A 128 Gb 3b/cell NAND Flash design using 20 nm planar-cell technology,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2013, pp. 218–219. doi: 10.1109/ISSCC.2013.6487707.

[4] S. Lee *et al.*, “A 128 Gb 2b/cell NAND Flash memory in 14 nm technology with t_{PROG}=640 μs and 800 MB/s I/O rate,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Jan./Feb. 2016, pp. 138–139. doi: 10.1109/ISSCC.2016.7417945.

[5] K. Prall and K. Parat, “25 nm 64 Gb MLC NAND technology and scaling challenges invited paper,” in *IEDM Tech. Dig.*, 2010, pp. 102–105. doi: 10.1109/IEDM.2010.5703300.

[6] 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. doi: 10.1109/IEDM.2011.6131518.

[7] N. R. Mielke, R. E. Frickey, I. Kalastirsky, M. Quan, D. Ustinov, and V. J. Vasudevan, “Reliability of solid-state drives based on NAND Flash memory,” *Proc. IEEE*, vol. 105, no. 1, pp. 1725–1750, Sep. 2017. doi: 10.1109/JPROC.2017.2725738.

[8] 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. doi: 10.1016/j.sse.2013.01.006.

[9] 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. doi: 10.1109/IRPS.2014.6860599.

[10] J. Lee, J. Jang, J. Lim, Y. G. Shin, K. Lee, and E. Jung, “A new ruler on the storage market: 3D-NAND Flash for high-density memory and its technology evolutions and challenges on the future,” in *IEDM Tech. Dig.*, 2016, pp. 284–287. doi: 10.1109/IEDM.2016.7838394.

[11] 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. doi: 10.1109/VLSIT.2007.4339708.

[12] J. Jang *et al.*, “Vertical cell array using TCAT (Terabit Cell Array Transistor) technology for ultra high density NAND Flash memory,” in *VLSI Symp. Tech. Dig.*, 2009, pp. 192–193.

[13] 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.*, 2012, pp. 211–214. doi: 10.1109/IEDM.2012.6479011.

[14] K. Parat and C. Dennison, “A floating gate based 3D NAND technology with CMOS under array,” in *IEDM Tech. Dig.*, 2015, pp. 48–51. doi: 10.1109/IEDM.2015.7409618.

[15] S. Lee *et al.*, “A 1Tb 4b/cell 64-stacked-WL 3D NAND Flash memory with 12 MB/s program throughput,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2018, pp. 340–341. doi: 10.1109/ISSCC.2018.8310323.

[16] H. Maejima *et al.*, “A 512 Gb 3b/cell 3D Flash memory on a 96-word-line-layer technology,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 2018, pp. 336–337. doi: 10.1109/ISSCC.2018.8310321.

[17] N. Shibata *et al.*, “A 1.33 Tb 4-bit/cell 3D-Flash memory on a 96-word-line-layer technology,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 210–211.

[18] D. Kang *et al.*, “A 512 Gb 3-bit/cell 3D 6^{th}-generation V-NAND Flash memory with 82 MB/s write throughput and 1.2 Gb/s interface,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 2019, pp. 216–217.

[19] C. Siau *et al.*, “A 512 Gb 3-bit/cell 3D Flash memory on 128-wordline-layer with 132 MB/s write performance featuring circuit-under-array technology,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 218–219.

[20] B. T. Park *et al.*, “ 32 nm 3-bit 32 Gb NAND Flash memory with DPT (double patterning technology) process for mass production,” in *Symp. VLSI Tech. Dig.*, 2010, pp. 125–126. doi: 10.1109/VLSIT.2010.5556196.

[21] H. Kim, S.-J. Ahn, Y. G. Shin, K. Lee, and E. Jung, “Evolution of NAND Flash memory: From 2D to 3D as a storage market leader,” in *Proc. IMW*, 2017, pp. 1–4. doi: 10.1109/IMW.2017.7939081.

[22] 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.*, 2015, pp. 374–377. doi: 10.1109/IEDM.2015.7409699.

[23] K. Parat and A. Goda, "Scaling trends in NAND Flash," in *IEDM Tech. Dig.*, 2018, pp. 27–30. doi: 10.1109/IEDM.2018.8614694.

[24] C. Monzio Compagnoni *et al.*, "First evidence for injection statistics accuracy limitations in NAND Flash constant-current Fowler-Nordheim programming," in *IEDM Tech. Dig.*, 2007, pp. 165–168. doi: 10.1109/IEDM.2007.4418892.

[25] 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. doi: 10.1109/TED.2008.2003332.

[26] A. Chimenton, P. Pellati, and P. Olivo, "Erratic bits in Flash memories under Fowler-Nordheim programming," *Jpn. J. Appl. Phys.*, vol. 42, pp. 2041–2043, Apr. 2003. doi: 10.1143/JJAP.42.2041.

[27] K. Seidel *et al.*, "Analysis of trap mechanisms responsible for Random Telegraph Noise and erratic programming on sub-50 nm floating gate Flash memories," in *Proc. NVMTS*, 2009, pp. 67–71. doi: 10.1109/NVMT.2009.5429788.

[28] A. Spessot, C. Monzio Compagnoni, F. Farina, A. Calderoni, A. S. Spinelli, and P. Fantini, "Effect of floating-gate polysilicon depletion on the erase efficiency of NAND Flash memories," *IEEE Electron Device Lett.*, vol. 31, no. 7, pp. 647–649, Jul. 2010. doi: 10.1109/LED.2010.2048194.

[29] R. Shirota *et al.*, "Analysis of the correlation between the programmed threshold-voltage distribution spread of NAND Flash memory devices and floating-gate impurity concentration," *IEEE Trans. Electron Devices*, vol. 58, no. 11, pp. 3712–3719, Nov. 2011. doi: 10.1109/TED.2011.2165073.

[30] N. Mielke *et al.*, "Bit error rate in NAND Flash memories," in *Proc. IRPS*, 2008, pp. 9–19. doi: 10.1109/RELPHY.2008.4558857.

[31] Y. Cai, Y. Luo, S. Ghose, and O. Mutlu, "Read disturb errors in MLC NAND Flash memory: Characterization, mitigation, and recovery," in *Proc. 45th IEEE/IFIP Int. Conf. Dependable Syst. Netw.*, Jun. 2015, pp. 438–449. doi: 10.1109/DSN.2015.49.

[32] K.-D. Suh *et al.*, "A 3.3 V 32 Mb NAND Flash memory with incremental step pulse programming scheme," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 1995, pp. 128–129.

[33] 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*, 2006, pp. 31–33. doi: 10.1109/.2006.1629481.

[34] 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. doi: 10.1109/55.998871.

[35] 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. doi: 10.1109/LED.2008.2009555.

[36] R. Moazzami and C. Hu, "Stress-induced current in thin silicon dioxide films," in *IEDM Tech. Dig.*, 1992, pp. 139–142. doi: 10.1109/IEDM.1992.307327.

[37] F. Arai, T. Maruyama, and R. Shirota, "Extended data retention process technology for highly reliable Flash EEPROMs of 10^6 to 10^7 W/E cycles," in *Proc. IRPS*, 1998, pp. 378–382. doi: 10.1109/RELPHY.1998.670672.

[38] D. Ielmini, A. S. Spinelli, A. L. Lacaita, and A. Modelli, "A statistical model for SILC in Flash memories," *IEEE Trans. Electron Devices*, vol. 49, no. 11, pp. 1955–1961, Nov. 2002. doi: 10.1109/TED.2002.804730.

[39] H. P. Belgal, N. Righos, I. Kalastirsky, J. J. Peterson, R. Shiner, and N. Mielke, "A new reliability model for post-cycling charge retention of Flash memories," in *Proc. IRPS*, 2002, pp. 7–20. doi: 10.1109/RELPHY.2002.996604.

[40] H. Kurata *et al.*, "The impact of random telegraph signals on the scaling of multilevel Flash memories," in *Symp. VLSI Circ. Dig.*, 2006, pp. 112–113. doi: 10.1109/VLSIC.2006.1705335.

[41] R. Gusmeroli *et al.*, "Defects spectroscopy in SiO_2 by statistical random telegraph noise analysis," in *IEDM Tech. Dig.*, 2006, pp. 483–486. doi: 10.1109/IEDM.2006.346819.

[42] 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. doi: 10.1109/TED.2007.910605.

[43] 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. doi: 10.1109/TED.2009.2024031.

[44] N. Mielke *et al.*, "Flash EEPROM threshold instabilities due to charge trapping during program/erase cycling," *IEEE Trans. Device Mater. Rel.*, vol. 4, no. 3, pp. 335–344, Sep. 2004. doi: 10.1109/TDMR.2004.836721.

[45] 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. doi: 10.1109/RELPHY.2006.251188.

[46] 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. doi: 10.1109/TED.2014.2327661.

[47] 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. doi: 10.1109/TED.2014.2327149.

[48] 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. doi: 10.1109/TED.2016.2617888.

[49] 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. doi: 10.1109/TED.2016.2617890.

[50] 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. Rel.*, vol. 4, no. 1, pp. 110–117, Mar. 2004. doi: 10.1109/TDMR.2004.824360.

[51] 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. doi: 10.1109/TED.2011.2150751.

[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. doi: 10.1109/TED.2008.2003230.

[53] 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. doi: 10.1109/LED.2009.2026658.

[54] A. Spessot *et al.*, "Variability effects on the V_T distribution of nanoscale NAND Flash memories," in *Proc. IRPS*, 2010, pp. 970–974. doi: 10.1109/IRPS.2010.5488695.

[55] D. Ielmini, A. S. Spinelli, A. L. Lacaita, L. Confalonieri, and A. Visconti, "New technique for fast characterization of SILC distribution in Flash arrays," in *Proc. IRPS*, 2001, pp. 73–80. doi: 10.1109/RELPHY.2001.922885.

[56] P. Cappelletti, R. Bez, A. Modelli, and A. Visconti, "What we have learned on Flash memory reliability in the last ten years," in *IEDM Tech. Dig.*, 2004, pp. 489–492. doi: 10.1109/IEDM.2004.1419196.

[57] 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, no. 4S, pp. 2598–2601, 2008. doi: 10.1143/JJAP.47.2598.

[58] D. James, "Recent advances in memory technology," in *Proc. ASMC*, 2013, pp. 386–395. doi: 10.1109/ASMC.2013.6552766.

[59] A. Maconi, S. M. Amoroso, C. Monzio Compagnoni, A. Mauri, A. S. Spinelli, and A. L. Lacaita, "Three-dimensional simulation of charge-trap memory programming—Part II: Variability," *IEEE Trans. Electron Devices*, vol. 58, no. 7, pp. 1872–1878, Jul. 2011. doi: 10.1109/TED.2011.2138709.

[60] 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. doi: 10.1109/IRPS.2011.5784588.

[61] D. Resnati, A. Goda, G. Nicosia, C. Miccoli, A. S. Spinelli, and C. Monzio Compagnoni, "Temperature effects in NAND Flash memories: A comparison between 2-D and 3-D arrays," *IEEE Electron Device Lett.*, vol. 38, no. 4, pp. 461–464, Apr. 2017. doi: 10.1109/LED.2017.2675160.

[62] 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.*, 2007, pp. 169–172. doi: 10.1109/IEDM.2007.4418893.

[63] 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. doi: 10.1109/LED.2013.2250477.

[64] G. Nicosia *et al.*, "Characterization and modeling of temperature effects in 3-D NAND Flash arrays—Part II: Random telegraph noise," *IEEE Trans. Electron Devices*, vol. 65, no. 8, pp. 3207–3213, Aug. 2018. doi: 10.1109/TED.2018.2839904.

[65] G. Nicosia, A. Goda, A. S. Spinelli, and C. Monzio Compagnoni, "Impact of cycling on random telegraph noise in 3-D NAND Flash arrays," *IEEE Electron Device Lett.*, vol. 39, no. 8, pp. 1175–1178, Aug. 2018. doi: 10.1109/LED.2018.2847341.

[66] T. Tanaka *et al.*, "A 768 Gb 3b/cell 3D-floating-gate NAND Flash memory," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Jan./Feb. 2016, pp. 142–143. doi: 10.1109/ISSCC.2016.7417947.

[67] 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. doi: 10.1109/.2006.1629483.

[68] 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. doi: 10.1109/JSSC.2015.2474117.

[69] 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.*, 2007, pp. 449–452. doi: 10.1109/IEDM.2007.4418970.

[70] G. Malavena, A. L. Lacaita, A. S. Spinelli, and C. Monzio Compagnoni, "Investigation and compact modeling of the time dynamics of the GIDL-assisted increase of the string potential in 3-D NAND Flash arrays," *IEEE Trans. Electron Devices*, vol. 65, no. 7, pp. 2804–2811, Jul. 2018. doi: 10.1109/TED.2018.2831902.

[71] W.-J. Tsai *et al.*, "Polycrystalline-silicon channel trap induced transient read instability in a 3D NAND Flash cell string," in *IEDM Tech. Dig.*, 2016, pp. 288–291. doi: 10.1109/IEDM.2016.7838395.

[72] Y. Kim and M. Kang, "Down-coupling phenomenon of floating channel in 3D NAND Flash memory," *IEEE Electron Device Lett.*, vol. 37, no. 12, pp. 1566–1569, Dec. 2016.

[73] R. Delhougne *et al.*, "First demonstration of monocrystalline silicon macaroni channel for 3-D NAND memory devices," in *VLSI Symp. Tech. Dig.*, 2018, pp. 203–204. doi: 10.1109/VLSIT.2018.8510635.

[74] D. Resnati *et al.*, "Characterization and modeling of temperature effects in 3-D NAND Flash arrays—Part I: Polysilicon-induced variability," *IEEE Trans. Electron Devices*, vol. 65, no. 8, pp. 3199–3206, Aug. 2018. doi: 10.1109/TED.2018.2838524.

[75] H.-J. Kang *et al.*, "Effect of traps on transient bit-line current behavior in word-line stacked NAND Flash memory with poly-Si body," in *VLSI Symp. Tech. Dig.*, pp. 1–2, 2014. doi: 10.1109/VLSIT.2014.6894348.

[76] N. Bavidge, M. Boero, P. Migliorato, and T. Shoimoda, "Switch-on transient behavior in low-temperature polycrystalline silicon thin-film transistors," *Appl. Phys. Lett.*, vol. 77, no. 23, pp. 3836–3838, 2000. doi: 10.1063/1.1329867.

[77] L. Michalas, G. J. Papaioannou, D. N. Kouvatsos, and A. T. Voutsas, "Investigation of the undershoot effect in polycrystalline silicon thin film transistors," *Solid-State Electron.*, vol. 52, no. 3, pp. 394–399, 2008. doi: 10.1016/j.sse.2007.10.018.

[78] 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, doi: 10.1016/j.sse.2012.04.013.

[79] H.-J. Kang *et al.*, "Comprehensive analysis of retention characteristics in 3-D NAND Flash memory cells with tube-type poly-Si channel structure," in *VLSI Symp. Tech. Dig.*, 2015, pp. 1–2. doi: 10.1109/VLSIT.2015.7223670.

[80] A. Maconi *et al.*, "Impact of lateral charge migration on the retention performance of planar and 3D SONOS devices," in *Proc. ESSDERC*, 2011, pp. 195–198. doi: 10.1109/ESSDERC.2011.6044201.

[81] B. Choi *et al.*, "Comprehensive evaluation of early retention (fast charge loss within a few seconds) characteristics in tube-type 3-D NAND Flash memory," in *VLSI Symp. Tech. Dig.*, 2016, pp. 1–2. doi: 10.1109/VLSIT.2016.7573385.

[82] C.-P. Chen, H.-T. Lue, C.-C. Hsieh, K.-P. Chang, K.-Y. Hsieh, and C.-Y. Lu, "Study of fast initial charge loss and it's impact on the programmed states V_T distribution of charge-trapping NAND Flash," in *IEDM Tech. Dig.*, 2010, pp. 118–121. doi: 10.1109/IEDM.2010.5703304.

[83] K.-S. Shim *et al.*, "Inherent issues and challenges of program disturbance of 3D NAND Flash cell," in *Proc. IMW*, 2012, pp. 1–4. doi: 10.1109/IMW.2012.6213659.

[84] B.-I. Choe, J.-K. Lee, B.-G. Park, and J.-H. Lee, "Suppression of read disturb fail caused by boosting hot carrier injection effect for 3-D stack NAND Flash memories," *IEEE Electron Device Lett.*, vol. 35, no. 1, pp. 42–44, Jan. 2014. doi: 10.1109/LED.2013.2288991.

[85] Y. Zhang, L. Jin, D. Jiang, X. Zou, H. Liu, and Z. Huo, "A novel read scheme for read disturbance suppression in 3D NAND Flash memory," *IEEE Electron Device Lett.*, vol. 38, no. 12, pp. 1669–1672, Dec. 2017. doi: 10.1109/LED.2018.2844404.

[86] Y. Zhang, L. Jin, X. Zou, H. Liu, A. Zhang, and Z. Huo, "A novel program scheme for program disturbance optimization in 3-D NAND Flash memory," *IEEE Electron Device Lett.*, vol. 39, no. 7, pp. 959–962, Jul. 2018. doi: 10.1109/LED.2018.2844404.