1T-1C Dynamic Random Access Memory Status, Challenges, and Prospects

Alessio Spessot and Hyungrock Oh

(Invited Paper)

***Abstract*— This article reviews the status, the challenges, and the perspective of 1T-1C dynamic random access memory (DRAM) chip. The basic principles of the DRAM are presented, introducing the key functional aspects and the structure of modern devices. We present the most relevant historical trends for different modules of the memory chip, such as access device and storage element, reviewing some of the technological challenges faced by industry to guarantee the device shrinking imposed by the economic law. The most recent solutions introduced by the industry in modern DRAM devices for the critical elements are presented. Finally, a survey of the most critical bottleneck for future development is presented, reviewing some of the potential trends and perspectives of DRAM development.**

***Index Terms*— 1T-1C, access device, dynamic random access memory (DRAM) chips, DRAM, vertical transistor.**

I. Introduction

THE concept of a one-transistor dynamic random access memory (DRAM) obtained combining a simple transistor and a small capacitor was envisioned in 1966 by Dr. Robert Dennard, a Fellow at the IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA [1]. Dennard and his team were working on early field-effect transistors and integrated circuits, and his attention to memory chips came from seeing another team's research on thin-film magnetic memory. Dennard claims he went home and started working on a simplified version of the memory, and in 1968, a patent for DRAM was granted to Dennard and IBM [2].

In 1970, a newly formed company called Intel Corp, Mountain View, CA, USA, in 1968 (now headquartered in Santa Clara, CA, USA) publicly released 1103, the first DRAM chip based on a 1-kb pMOS DRAM [3]. By 1972, it was the best selling semiconductor memory chip in the world, defeating the magnetic core-type memory [4].

Nowadays, DRAM chips are widely used in electronic devices, thanks to high-speed operations, large integration density, and excellent reliability [5], [6]. In the past decades, we have assisted in an exponential growth of the number of memory cells per chip used. The major strategy to realize such growth is the perpetual memory area cell scaling [7]. Due to its intrinsic simple nature, DRAM is particularly suitable for scaling, and its excellent scalability contributed to its widespread success.

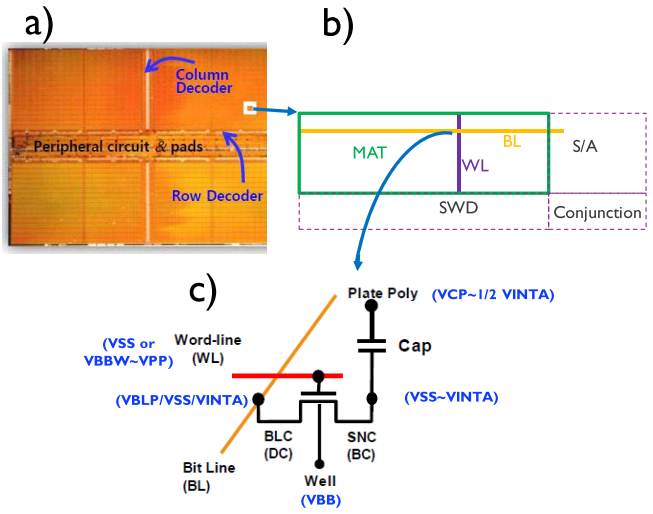


Fig. 1. (a) DRAM chip. array area, column decoder, and row decoder are visible. (b) Sketch of a memory array (MAT). WLs and BLs are connected to SWDs and sense amplifier (S/A), respectively. (c) Basic schematic of the 1T-1C circuit. Gate of the access device is connected to the WL, while the source and drain terminals are connected to BL and capacitor. The most relevant bias is shown in blue. Adapted from [5] and [9].

Today, technological efforts are focusing on enabling further cell area scaling. As the DRAM cell is reduced, innovations in many technological aspects are getting more crucial than ever since we need to overcome the limits raising with the 10-nm technology nodes and beyond [8].

An example of a modern DRAM chip is shown in Fig. 1(a), where the key functional elements of the memory are visible.

This article is organized as follows. In Section II, we review the key functional aspects of a DRAM cell. Section III reviews some of the most relevant historical trends and latest innovations, which serve as a basis to understand the key technical challenges that modern research has to tackle. In Section IV, we discuss some of such critical technical aspects, reviewing the proposed solutions mentioned in the literature and discussing their different levels of maturity.

II. DRAM Key Functional Aspects

A. DRAM Chip Architecture

A DRAM cell is conceptually a very simple structure, based on a 1 selector transistor and 1 capacitor (1T-1C), which

Manuscript received December 19, 2019; accepted December 23, 2019. Date of publication January 30, 2020; date of current version March 24, 2020. The review of this article was arranged by Editor T. Kim. *(Corresponding author: Alessio Spessot.)*

The authors are with imec, 3001 Leuven, Belgium (e-mail: alessio.spessot@imec.be; hyungrock.oh@imec.be).

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.2020.2963911

acts as a storage memory element [Fig. 1(c)]. The transistor, also called access transistor or access device, has the gate terminal connected to the word line (WL). The drain terminal is connected to the bit line (BL) by a BL contact (BLC), and the other terminal is connected to the capacitor by a storage node contact (SNC). The access device acts as a switch, and the capacitor can store the bit as a positive or negative electrical charge. The memory state can be read by sensing the stored charge on the capacitor via the BL, which is set to an operating bias when the transistor is closed. When the transistor is then switched on, the stored charge flows into the BL, generating a potential change that can be detected and amplified by a sense amplifier connected to the BL [9].

This basic structure is very simple and small, contributing to the ubiquitous diffusion of this memory device, but has some disadvantages. The charge cannot remain in the small capacitor forever, due to the leakage current from or to the access devices, making it lose its well-defined charge state over time [10]. To overcome this problem, DRAM memories are periodically refreshed, reading the content of the memory and writing it back. This is where the name "Dynamic" means in the DRAM context.

As shown in Fig. 1(a), the DRAM chip can be divided into three major parts, which are identified as cell array, core part, and peripheral part. In a modern mobile DRAM chip, the array area is composed of a memory cell array to store the data and is the more significant area contributor since its portion amounts to 50%–55% of the full-chip area [9]. The second one is the core area, which is composed of a row and column decoder, a section WL driver (SWD), a BL sense amplifier (BLSA), and the conjunction (CJT) area formed in the cross region of BLSA and SWD to generate or pass through the control signals for BLSA and data delivery on the I/O lines [Fig. 1(b)]. This section of the chip manages the read and write, decoding, and data restoring and typically occupies 25%–30% of the chip area. The last one is the peripheral part, which is formed by the control logic, I/O interface, and dc circuits that account for the remaining ∼20% of the area [9].

In general, a DRAM chip is hierarchically composed of rank, bank, and cell array also called MAT [Fig. 1(b)]. The smallest element is the DRAM cell array. Each vertical column of the cells is connected to a BLSA, and each horizontal row of the cells is connected to an SWD. In typical modern devices, cell array consists of 1024 columns × 512–1024 rows. A larger number of cells per MAT are beneficial to reduce the chip size and to enhance the cell efficiency, and therefore, larger MATs are preferred from a cost perspective. However, for a fixed MAT density, the number of cells in BL and WL determines the parasitic loadings, which strongly affects the core performance (speed and power consumption). Therefore, special attention needs to be reserved for the tradeoff between the sensing margin and performance due to the increase in loading.

B. Voltages in DRAM Cell

Fig. 2 shows the various operating voltages required in a DRAM device. Only VDD is applied to the DRAM as external inputs, while all the others are generated by internal voltage generator circuits. All these voltages are crucial for DRAM device performance, so they should be precisely generated from VDD with different trimming methods and applied to the relevant chip area. VINTA and VINT are the key dc levels, which are provided to the core area and peripheral area, respectively. VINTA, which is the core area dedicated voltage, drives the BL during read/write operations. More precisely, data transfer during the read/write operation occurs through the BL, where the swing level is driven from the ground voltage (VSS) (low level, corresponding to data "0") to VINTA (high level, corresponding to data "1"). In addition, VINTA is strongly related to the sensing margin, so higher VINTA increases the sensing margin at the expense of power increase. DRAM operation starts with the WL enablement, which is driven by the word line voltage (VPP) level. VPP is the bias applied to the gate voltage of the access transistor. This bias is ∼3.0 V [5], which is higher than the external VDD and is then positively boosted by the pump circuitry to ensure the ON-current required meeting the write-speed specifications.

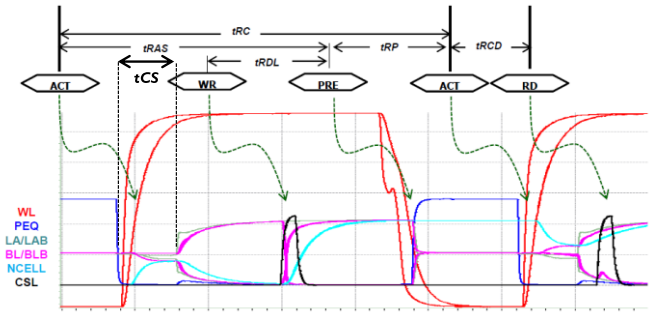
Bias type		Level @DDR3	Note
VDD	External voltage	1.5V	To operate peripheral area
VINT	Internal voltage	Same as VDD or ≤ VDD	
VPP	Word line voltage	~ 3.0V	To turn on cell transistor
VINTA	Storage node voltage	1.0 ~ 1.3V	Data '1' stored when storage node's voltage is VINTA
VCP	Capacitor plate voltage	~ ½ VINTA	To reduce electric field applied to cell capacitor
VBLP	Bit line pre-charge voltage	~ ½ VINTA	Stand-by voltage level of Bit line
VBBW	Negative Word line voltage	-0.1~-0.4V	To turn off cell transistor if VBBW is available
VBB	Cell transistor's body voltage	-0.5 ~ -0.8V	To reduce off-current of cell transistor
VSS	Ground voltage	0V	Ground level

Fig. 2. Bias levels used in the modern DRAM, with an example of values for a DDR3. Adapted from [5].

The leakage of the cell transistor is another crucial parameter since it affects the refresh time. To reduce the OFF-current of cell transistor, a negative-boosted WL voltage (called VBBW) is applied. Typically, VBBW is lower than VSS and depends on the performance of the cell transistor. Consequently, the WL swing level ranges from VBBW to VPP, which ensures the high current by VPP and lower OFF-current by VBBW. As in the case of VBBW, the cell transistor's body voltage (VBB) is applied to the bulk of the access transistor to decrease the OFF-current. By applying the VBB level, the cell transistor's threshold voltage can be increased and the leakage current is limited. As shown in the bit cell configuration, one node of the cell capacitor is connected to the drain (source) of the cell transistor, and the other node is connected to the cell plate which is a common node in the cell array.

The plate voltage is commonly called VCP (cell plate voltage), and its static level is fixed as half of the VINTA to reduce the cell capacitor stress and improve its reliability. Typical values are VINTA ∼ 1.0 V and VCP ∼ 0.5 V [5].

It is relevant to mention that the internal reference level is designed to compensate for the process–voltage–temperature (PVT) variation to guarantee the reliability of the device operations. Examples of circuital solutions to accounting for the PVT variations are reported in [11].



tRC: ACT to ACT/REF command period
tRAS: ACT to PRE command period
tRP: PRECHARGE command period
tRCD: ACT to READ/write delay
tRDL: Last data-in to PRE command period
tCS: charge-sharing time

Fig. 3. Basic DRAM cell operations. ACT, write and read, and precharge phases are visible. Signal pulses for WL, precharging (PEQ), sense amplifier (LA/LAB), BLs, and column select line are visible. Adapted from [9].

C. Basic Cell Operation

Before going into the details of the operations, there are a few crucial time parameters that define the internal core operation such as tRAS, tRP, tRC, tRCD, and tRDL, which are shown in Fig. 3.

A typical DRAM cell operation is triggered by WL enablement, which is also called activation (ACT) command. Obviously, prior to WL activation, the DRAM cell remained in the precharged state fixed by the precharging circuitry. As shown in Fig. 3, once the ACT command is introduced, the precharge circuit disables the precharging signal [equalizing pulse signal (PEQ)] for core operation, and also, the WL selected by the address decoding is enabled with the boosted voltage (VPP) that is applied to the gate of access device transistor. As mentioned before, the VPP level is the highest dc level applied to the cell transistor during the operation and should be enabled until the end of internal operation. The timing of this operation is determined by the tRAS parameter. During the WL activation, charge sharing takes place between the BL and the storage node. This shared level can be enlarged by the sense amplifier's enablement that is controlled by LA and LAB signals, which supply the power to the sense amplifier, followed by the column select line (CSL) signal (LA and LAB denote one of the SA enabling signals used to enable nMOS sense amplifier (NSA) and pMOS sense amplifier (PSA) located in the BLSA, respectively). This CSL signal triggers that data transfers between BLs and I/O lines controlled by the column address. When a read command is given, the data are transferred from the memory cell array. In the case of the write operation, the written data are delivered from the I/O lines to the memory array, which forces the BL to have a full swing from VSS (data "0") to VINTA (data "1"). The time tRDL is allowed to complete the write operation, ensuring that the full swing of BL during the write operation can take place. On the contrary, read operation should be controlled by the critical timing parameters such as tRC and tRCD, as will be detailed in the read section.

After finishing the read/write operation, a precharging operation is required to restore the BL precharge state; the minimum required time for precharging is determined by tRP. Consequently, the read cycle (tRC) can be determined with tRAS and tRP. All these parameters should be met to ensure reliable DRAM operations.

D. Write

Write operation is initiated from the WL activation process, as in the case of the read operation. Fig. 3 illustrates the biasing condition during the write operation. At this time, the boosted VPP voltage is applied to the WL to transfer the data into the cell from the BL. Such VPP level, that as was mention is the highest applied to the cell, can be defined as follows:

$$\mathrm{VPP} \geq \mathrm{VINTA} + \mathrm{Vt} + \mathrm{VINTA} \times \gamma \tag{1}$$

where VINTA is the array-dedicated voltage, and Vt and γ are the threshold voltage and body-effect coefficient of the cell transistor. Once WL is driven to VPP by the activation command, a high level (VINTA) or low level (VSS) is applied to the BL through the BLSA. After that, the storage node of the selected cell can preserve the written data through the cell transistor. As illustrated in Fig. 4(a), the write path includes the BL, the access transistor channel, and the storage node. Therefore, the total resistance that affects the write operation includes the BL resistance (R_{BL}), the BLC resistance (R_{BLC}), the access transistor channel resistance (Rch), and the SNC resistance (R_{SNC}). In particular, the channel resistance (Rch) depends on the boosted WL voltage (VPP), the gate oxide thickness, and the mobility of access transistor. On the other hand, contact dimension and plug materials influence R_{SNC} and R_{BLC}. The time period required to write data to overdrive the sense amplifiers and written through into the DRAM cells is defined as write recovery time (tWR). It is clear that all the above-mentioned resistances should be controlled to avoid that the tWR is exceeded. tWR is the maximum time allowed to complete the writing of the selected cell, and a precharge command can be released only after the correct data values have been restored to the DRAM cells.

E. Retention

We have mentioned that the DRAM bit cell is composed of 1T-1C and the SNC located between cell transistor and cell capacitor [Fig. 1(c)]. DRAM capacitor cannot retain the data permanently due to multiple leakage current paths [Fig. 4(c)], and therefore, periodical refresh is needed (typical refresh time is 64 ms according to JEDEC specifications). As shown in Fig. 4(b), retention time is widely distributed across different cells of a DRAM chip, and only a few cells are approaching the refresh limit. However, these weak cells are the real limiting factor for the refresh time. Proper screening of the weak cells and fabrication of redundant cells is needed to enable the device functionality.

The main leakage current contributors, as shown in Fig. 4(c), are as follows: 1) junction leakage from the SNC; 2) gate-induced drain leakage (GIDL) current generated by the access device; and 3) OFF-current leakage between SNC and

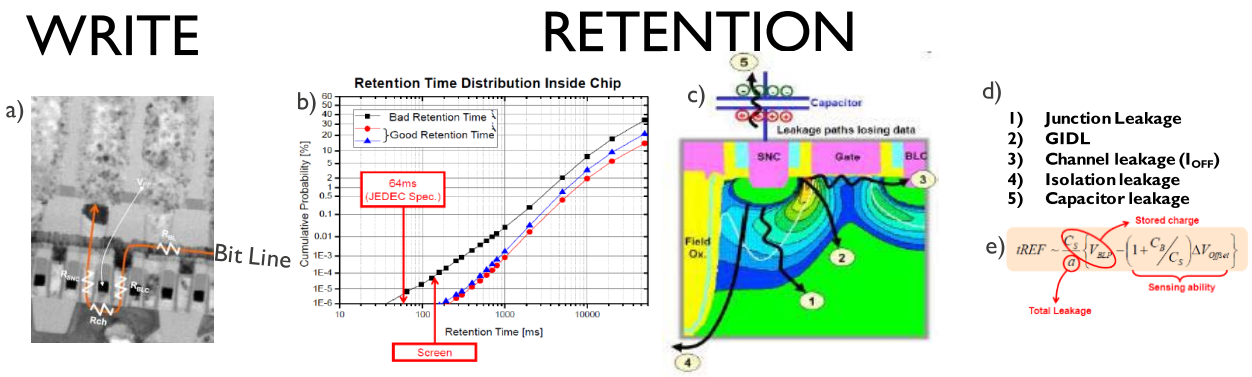


Fig. 4. (a) Write path in the DRAM cell. Current flows from BLs into the access device channel till the capacitor plate. (b) Retention time distribution inside a DRAM chip [5]. Only a few cells are typically exceeding the JEDEC specifications [9]. (c) Major leakage paths listed (d) as separate items. (e) Total leakage contribution to retention time. Adapted from [5].

BLC. Additional leakage contributors are as follows: 1) leakage current under the field oxide from SNC and 2) leakage of the cell capacitor [Fig 4(d)].

The equation reported in Fig. 4(e) shows the relation among the most crucial factors, which can be grouped into three main categories: stored charge, total leakage, and sensing ability which determine the retention time. More specifically, higher storage node voltage (VINTA), higher cell capacitance (C_s), and lower leakage are helpful to increase the retention time. At the same time, higher sensing ability is desired, and this can be obtained by reducing the ratio of BL capacitance (C_{BL}) to C_s and reducing the sense amplifier offset voltage ($\Delta V_{\mathrm{offset}}$).

Practically, higher VINTA is limited by the maximum operating voltage within the circuitry and the consequent power loss and reliability. The lower ration of C_{BL}-to-C_s is required to increase the retention time: C_s needs to be enlarged, despite the dimensional scaling trend that tends to constantly reduce it, and reduced C_{BL} coupling is desired. Another way to improve the retention time is to minimize the leakage current below critical value or to boost the sensing ability that needs to supervise the offset voltage of S/A. In Section III, we will describe the technological solutions adopted so far, and in Section IV, we will review some of the challenges for future technology nodes.

F. Read/Sensing

The basic design of the voltage latch-typed BLSA is based on two cross-coupled CMOS inverters with strong positive feedback. Two NSAs and two PSAs constitute the cross-coupled CMOS. Each S/A is connected with two BLs, and one selected BL compared with an unselected BL which plays the role of reference voltage with a precharged level.

The first step of the sensing is the precharge to the equalized voltage [Bit line precharge voltage (VBLP)] before the activation by the enablement of the BL equalizing signal (PEQ in Fig. 3). The applied equalized voltage is approximately half of the VINTA, which biases the BLs at the same level.

The stored voltage at the cell capacitor is VINTA for data "1" and VSS for data "0" by writing operation.

When the WL is enabled (ACT in Fig. 3), charge sharing takes place between SNC and selected BL. The stored data (VINTA or VSS) are destroyed, requiring a mandatory restore operation at the end of sensing operation. At this moment, charge-sharing time (tCS in Fig. 3) is critical for the following sensing operation, which strongly depends on not only the ratio of BL capacitance (C_{BL}) and cell capacitance but also VINTA. The charge-sharing time can be defined as the time required to achieve enough voltage difference (also called sensing margin) between BL and BL bar (BLB) (see Fig. 3), in other words, data "1" and data "0." After tCS, when the source voltage is applied to the paired pMOS and nMOS through the LA and LAB signals (Fig. 3), the sensing operation gets started. At this moment, the source voltages are array-dedicated voltage (VINTA) and VSS, respectively. Due to the sensing operation, the signal difference can be enlarged and stable sensing operation can be achieved. Based on the enlarged BL developing, BLSA can be ready to take read/write command triggered by CSL enablement. At the same time, destructed data during the charge sharing can be restored. After the restore operation, enabled WL should be turned off, and at the same time, BL and BLB should go back to the precharge level though the precharging operation takes a time window named tRP. Fig. 3 shows the waveform during data "1" sensing operation. As mentioned above, after activating the selected WL with ACTIVE command (ACT in Fig. 3), BLB from unselected MAT maintains VBLP as a reference voltage, while the BL is connected to the accessed cell capacitor with VINTA. Thus, the charge-sharing operation occurs, and due to the charge conservation, the voltage difference on BL can be calculated as follows:

$$\Delta V_{\mathrm{BL}} = \frac{(V_{\mathrm{cell}} - V_{\mathrm{BLP}})}{\left(1 + \frac{C_{\mathrm{BL}}}{C_s}\right)} \tag{2}$$

where V_{cell} can be VINTA for data "1" or VSS for data "0," C_s is the cell capacitance, C_{BL} is the capacitance of the BL, and VBLP is a precharged BL level. After charge sharing, LA and LAB are changed to VINTA/VSS, respectively. This signaling provokes the sensing operation, and thereafter, data "1" sensing on BL is amplified to VINTA by the ON-state of LAB and BLB as a reference is amplified to VSS by the ON-state of LA. Once the read command arrives after the specific time, the amplified output of BL and BLB are delivered to the

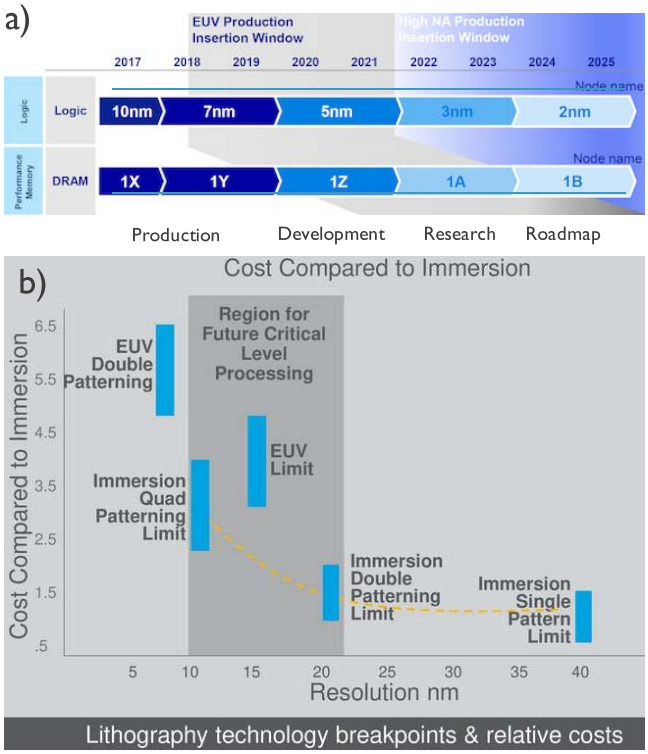


Fig. 5. (a) Forecast of different DRAM technology nodes introduction [12]. EUV is expected to enter HVM in 1Z generation. (b) Cost of ownership of immersion and EUV patterning relative to the cost and the resolution limit [13].

data path. This critical time is defined as tRCD with respect to the column access signal (CAS). Eventually, the precharge command is followed by a tRP timing, which equalizes the amplified BL and BLB levels.

III. Historical Evolution

In this section, we review some of the most relevant historical trends and the latest innovation that is being already introduced by industry in the most recent technology nodes. At present, the 20-nm technology node is an industry mature node, commercially available via multiple companies. Products based on the technology nodes 1X and 1Y are in production across the major DRAM makers [11]. As shown in Fig. 5(a) [11], the technology node 1Z is under development, and nodes 1A (or 1α) and 1B (or 1β) are explored in the research and development as future nodes. At present, some DRAM companies had already announced the architectural pathfinding of 1-γ generation [13]. Section III reviews all the solutions adopted till nodes 1X and 1Y. The technology innovation that might be introduced with the node 1Z and beyond will be discussed in Section IV.

A. Cell Access Device History Review

As we have seen in the previous section, the access cell transistor is the transistor that connects the BL data path to the cell capacitor that stores the charge [Fig. 1(c)], and its gate bias is controlled by the WL bias (V_{WL}). From an electrical perspective, the desired specifications are different from high-performance logic devices because a higher I_{ON}/I_{OFF} ratio ($\sim 10^8$) is required [8], [14]. The low leakage is required to prevent the discharging of the capacitor, and the high ON-current is expected to write the data in a short time. With the shrinking dimensions of the DRAM cell, the OFF-current is increasing due to the degraded short-channel effect, and the ON-current is limited by the reduced effective width [5]. Also, the channel doping concentration increases with the dimensional scaling, increasing the electric field and the junction leakage current, which lowers the retention time.

A simple and effective way to overcome the short channel effect is to increase the channel length (*Leff*). To achieve this scope, the historical trend moved away from a planar architecture to a more complex 3-D type of access device. As final evolution, a possible full vertical integration can be envisioned for future technology nodes, as will be presented in Section IV. In the following, we will review the major historical evolution of the access device, which is also shown in Fig. 6.

1) Planar Asymmetric Junction: One of the key innovations introduced with the 120-nm technology node was the asymmetric junctions [15]. At that time, the access device was still planar type, and the source and drain junction profiles were independently optimized. The junction profile at the storage node was graded to reduce the electric field, which minimizes the junction leakage current and thereby improves the data retention time. On the contrary, the junction profile at the BL direct contact node, which acts as a drain, was designed to be shallower and suppress the short-channel effects of a cell transistor.

2) Stepped Gate STAR: In the evolution toward longer channels by using the 3-D structure, in 2005, a novel step-gated asymmetric (STAR) cell transistors device has been proposed [16], where the channel length increase is obtained by recessing half of the channel and creating an asymmetric junction. This approach was proposed for the 100-nm technology node and was compatible with the lower dimension. Although it does not require the fabrication of deep recess channel, it suffers from poor scalability due to the large V_{th} variations, caused by the misalign effect between the gate and active area [5].

3) Recess Gate RCAT: The ultimate extension of the channel length comes with 3-D structures. Recess channel gate transistor (RCAT) has been introduced by Samsung in the 88-nm technology to increase the effective channel length without compromising the lateral footprint [17]. The recessed channel is obtained by growing the oxide on an etched Si surface, reducing the S/D resistance at the same time, and enhancing the carrier mobility, which can compensate for increased *Leff*. Lower substrate doping concentration and smaller electric field on the storage node junction can be achieved [17]. The lower electrical field generates a significant improvement of data retention time [5], while at the same time, electrical characteristics such as DIBL, break down, junction leakage, and cell channel resistance are improved. Thanks to RCAT, significant improvements in both static and dynamic retention times were achieved. A natural extension of the concept was

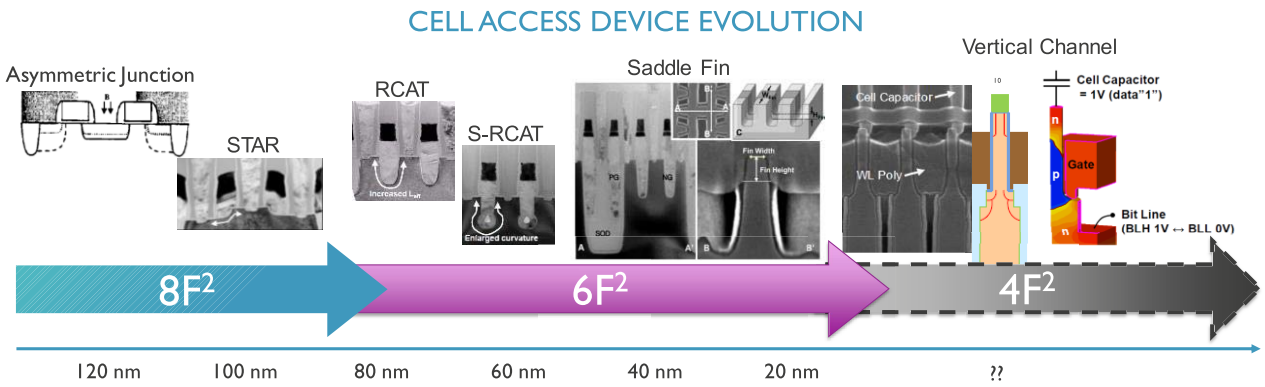


Fig. 6. Review of the historical evolution trend for the cell access device. Various cell access device options are shown. The $4F^2$ is enabled by the vertical channel. Corresponding technology nodes are included. Adapted from [5], [9], [15]–[18], [20], [22], and [23].

the fabrication of deeper recess, needed to extend DRAM technology down to 80 nm and beyond [9].

4) Sphere S-RCAT: A further extension of the *Leff* was achieved by introducing the sphere-shaped-recess-channel-array transistor (S-RCAT) (see Fig. 6) [18]. A simple scaling-down of RCAT would induce worse gate controllability due to the curvature effect. Increasing the curvature radius Rs (with a sphere-shaped channel) helps to maintain the gate controllability and long-channel length when scaling [9].

Thanks to an even enlarged curvature, the S-RCAT can decrease DIBL, body effect, and V_{th}, resulting in the increased retention time [18].

5) Saddle Fin: FinFET was proposed as access devices (e.g., a so-called Omega-FET) to reduce the leakage, thanks to the improved electrostatics [19]. However, the required narrow body width (~2/3 Lg) creates some limitations to the practical utilization in high-volume manufacturing (HVM) [20].

The saddle fin cell transistors have been introduced with the 50-nm technology node, and it is still the device used for the 20-nm technology and beyond. This type of access device combines the recess channel of the RCAT in the channel length direction with the FinFET structure in the channel width direction, and it is fabricated by etching both the field oxide and the active silicon. The combined electrical benefit of both structures is obtained, such an improved short-channel effect, DIBL, and active current due to the partial triple gate [22], which all increases the retention time [20]. In one of the first articles showing this concept, a fin height of about ~40 nm was shown [23] for a sub-50-nm technology.

In terms of achievable specifications, Lee *et al.* [22] showed DIBL ~6 mV/V and subthreshold swing ~85 mV/dec in a 44-nm technology, respectively, thanks to the longer *Leff*.

Other improvements, such as superior gate controllability with a gate-shielded channel region, lead to the improvement of neighbor-gate effects [23]. It is relevant to also observe that the cell-V_T in a fully inverted channel region is mainly controlled by the recess channel single-gate structure. Due to the nature of saddle fin cell transistor, the sensitivity of the V_T versus fin width (W_{Fin}) and fin height (H_{Fin}) is opposite, and it is also more sensitive to W_{Fin} variations than H_{Fin}. These two combined effects enable larger process margins on saddle-fin

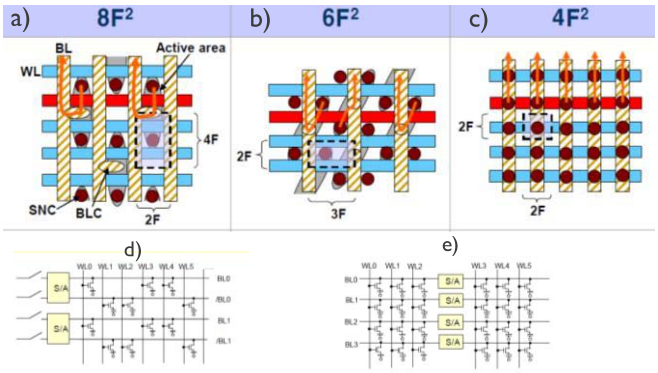


Fig. 7. (a) $8F^2$, (b) $6F^2$, and (c) $4F^2$ design architecture. (d) Folded BL sensing scheme is used in the $8F^2$. (e) Open BL architecture sensing is used in $6F^2$. The $4F^2$ will likely adopt an open BL sensing. Adapted from [5] and [9].

fabrication using the etching process [22], making it still the reference access device for modern DRAM.

B. Cell Architecture

In the traditional DRAM technology, the cell design architecture was based on an $8F^2$ geometry, where F is the minimum feature size for a given technology node. As shown in Fig. 7(a), this design is based on a folded sensing scheme [Fig. 7(d)], where two physically adjacent BLs are connected to the same S/A. Two memory bit cells are connected to the same BL, sharing the drain node of the cell transistor. The required BL pitch is 2F, while the WL pitch is 4F, resulting in an $8F^2$ bit cell area.

The reliability of the operations is one of the major advantages of the $8F^2$, since it offers large noise immunity [5]. Since the paired BLs are connected to the same S/A, they show a similar noise susceptibility and, therefore, higher immunity to external noise. On top of that, all the adjacent BLs to the activated BL are precharged to a reference fixed level, which shields the activated BL itself. Therefore, the impact of the S/A mismatch is reduced.

Since the 80–90-nm technology node [6], [9], the demand for further scaling forced the conversion from an $8F^2$ to a $6F^2$ design architecture [Fig. 7(b)]. The $6F^2$ is based on an

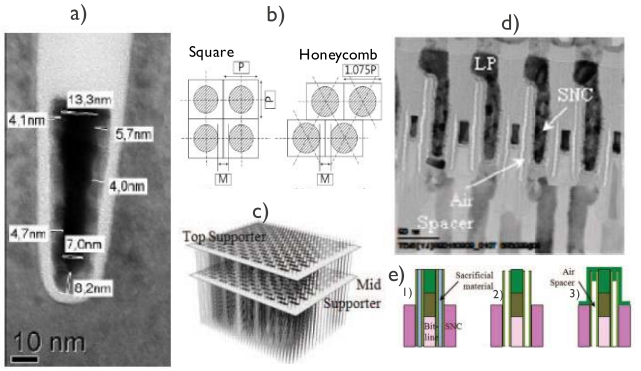


Fig. 8. Different innovations currently used by the state-of-the-art DRAM industry are shown. Scaling enablers like (a) buried WL [14] and (b) honeycomb structures [24] are shown. (c) Supporters enhance mechanical stability [24]. Capacitance reduction can be achieved by (d) air gap [24] with (e) example of integration flow [24], showing: 1) the sacrificial material formation; 2) removal; and 3) capping Si_3N_4 deposition to form air spacer.

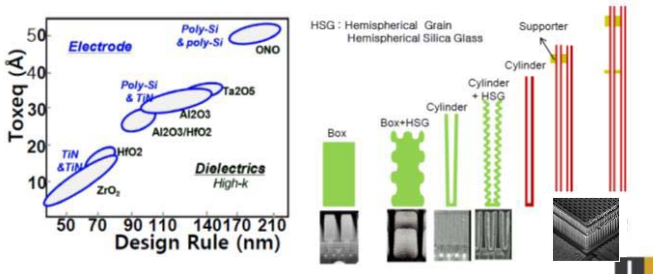


Fig. 9. Material options for capacitor [9] (left). Alternatives of capacitor shape used in the industry (right). Adapted from [9] and [24].

open BL architecture [Fig. 7(e)], with one BL fabricated on the left side of the S/A and another BL on the right side. The S/A pitch can be reduced from 4BL pitch to 2BL pitch [5]. In general, the open BL architecture offers a high degree of regularity, resulting in closer packing of memory bit cells and consequent area reduction. Such area benefit needs to be balanced by higher integration difficulties [6] and larger noise than the $8F^2$.

Another disadvantage of the $6F^2$ is that it requires a dummy array at the edges of the DRAM array to ensure the BL loading matching. Also, since the coupled BLs are coming from different array segments (MAT), they can suffer from the following: 1) noise induced by the adjacent BLs in the selected MAT, which is operating at the same time with different data and 2) process-induced variability, which is expected to be higher in two different MATs with respect to BL located in the same MAT.

To mitigate the noise vulnerability, a buried WL placed below the Si surface is currently used since the ∼40-nm technology [14]. By using a Ti/W-buried WL [Fig. 8(a)], a low resistive interconnect and the metal gate of the array transistor are formed. The buried WL cell can offer two times smaller BL capacitance and three times smaller WL capacitance per cell compared to the conventional cell structure because of the inherently smaller BL to WL coupling [14]. The smaller parasitic capacitances result in faster cell access, less power consumption, and improved signal margin. The drawback of this architecture is the increased gate work function (WF), which might induce a GIDL leakage [5].

A significant component of the BL capacitance is linked to the BL to storage node cell [5], which might account for about half of the entire BL capacitance. To reduce such a coupling between the BL and the storage node cell, an air gap has been proposed since the 20-nm technology [24]. The fabrication of this module is based on a sacrificial material, which is located between the silicon nitride spacers [see Fig. 8(d)]. The sacrificial material is removed by isotropic etching, and the air gap generated is sealed by a capping layer of silicon nitride [Fig. 8(e)]. Interestingly, it has been shown that the voltage breakdown of an air-gap-based spacer increases by reducing the air volume, since the number of molecules that can be involved in the avalanche is reduced. Experimental results show that an improvement of ∼30% in the breakdown is achieved by air-gap spacers with respect to the Si_3N_4 spacers, accompanied by a capacitance reduction of ∼34%.

C. Cell Capacitor Technology Innovations/Trends

We have seen that preventing a drastic C_s reduction with the technology shrinking is one of the biggest challenges for modern DRAM. In the following, we will review some of the solutions used in the industry to mitigate such C_s reduction and present in a class 20-nm technology [24].

1) Material Selection: It is important to realize that the requirements for a capacitor dielectric in DRAM are significantly different from those for logic gate dielectrics [25]. In fact: 1) they are not in contact with Si; 2) the capacitor electrodes are metals, so the band offset requirement is easier; 3) the capacitor is a back-end component so that it only needs to withstand lower temperature processing [25]; and 4) it should be resistant to hydrogen-induced degradation, and usually this requires forming a hydrogen diffusion barrier around it.

Till the 45-nm technology node, HfO_2 was the election material for the DRAM capacitor technology. After it, ZrO_2 dielectric material has been the basis for a decade of DRAM capacitor technology, down to a 25-nm node [6] (Fig. 9, left). In particular, the trilayer structure made of tetragonal (or cubic) ZrO_2 ($k \sim 40$), amorphous Al_2O_3 ($k \sim 9$), and tetragonal (or cubic) ZrO_2 again, which is also called ZAZ, combined with TiN-based electrodes has been widely used in the industry [26]. It should be noted that further scaling the ZAZ nanolaminate for sub-20 nm is challenging from a leakage perspective, due to the reduced physical thickness of the dielectric. We will discuss further material options in Section IV.

2) Deposition Technique: ALD: One of the requirements for the dielectric layer utilized in the fabrication of the capacitor is its conformality in the entire covered surface. Even a tiny difference in the layer thickness can generate a significant E_{field} difference, which is detrimental to leakage control. The most suitable technique to achieve these expectations is represented by atomic layer deposition (ALD), which can

grow a high-quality dielectric layer in a 3-D structure like a capacitor and at relatively low temperature [6].

3) Capacitor Shape: Historically, different shapes of the storage element have been considered to maximize the effective area. Examples of different solutions proposed are shown in Fig. 9 (right).

For the technology nodes beyond 40 nm, the lateral size of the capacitor is significantly larger than the physical thickness of the layers, allowing the usage of a cylindrical capacitor to maximize the effective area [6].

4) Honeycomb Structure (HCS): A honeycomb structure (HCS) has been proposed as a solution to maximize C_s for a given cell size [Fig. 8(b)]. Thanks to different packing patterns, a relative increase of 7.5% of the pitch between two adjacent cell capacitors is reached in the HCS with respect to a square structure (SS) [24]. This corresponds to a potential storage node diameter increase of +11% in the HCS with respect to the SS, which directly translates in higher C_s for a certain capacitor height without increasing the high aspect ratio etching capability [see Fig. 8(b)]. With the same dielectric material, the C_s of the HCS is 21% larger than that of the SS.

5) Supporters to Enable Tall Cylinder: Another way to increase the cell capacitance is to increase the height of the container. One of the problems of this approach is linked to the mechanical stability of the capacitor itself, which can suffer from bending. This is mitigated by introducing a silicon nitride net, which enhanced the mechanical stability [27]. More recently, two supporters have been introduced, one positioned close to the top of the capacitor and the other located approximately in the middle of the DRAM capacitor fabrication [Fig. 8(c)]. Such a solution can allow a significant height increase [24].

IV. Key Technical Aspects and Future Outlook on the Challenges

It was shown that to be economically viable, the expected bit growth should be >50% generation by generation considering the amortization of the required investment to keep pace with the expected cost reduction trend year by year [5]. This expectation is also coming with a constant pressure to decrease the memory size, which is accompanied by a continuous complexity increase in the modern technology nodes. We will illustrate the major technical challenges that are expected in the coming years for future DRAM technologies going toward node 10 nm and below in the following section.

A. Low Capacitance

Fig. 10 shows the expected trend of cell capacitance C_s reduction [9], [28]. The dramatic capacitance reduction associated with scaling represents one of the fundamental problems for further extends the scaling potential.

1) Pillar, Mechanical Stability, and Coverage Uniformity: We have shown that cylindrical capacitors have been used to maximize the effective area of the storage node, and all the state-of-the-art devices are currently based on this technology. On the contrary, for technology nodes beyond 20 nm, the cylindrical structure cannot fit anymore, and pillar capacitors need to be utilized to leverage their smaller feature size. Consequently, a significant increase in the aspect ratio of the pillar height is required even higher than the cylinder case to mitigate the reduction of C_s. Some projections have been shown concerning the expected capacitance required for future DRAM technology nodes, projecting aspect ratio (A/R) > 2.5 times worse than the cylinder [24].

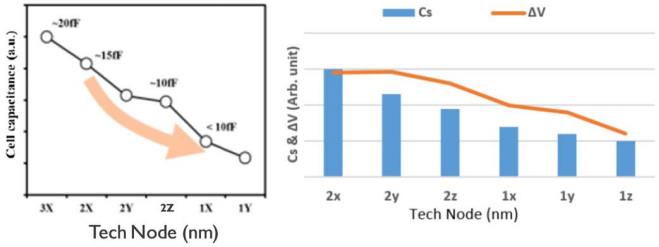


Fig. 10. Expected cell capacitance (C_S) reduction versus technology nodes down to 1Y, with the estimation of the node capacitance (left). Adapted from [8]. C_S and data sensing signal (ΔV) margin (a.u.) versus technology nodes, down to the 1Z [28] (right).

If we want to quantify some numbers, according to [29], for the 1Y node, we can assume a hexagonal pitch of 50 nm and a capacitance target of 8 fF. To respect this target, a cylindrical capacitor requires a hole CD of 32 nm, while the pillar imposes a maximum pillar etch CD of 26 nm. This translates into an aspect ratio >30 for a capacitor and >60 for a pillar, by using a conventional ZAZ material (k∼40) [29]. Clearly, such high A/R values impose critical concern on the mechanical stability and uniformity of the step coverage of a deposited film [28]. Even if multiple supporters can mitigate the mechanical stability concern, as shown in the previous section, the uniformity of the coverage remains a challenge. Partial mitigation can arrive by using double-pillar stacked [9]. A/R ∼ 100 has been mentioned as the maximum sustainable level [6], and such value will be approached soon going down to 1Z and 1A technology nodes without material innovation.

2) Material Innovation for Enhanced Dielectric Performance: Material innovation can help in increasing the dielectric constant and generating a larger bandgap [5]. Even in a pillar structure, the physical thickness of the dielectric needs to be limited to 5 nm and below [6]. In terms of required specifications, a new high-k material for the future DRAM capacitor should enable a low equivalent oxide thickness (EOT) (<0.5 nm) for a physical thickness of <5 nm and ultralow Jg (∼10–7 A/cm^2 at an operating voltage) [6].

Research has been conducted to reduce the thickness of the dielectric material in the capacitor by using a material with the higher dielectric constant. Two of the material that are considered very promising are TiO_2 [30] and $SrTiO_3$ [31], grown by ALD. It has been shown that [25] the k of the oxide is inversely proportional to its bandgap, and in fact, both materials show high-k combined with narrow bandgap (3.2–3.3 eV). In particular, the high k value of SrTiO3 (k > 100) has been shown in combination with capped Jg by controlling the grain size [32] and the stoichiometry with Sr enrichment [33]. Recently, a combination of $SrTiO_3$ deposited on the top of

ALD-Ru has reached a high dielectric constant ($k \sim 118$) and low leakage, which makes this one of the candidates for the future capacitor material [29]. There are, however, some challenges also in this solution. The demonstrated thickness of the layer was ∼11 nm, which is promising but still higher than what needs to be achieved in future technologies. Also, the dielectric properties were shown only in the planar capacitor structure, and the challenge is to demonstrate that the interesting characteristics are kept in a 3-D integration with high A/R.

Another potentially interesting candidate is TiO_2, which can crystallize in two phases, anatase ($k \sim 40$) or rutile ($k \sim 80$). Even if the rutile phase was considered in the past, a high-temperature phase not compatible with the thermal budget of the DRAM fabrication, EOT, and *Jg* reduction have been shown by growing rutile TiO_2 on the top of RuO_2 conductive oxide at a deposition temperature lower than 300 °C [6], [30]. However, the scalability of the physical thickness seems limited to be 10–12 nm, which reduced the hopes to introduce it as a material for future technology nodes.

3) Doping Dielectric: A reduction of the physical thickness can be obtained by material doping [6]. Thickness down to 7 nm with controlled *Jg* has been demonstrated by using the Al-doped TiO_2 dielectric [34].

A doping approach has also been considered for other materials, in particular for HfO_2 or ZrO_2 dielectrics [6]. Dopants with larger ionic radius and lower electronegativity than the corresponding host oxide generate higher k and *Jg* reduction for HfO_2. Rare Earth doping, such as Gd, Er, and Dy, has been proposed as a dopant to reduce the HfO_2 leakage compared with pure HfO_2 [35]. The explanation of the reduced leakage current and equivalent oxide thickness is linked to the stabilization of the higher permittivity tetragonal phase. For ZrO_2 dielectrics, La has been used as a dopant element to achieve $k > 40$ [36]. To the authors' knowledge, these approaches have not been integrated with DRAM capacitors, due to difficulties in implanting species in high A/R 3-D devices [37].

4) Electrodes in DRAM Capacitor: Electrode engineering is another parameter to be leveraged in order to reduce *Jg*. In fact, better dielectric performance can be achieved in a DRAM capacitor by electrodes that have high WF and a sharp interface between the electrode and the dielectric [6]. Currently, TiN grown by ALD is used as the electrode in the DRAM capacitor. However, TiN WF is insufficient to suppress *Jg* at thin dielectric thickness required for 1Z-nm technology nodes and beyond. New electrodes such as noble metals and conducting oxide have been investigated. Ruthenium (Ru) is considered one of the most promising candidates for DRAM capacitor electrodes [6]. Thanks to its relatively high WF (4.8 eV), Ru can suppress *Jg* [38], and the compatibility with dry etch is an advantage for the electrode patterning. One of the drawbacks of Ru is its high surface energy, which makes it difficult to fabricate a continuous and smooth Ru layer of <5-nm thickness. Also, morphological issues (e.g., blisters), which are often occurring on ALD Ru on oxides, need to be addressed to allow the HVM application of Ru as electrode [39].

Considering material with high WF to reduce the leakage, conducting oxides such as RuO_2 and $SrRuO_3$ have WF even higher than the Ru itself [40]. The structural coherency showed by, e.g., rutile TiO_2/RuO_2 [41] and $SrTiO_3/SrRuO_3$ [42] provides EOT and *Jg* reduction. However, these conducting oxides have not yet used in commercial DRAM capacitors due to practical implementation problems. Since the bonding between Ruthenium and oxygen is quite weak, these Ru oxides are easily reduced during the back-end process.

Another candidate proposed as an oxide electrode with high WF and thermal reduction resistant is Ta-doped SnO_2 [43], which has shown promising dielectric properties, high WF comparable to the RuO_2, and thermal stability. In fact, it has been shown that structural and chemical stabilities have been preserved till annealing at 400 °C [43].

B. Access Device Dual WF

It has been shown that high-effective WF generates high band-to-band generation on the drain side, which causes a significant increase in the GIDL that dominates the leakage in the OFF-region [44].

In fact, WF of gate metal has a huge effect on GIDL current. Using a material with a lower WF along all the channel of the access device metal gate causes an unacceptable I_{OFF} increase due to lower V_{th}. However, if higher eWF material is used only on the top part of the recess channel of the access device, the GIDL current is limited and the drive current remains practically unaltered. As shown in [44], by using a combination of TiN with lower (4.5 eV) WF on the top of the recess channel and higher WF (4.66 eV) at the bottom, I_{ON} remains practically unaltered, while I_{OFF} is significantly reduced (e.g., ∼1/400 in $V_g = -0.5$ V at $V_d = 1.5$ V [44]).

Several methods of WF shifting were reported in the literature, and different fabrication techniques based on the combination of metal and poly gate for dual WF can be found [45], [46]. Also, ion implantation of WF species by using a tilted angle has been proposed as a localized way to achieve desired WF in a cost-effective way [44].

C. Device Structure: $4F^2$ and Vertical Transistor

Considering the cell structure, the $4F^2$ is the most compact cell architecture that can allow an area reduction of 33% with respect to a $6F^2$ architecture. Vertical gate (VG) cell (Fig. 6) is considered a promising candidate to enable the $4F^2$ transition. In the literature, there are demonstrations of $4F^2$ cell architecture to further scale down the DRAM cell [44], by using a 30-nm process technology and a VG transistor, which offers superior driving capability than a conventional saddle transistor (Fig. 6).

In a $4F^2$ design, at each intersection of the WL and BL, there is one transistor and one capacitor [Fig. 7(c)]. The vertical pillar is placed on the buried BL and underneath the storage node. This cell design has two advantages since it is highly favorable to lower the BL capacitance, due to the distinct arrangement of the buried BL, and can offer a large on/off signal ratio [44]. One of the major drawbacks of this architecture is the floating body effect, which remains a significant obstacle even if the body of the VG transistor can be connected to the well in a buried BL structure [48].

In fact, at the floating state, GIDL holes accumulated between the storage node and gate increase the body potential, which in turn reduce the V_{th} of a transistor and result in the off-leakage failure and dynamic retention time degradation [48].

Therefore, various approaches need to be considered to reduce the floating body effect in the VG cell scheme. Among the proposed device solutions, we can find the following in [48]: 1) the minimization of the GIDL by using storage node junction engineering [44] or buried body engineering method [49]; 2) BL junction leakage increase by BL junction engineering; 3) hole barrier height reduction between the body and the BL using SiGe layer; and 4) parasitic bipolar gain reduction using the dimension control [50]. Concerning the design solutions to improve the floating body effect biases operation optimization and purging hole charges within the body for a specified time interval are mentioned in [48].

D. Peripheral Transistor Performance

Peripheral devices needed in DRAM memory need to be: 1) low-leakage; 2) cost-effective; and 3) thermally resistant to the high thermal budget that are required by the front-end-of-line (FEOL) and back-end-of-line (BEOL) steps. The combination of all these specifications makes in practice the direct copy of the technological solutions found in logic impossible [51]. Today, the state-of-the-art DRAM technology is still using the planar transistor based on silicon oxynitride (SiON), but to keep pace with the expected system speed improvement, more advanced solutions need to be found [52]. In the following, we will review all the solutions that have been proposed in the literature but have not been implemented yet by the industry.

It is interesting to note that there are three fundamental types of transistors used in a DRAM chip, which have different requirements, as we have seen in Section II, and that can be grouped under the following categories: 1) regular logic, requiring good short-channel control; 2) sense amplifier (S/A), which ideally shows low mismatch and low V_{th}; and 3) thicker oxide to sustain higher bias, where reliability needs to be addressed [5], [9]. Clearly, the needs for the three categories are different, but all have to be accounted for while designing a single periphery platform, which has to remain cost-effective at the same time. It is important to note that the S/A mismatch represents a significant contributor in the overall sensing margin, which is reducing node to node (Fig. 10, right). Nowadays, there are design solutions which have been proposed, at the expense of area and power [28], or process solutions to minimize the random dopant fluctuation, thanks to the reduced dose of halo implant and new implantation schemes which reduced the transient enhanced diffusion (TED) [8].

One of the obvious candidates to improve the electrostatic and, therefore, the electrical characteristics of the logic type is to replace the SiON by high-k/metal gate (HKMG) dielectrics. In 2007, Intel introduced the first high-k-based device, which was fabricated by using a replacement metal gate (RMG) scheme [53]. This specific approach cannot be directly coupled with the thermal budget required by the DRAM, making it gate first integration more suitable for memory applications [54]. Different gate-stack integration proposals have been shown in the literature as compatible with a DRAM, since it induces fermi-level pinning and potential gate leakage issue [55]. This makes the gate first integration scheme more suitable for memory application. Both HfO_2 [54] and HfSiON [56] have been proposed as HVM compatible. HfO_2 is expected to be more scalable for future technology nodes, and HfSiON is mentioned to be more thermally stable [57].

To achieve a wider eWF separation gap between nMOS and pMOS, doping material is desired into the HKMG stack. La and Mg have been proposed as the most promising doping dielectric material for the nMOS devices. La offers a higher achievable V_{th} shift, while Mg in a sandwich of TiN/Mg/TiN offers a more robust thermal stability [54]. Al_2O_3 [58] and SiGe channel [56] have been proposed for the pMOS device DRAM compatible. There are RMG gate-stack proposals which are compatible with the memory flow [55], with TiAl and Ta/Ti used as dipole sources for nMOS and pMOS, respectively. However, the implementation of the RMG scheme has been limited to the capacitor flow.

More recently, a diffusion and replacement metal gate (D&GR) scheme has been proposed [59] to reduce the gate asymmetry between the two device polarities and enhance the thermal stability of the overall flow [60]. The reliability of the HKMG devices has been shown to N/P-BTI, HCD, TDDB, and OFF-state stress [61], [62] for different integration schemes.

On top of the gate stack, all the other modules need to be thermally stable and optimized for memory applications. Contact resistance is one of the bottlenecks when the high thermal budget is considered. A reduced access resistance can be achieved by replacing the silicidation through contact holes by a complete silicidation of the source and drain areas [63]. However, Ni(Pt) silicide typically used in logic devices suffers from too limited thermal stability, imposing the optimization of a dedicated thermally stable silicide (TSS) module [64].

In fact, the thermal stability of a silicide increases with the C incorporation, but this improvement often comes at the expenses of enhanced nMOS V_{th}-Lgate roll-off and pMOS device performance degradation [65]. TSS based on optimized NiPt concentration and combined with preamorphization implant and anneal silicide stabilization steps have been integrated with a DRAM flow without electrical detrimental impact [64].

As a device candidate, FinFET has been proposed as advanced peripheral transistors [52]. In fact, these devices can reduce the area due to the higher effective width per footprint and, at the same time, provide a benefit in terms of mismatch that will benefit the S/A. Significant power-performance benefit with respect to the planar conventional and HKMG devices has been shown for FinFET, keeping the overall cost under control and translating into system benefit [52].

E. Chip Cost Issue and EUV Lithography

The cost of a DRAM chip is mainly determined by the wafer process cost and the total number of bits per wafer [28]. To enable the tremendous geometrical scaling expected from the devices, an expensive 193i lithography process, such as

immersion double patterning and quadruple patterning, has been proposed, and their number is increasing generation by generation [28]. Data reported in the literature show that the reported number of a process step is increasing by >50% from node 2X to node 1Z, with a significant increase of double and quadruple patterning steps [28].

New patterning methods based on extreme ultraviolet (EUV) lithography and nanoimprinting lithography have been considered [66], [67]. At present, none of the DRAM makers have been introduced EUV in HVM, but there are speculations that some DRAM makers will switch from ArF to EUV with the introduction of the generation 1Z (Fig. 5(a)] [12]. It is interesting to note that according to some reports, the single print EUV cost is higher than the quadruple patterning and offers only lower resolution [Fig. 5(b)] [13]. To outperform the resolution of immersion quadruple patterning, EUV double patterning is required, at the expense of higher cost of ownership with respect to the quadruple immersion itself [13].

V. Conclusion

We have reviewed the status, the challenges, and the perspective of 1T-1C DRAM memory chip. Starting from the basic principles of the DRAM, we reviewed the key functional aspects of the memory chip. We have introduced the most relevant historical industry trends to understand how they are related to the challenges faced by modern devices and how the scaling can continue. Finally, the potential critical show stoppers for the future development are discussed to understand where the efforts of DRAM research and development need to focus.

Acknowledgment

The authors would like to thank the imec Core Partners Program for the support. They would also like to thank N. Horiguchi, A. Furnemont, M. H. Na, E. Dentoni Litta, R. Ritzenthaler, and M. Popovici from imec, P. Fazan and C. Mouli from Micron, and C. Kim, Y. Son, and Y. Ji from SK Hynix for the interesting discussions.

References

[1] *DRAM the Invention of On-Demand Data*. Accessed: Jan. 22, 2020. [Online]. Available: https://www.ibm.com/ibm/history/ibm100/us/en/icons/dram/

[2] R. H. Dennard, "Field-effect transistor memory," U.S. Patent 3 387 286 A, Jun. 4, 1968.

[3] *THE Intel Memory Design Handbook*, Intel, Santa Clara, CA, USA, Aug. 1973.

[4] *Who Invented the Intel 1103 DRAM Chip?* Accessed: Jan. 22, 2020. [Online]. Available: https://www.thoughtco.com/who-invented-the-intel-1103-dram-chip-4078677

[5] S. Cha, "DRAM technology-history & challenges," in *IEDM Tech. Dig.*, 2011.

[6] S. K. Kim and M. Popovici, "Future of dynamic random-access memory as main memory," *MRS Bull.*, vol. 43, no. 5, pp. 334–339, May 2018, doi: 10.1557/mrs.2018.95.

[7] C. S. Hwang, "Prospective of semiconductor memory devices: From memory system to materials," *Adv. Electron. Mater.*, vol. 1, Jun. 2015, Art. no. 1400056.

[8] S.-K. Park, "Technology scaling challenge and future prospects of DRAM and NAND flash memory," in *Proc. IEEE Int. Memory Workshop (IMW)*, May 2015.

[9] D. Woo, "DRAM—Challenging history and future," in *IEDM Tech. Dig.*, 2018.

[10] B. Keeth and R. J. Baker, *DRAM Circuit Design*. Hoboken, NJ, USA: Wiley, 2007.

[11] J.-Y. Sim, "Circuit design of DRAM for mobile generation," *J. Semicond. Technol. Sci.*, vol. 7, no. 1, pp. 1–10, Mar. 2007.

[12] M. van den Brink, "Industry roadmap and technology strategy," *ASML Investor Day*, vol. 2018, p. 9, Nov. 2018.

[13] S. D. Boer, "Accelerating memory and storage innovation," Micron, Boise, ID, USA, 2019, p. 13.

[14] T. Schloesser, "A $6F^2$ buried wordline DRAM cell for 40 nm and beyond," in *IEDM Tech. Dig.*, San Francisco, CA, USA, 2008, p. 1.

[15] S. J. Ahn *et al.*, "Novel DRAM cell transistor with asymmetric source and drain junction profiles improving data retention characteristics," in *Symp. VLSI Technol., Dig. Tech. Papers*, Honolulu, HI, USA, 2002, pp. 176–177, doi: 10.1109/VLSIT.2002.1015441.

[16] M. W. Jang *et al.*, "Enhancement of data retention time in DRAM using step gated asymmetric (STAR) cell transistors," in *Proc. 35th Eur. Solid-State Device Res. Conf. (ESSDERC)*, Grenoble, France, 2005, pp. 189–192, doi: 10.1109/ESSDER.2005.1546617.

[17] J. Y. Kim *et al.*, "The breakthrough in data retention time of DRAM using recess-channel-array transistor (RCAT) for 88 nm feature size and beyond," in *Symp. VLSI Technol., Dig. Tech. Papers*, Kyoto, Japan, 2003, pp. 11–12, doi: 10.1109/VLSIT.2003.1221061.

[18] J. Y. Kim *et al.*, "The excellent scalability of the RCAT (recess-channel-array-transistor) technology for sub-70 nm DRAM feature size and beyond," in *Proc. IEEE VLSI-TSA Int. Symp. VLSI Technol. (VLSI-TSA-Tech)*, Hsinchu, Taiwan, Apr. 2005, pp. 33–34, doi: 10.1109/VTSA.2005.1497071.

[19] T. Park *et al.*, "Fabrication of body-tied FinFETs (Omega MOSFETs) using bulk Si wafers," in *Symp. VLSI Technol., Dig. Tech. Papers*, 2003, pp. 135–136.

[20] K.-H. Park, K.-R. Han, and J.-H. Lee, "Highly scalable saddle MOSFET for high-density and high-performance DRAM," *IEEE Electron Device Lett.*, vol. 26, no. 9, pp. 690–692, Sep. 2005.

[21] S.-W. Park *et al.*, "Highly scalable saddle-fin (S-Fin) transistor for sub-50 nm DRAM technology," in *Symp. VLSI Technol., Dig. Tech. Papers*, Honolulu, HI, USA, 2006, pp. 32–33, doi: 10.1109/VLSIT.2006.1705202.

[22] H. Lee *et al.*, "Fully integrated and functioned 44 nm DRAM technology for 1 GB DRAM," in *Proc. Symp. VLSI Technol.*, Honolulu, HI, USA, 2008, pp. 86–87, doi: 10.1109/VLSIT.2008.4588572.

[23] S. W. Chung *et al.*, "Highly scalable saddle-Fin (S-Fin) transistor for sub-50nm DRAM technology," in *Proc. Symp. VLSI Technol.*, 2006, p. 32.

[24] J. M. Park *et al.*, "20 nm DRAM: A beginning of another revolution," in *IEDM Tech. Dig.*, 2015, doi: 10.1109/IEDM.2015.7409774.

[25] J. Robertson, "High dielectric constant oxides," *Eur. Phys. J.-Appl. Phys.*, vol. 28, pp. 265–291, Dec. 2004.

[26] D.-S. Kil *et al.*, "Development of new TiN/ZrO_2/Al_2O_3/ZrO_2/TiN capacitors extendable to 45 nm generation DRAMs replacing HfO_2 based dielectrics," in *Symp. VLSI Technol., Dig. Tech. Papers*, 2006, pp. 38–39.

[27] K. Kim, "Technology for sub-50 nm DRAM and NAND flash manufacturing," in *IEDM Tech. Dig.*, 2005, pp. 323–326, doi: 10.1109/IEDM.2005.1609340.

[28] S. H. Lee, "Technology scaling challenges and opportunities of memory devices," in *IEDM Tech. Dig.*, 2016.

[29] M. Popovici *et al.*, "High-performance (EOT<0.4 nm, Jg~10^{-7} A/cm^2) ALD-deposited Ru\$SrTiO_3$ stack for next generations DRAM pillar capacitor," in *IEDM Tech. Dig.*, San Francisco, CA, USA, 2018, pp. 2.7.1–2.7.4, doi: 10.1109/IEDM.2018.8614673.

[30] S. K. Kim, W.-D. Kim, K.-M. Kim, C. S. Hwang, and J. Jeong, "High dielectric constant TiO_2 thin films on a Ru electrode grown at 250°C by atomic-layer deposition," *Appl. Phys. Lett.*, vol. 85, no. 18, pp. 4112–4114, Nov. 2004.

[31] N. Menou *et al.*, "Composition influence on the physical and electrical properties of $Sr_xTi_{1-x}O_y$-based metal-insulator-metal capacitors prepared by atomic layer deposition using TiN bottom electrodes," *J. Appl. Phys.*, vol. 106, no. 9, Nov. 2009, Art. no. 094101.

[32] O. S. Kwon, S. W. Lee, J. H. Han, and C. S. Hwang, "Atomic layer deposition and electrical properties of $SrTiO_3$ thin films grown using $Sr(C_{11}H_{19}O_2)_2$, Ti(Oi-C_3H_7)4, and H_2O," *J. Electrochem. Soc.*, vol. 154, p. G127, Apr. 2007.

[33] J. Swerts *et al.*, "Leakage control in 0.4-nm EOT Ru/$SrTiO_x$/Ru metal-insulator-metal capacitors: Process implications," *IEEE Electron Device Lett.*, vol. 35, no. 7, pp. 753–755, Jul. 2014.

[34] S. K. Kim, S. W. Lee, J. H. Han, B. Lee, S. Han, and C. S. Hwang, "Capacitors with an equivalent oxide thickness of <0.5 nm for nanoscale electronic semiconductor memory," *Adv. Funct. Mater.*, vol. 20, no. 18, pp. 2989–3003, Sep. 2010.

[35] S. Govindarajan *et al.*, "Higher permittivity rare earth doped HfO_2 for sub-45-nm metal-insulator-semiconductor devices," *Appl. Phys. Lett.*, vol. 91, no. 6, Aug. 2007, Art. no. 062906, doi: 10.1063/1.2768002.

[36] L. Lamagna *et al.*, "Thermally induced permittivity enhancement in La-doped ZrO_2 grown by atomic layer deposition on Ge(100)," *Appl. Phys. Lett.*, vol. 95, no. 12, Sep. 2009, Art. no. 122902, doi: 10.1063/1.3227669.

[37] Q. A. Acton, *Advances in Nanotechnology Research and Application*. Atlanta, GA, USA: ScholarlyEditions, 2013.

[38] M. Popovici *et al.*, "Low leakage Ru-strontium titanate-Ru metal-insulator-metal capacitors for sub-20 nm technology node in dynamic random access memory," *Appl. Phys. Lett.*, vol. 104, no. 8, Feb. 2014, Art. no. 082908.

[39] J.-Y. Kim *et al.*, "Ru films from bis (ethylcyclopentadienyl) ruthenium using ozone as a reactant by atomic layer deposition for capacitor electrodes," *J. Electrochem. Soc.*, vol. 159, no. 6, pp. H560–H564, 2012.

[40] S. K. Kim *et al.*, "Al-doped TiO_2 films with ultralow leakage currents for next generation DRAM capacitors," *Adv. Mater.*, vol. 20, no. 8, pp. 1429–1435, Apr. 2008, doi: 10.1002/adma.200701085.

[41] J. H. Han *et al.*, "Improvement in the leakage current characteristic of metal-insulator-metal capacitor by adopting RuO_2 film as bottom electrode," *Appl. Phys. Lett.*, vol. 99, no. 2, Jul. 2011, Art. no. 022901.

[42] D. Popescu *et al.*, "Feasibility study of $SrRuO_3/SrTiO_3/SrRuO_3$ thin film capacitors in DRAM applications," *IEEE Trans. Electron Devices*, vol. 61, no. 6, pp. 2130–2135, Jun. 2014.

[43] C. J. Cho *et al.*, "Ta-Doped SnO_2 as a reduction-resistant oxide electrode for DRAM capacitors," *J. Mater. Chem. C*, vol. 5, no. 36, pp. 9405–9411, Aug. 2017.

[44] S. K. Gautam *et al.*, "Reduction of GIDL using dual work-function metal gate in DRAM," in *Proc. IEEE 8th Int. Memory Workshop (IMW)*, Paris, France, May 2016, pp. 1–4, doi: 10.1109/IMW.2016.7495287.

[45] J.-Y. Min, S.-H. Lee, H. Hwang, S.-Y. Choi, S. Kang, and D. Woo, "Recess gate transistor," U.S. Patent 8 012 828 B2, Sep. 6, 2008.

[46] V. Ananthan and S. D. Tang, "Dual work function recessed access device and methods of forming," U.S. Patent 8 008 144 B2, Aug. 30, 2011.

[47] H. Chung *et al.*, "Novel $4F^2$ DRAM cell with vertical pillar transistor (VPT)," in *Proc. Eur. Solid-State Device Res. Conf. (ESSDERC)*, Helsinki, Finland, Sep. 2011, pp. 211–214, doi: 10.1109/ESSDERC.2011.6044197.

[48] S. Hong, "Memory technology trend and future challenges," in *IEDM Tech. Dig.*, San Francisco, CA, USA, 2010, pp. 12.4.1–12.4.4, doi: 10.1109/IEDM.2010.5703348.

[49] Y. Cho *et al.*, "Suppression of the floating-body effect of vertical-cell DRAM with the buried body engineering method," *IEEE Trans. Electron Devices*, vol. 65, no. 8, pp. 3237–3242, Aug. 2018.

[50] C. Date and J. Plummer, "Suppression of the floating-body effect using SiGe layers in vertical surrounding-gate MOSFETs," *IEEE Trans. Electron Devices*, vol. 48, no. 12, pp. 2684–2689, Dec. 2001.

[51] A. Spessot, R. Ritzenthaler, T. Schram, N. Horiguchi, and P. Fazan, "Optimized material solutions for advanced DRAM peripheral transistors," *Phys. Status Solidi A*, vol. 213, no. 2, pp. 245–254, Feb. 2016, doi: 10.1002/pssa.201532791.

[52] A. Spessot *et al.*, "Cost effective FinFET platform for stand alone DRAM 1Y and beyond memory periphery," in *Proc. IEEE Int. Memory Workshop (IMW)*, Kyoto, Japan, May 2018, pp. 1–4, doi: 10.1109/IMW.2018.8388823.

[53] K. Mistry *et al.*, "A 45 nm logic technology with high-k+ metal gate transistors, strained silicon, 9 Cu interconnect layers, 193 nm dry patterning, and 100% Pb-free packaging," in *IEDM Tech. Dig.*, 2007, pp. 247–250, doi: 10.1109/IEDM.2007.4418914.

[54] R. Ritzenthaler *et al.*, "A low-power HKMG CMOS platform compatible with dram node 2× and beyond," *IEEE Trans. Electron Devices*, vol. 61, no. 8, pp. 2935–2943, Aug. 2014, doi: 10.1109/TED.2014.2331371.

[55] R. Ritzenthaler *et al.*, "Low-power DRAM-compatible replacement gate high-k/metal gate stacks," *Solid-State Electron.*, vol. 84, pp. 22–27, Jun. 2013.

[56] M. Sung *et al.*, "Gate-first high-k/metal gate DRAM technology for low power and high performance products," in *IEDM Tech. Dig.*, 2015.

[57] M. R. Visokay, J. J. Chambers, A. L. P. Rotondaro, A. Shanware, and L. Colombo, "Application of HfSiON as a gate dielectric material," *Appl. Phys. Lett.*, vol. 80, no. 17, pp. 3183–3185, Apr. 2002, doi: 10.1063/1.1476397.

[58] R. Ritzenthaler *et al.*, "Thermal budget impact on HKMG Al_2O_3 and La gate stacks for advanced DRAM periphery transistors," in *Proc. IEEE Workshop Microelectron. Electron Devices (WMED)* Apr. 2014, doi: 10.1109/WMED.2014.6818721.

[59] T. Schram *et al.*, "Method for manufacturing a dual work function semiconductor device," U.S. Patent 140 106 556, Apr. 17, 2014.

[60] R. Ritzenthaler *et al.*, "A new high-k/metal gate CMOS integration scheme (Diffusion and Gate Replacement) suppressing gate height asymmetry and compatible with high-thermal budget memory technologies," in *IEDM Tech. Dig.*, San Francisco, CA, USA, 2014, pp. 32.3.1–32.3.4, doi: 10.1109/IEDM.2014.7047154.

[61] A. Spessot *et al.*, "Impact of off state stress on advanced high-K metal gate NMOSFETs," in *Proc. 44th Eur. Solid State Device Res. Conf. (ESSDERC)*, Venice, Italy, 2014, pp. 365–368.

[62] M. Cho *et al.*, "Off-state stress degradation mechanism on advanced p-MOSFETs," in *Proc. Int. Conf. IC Design Technol. (ICICDT)*, Leuven, U.K., 2015, pp. 1–4.

[63] C. Ortolland *et al.*, "Carbon-based thermal stabilization techniques for junction and silicide engineering for high performance CMOS periphery in memory applications," in *Proc. Int. Conf. Ultimate Integr. Silicon*, 2009, doi: 10.1109/ULIS.2009.4897559.

[64] T. Schram, A. Spessot, R. Ritzenthaler, E. Rossel, C. Caillat, and N. Horiguchi, "Ni(Pt) silicide with improved thermal stability for application in DRAM periphery and replacement metal gate devices," *Microelectron. Eng.*, vol. 120, pp. 157–162, May 2014.

[65] A. Veloso *et al.*, "Gate-last vs. gate-first technology for aggressively scaled EOT logic/RF CMOS," in *Symp. VLSI Technol., Dig. Tech. Papers*, Honolulu, HI, USA, Jun. 2011, pp. 34–35.

[66] H. K. Kim, "Future of memory devices and EUV lithograph," in *Proc. Int. Symp. EUVL*, 2009.

[67] T. Higashiki, T. Nakasugi, and I. Yoneda, "Nanoimprint lithography and future patterning for semiconductor devices," *J. Micro/Nanolithography, MEMS, MOEMS*, vol. 10, no. 4, 2011, Art. no. 043008.