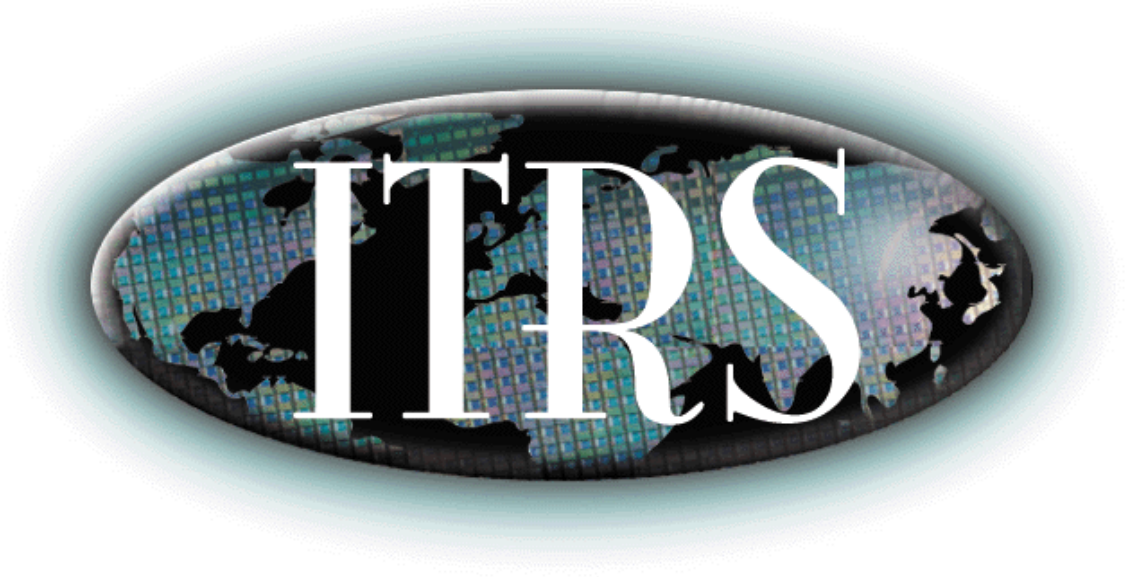


INTERNATIONAL TECHNOLOGY ROADMAP FOR SEMICONDUCTORS

2013 EDITION

PROCESS INTEGRATION, DEVICES, AND STRUCTURES

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

PROCESS INTEGRATION, DEVICES, AND STRUCTURES

1 SCOPE

The *Process Integration, Devices, and Structures (PIDS)* chapter deals with the main IC devices and structures, with overall IC process-flow integration, and with the reliability tradeoffs associated with new options. Physical and electrical requirements and characteristics are emphasized within PIDS. Parameters such as physical dimensions and key device electrical parameters including performance, leakage, and reliability criteria are considered. The focus is on nominal targets, although statistical tolerances are briefly discussed as well. Key technical challenges facing the industry in this area are addressed, and some of the best-known potential solutions to these challenges are discussed. The chapter is subdivided into the following major subsections: logic, DRAM, non-volatile memory (NVM), and reliability.

The main goals of the ITRS include identifying key technical requirements and challenges critical to sustain the historical scaling of CMOS technology per Moore's Law and stimulating the needed research and development to meet the key challenges. The objective of listing and discussing potential solutions in this chapter is to provide the best current guidance about approaches that address the key technical challenges. However, the potential solutions listed here are not comprehensive, nor are they necessarily the most optimal ones. Given these limitations, the potential solutions in the ITRS are meant to stimulate but not limit research exploring novel and different approaches.

1.1 LOGIC

A major portion of semiconductor device production is devoted to digital logic. In this section, both high-performance logic and low-power logic which is typically for mobile applications are included and detailed technology requirements and potential solutions are considered for both types separately. Key considerations are speed, power, density requirements, and goals. One key theme is continued scaling of the MOSFETs for leading-edge logic technology in order to maintain historical trends of improved device performance. This scaling is driving the industry toward a number of major technological innovations, including material and process changes such as higher-K gate dielectrics and strain enhancement, and in the near future, new structures such as gate-all-around (nanowire) and alternate high-mobility channel materials. These innovations are expected to be introduced at a rapid pace, and hence understanding, modeling, and implementation into manufacturing in a timely manner is expected to be a major issue for the industry.

1.2 DRAM

CMOS logic and memory together form the predominant majority of semiconductor device production. The types of memory considered in this chapter are DRAM and non-volatile memory (NVM). The emphasis is on commodity, stand-alone chips, since those chips tend to drive the memory technology. However, embedded memory chips are expected to follow the same trends as the commodity memory chips, usually with some time lag. For both DRAM and NVM, detailed technology requirements and potential solutions are considered.

For DRAM, the main goal is to continue to scale the foot-print of the 1T-1C cell, to the practical limit of $4F^2$. The issues are vertical transistor structures, high-κ dielectrics to improve the capacitance density, and meanwhile keeping the leakage low.

1.3 NON-VOLATILE MEMORY

The NVM discussion in this chapter is limited to devices that can be written and read many times; hence read-only memory (ROM) and one-time-programmable (OTP) memory are not included although many such memories are important both for standalone and embedded applications. The current mainstream NVM is Flash memory. NAND and NOR flash memories are used for quite different applications – data storage for NAND and code storage for NOR flash. There are serious issues with scaling for both NOR and NAND flash memories that are dealt with at some length in the chapter. Other non-charge-storage types of NVM are also considered, including ferroelectric RAM (FeRAM), magnetic RAM (MRAM), and phase-change RAM (PCRAM), all are in volume production. These emerging memories promise to continue NVM scaling beyond Flash memories. However, because NAND Flash and to some extent NOR Flash are still dominating the applications emerging memories have been used in specialty applications

and have not yet fulfilled their original promise to become dominating mainstream high-density NVM. Starting in 2013 edition, resistive memory (ReRAM) is added to the PIDS chapter as a potential solution.

1.4 RELIABILITY

Reliability is a critical aspect of process integration. Emerging technology generations require the introduction of new materials and processes at a rate that exceeds current capabilities for gathering and generating the required database to ensure product reliability. Consequently, process integration is often performed without the benefit of extended learning, which will make it difficult to maintain current reliability levels. Uncertainties in reliability can lead to performance, cost, and time-to-market penalties. Insufficient reliability margin can lead to field failures that are costly to fix and damaging to reputation. These issues place difficult challenges on testing and reliability modeling. This chapter discusses many reliability issues. The goal is to identify the challenges that are in need of significant research and development.

2 DIFFICULT CHALLENGES

The goal of the semiconductor industry is to be able to continue to scale the technology in overall performance. The performance of the components and the final chip can be measured in many different ways; higher speed, higher density, lower power, more functionality, etc. Traditionally, dimensional scaling had been adequate to bring about these aforementioned performance merits but it is no longer so. Processing modules, tools, material properties, etc., are presenting difficult challenges to continue scaling. We have identified these difficult challenges and summarized in Table PIDS1 below. These challenges are divided into near-term 2013-2020 and long-term 2021-2028.

Table PIDS1 Process Integration Difficult Challenges

Near-Term 2013-2020	*Summary of Issues*
1. Scaling Si CMOS	• Scaling of fully depleted SOI and multi-gate (MG) structures • Implementation of gate-all-around (nanowire) structures • Controlling source/drain series resistance within tolerable limits • Further scaling of EOT with higher K materials (K > 30) • Threshold voltage tuning and control with metal gate and high-K stack • Inducing adequate strain in advanced structures
2. Implementation of high-mobility CMOS channel materials	• Basic issues same as Si devices listed above • High-K gate dielectrics and interface state (D_{it}) control • CMOS (*n*- and *p*-channel) solution with monolithic material integration • Epitaxy of lattice-mismatched materials on Si substrate • Process complexity and compatibility with significant thermal budget limitations
3. Scaling of DRAM and SRAM	• DRAM— • Adequate storage capacitance with reduced feature size; implementing high-κ dielectrics • Low leakage in access transistor and storage capacitor; implementing buried gate type/saddle fin type FET • Low resistance for bit- and word-lines to ensure desired speed • Improve bit density and lower production cost in driving toward $4F^2$ cell size • SRAM— • Maintain adequate noise margin and control key instabilities and soft-error rate • Difficult lithography and etch issues
4. Scaling high-density non-volatile memory	• Endurance, noise margin, and reliability requirements • Multi-level at < 20 nm nodes and 4-bit/cell MLC • Non-scalability of tunnel dielectric and interpoly dielectric in flash memory – difficulty of maintaining high gate coupling ratio for floating-gate flash • Few electron storage and word line breakdown voltage limitations • Cost of multi-patterning lithography • Implement 3-D NAND flash cost effectively • Solve memory latency gap in systems

Table PIDS1 Process Integration Difficult Challenges	
5. Reliability due to material, process, and structural changes, and novel applications.	• TDDB, NBTI, PBTI, HCI, RTN in scaled and non-planar devices • Gate to contact breakdown • Increasing statistical variation of intrinsic failure mechanisms in scaled and non-planar devices • 3D interconnect reliability challenges • Reduced reliability margins drive need for improved understanding of reliability at circuit level • Reliability of embedded electronics in extreme or critical environments (medical, automotive, grid...)
Long-Term 2021-2028	• *Summary of Issues*
1. Implementation of advanced multi-gate structures	• Fabrication of advanced non-planar multi-gate and nanowire MOSFETs to below 10 nm gate length • Control of short-channel effects • Source/drain engineering to control parasitic resistance • Strain enhanced thermal velocity and quasi-ballistic transport
2. Identification and implementation of new memory structures	• Scaling storage capacitor for DRAM • DRAM and SRAM replacement solutions • Cost effective installation of high density 3-D NAND (512 Gb – 4 Tb) with high layer numbers or tight cell pitch • Implementing non-charge-storage type of NVM cost effectively • Low-cost, high-density, low-power, fast-latency memory for large systems
3. Reliability of novel devices, structures, and materials.	• Understand and control the failure mechanisms associated with new materials and structures for both transistor and interconnect • Shift to system level reliability perspective with unreliable devices • Muon-induced soft error rate
4. Power scaling	• V_{dd} scaling while supplying sufficient current drive • Controlling subthreshold current or/and subthreshold slope • Margin issues for low V_{dd}
5. Integration for functional diversification	• Integration of multiple functions onto Si CMOS platform • 3-D integration

2.1 NEAR-TERM 2013-2020

[1] Scaling of Si CMOS—

Implementation of fully depleted SOI and multi-gate will be challenging. Since such devices will typically have lightly doped channels, the threshold voltage will not be controlled by the channel doping. The problems associated with high channel doping and stochastic dopant variation in planar bulk MOSFETs will be alleviated, but numerous new challenges are expected. Among the most critical will be controlling the thickness and its variability for these ultra-thin bodies, and establishing a cost-effective method for reliably setting the threshold voltage. Additionally for multi-gate structures, the channel surface roughness may present problems in carrier transport and reliability. These issues will be more severe in nanowire structures.

Controlling source/drain series resistance within tolerable limits will be significant issues. Due to the increase of current density, the demand for lower resistance with smaller dimensions at the same time poses a great challenge. This problem becomes even more severe with thin bodies in SOI and multi-gate structures, and in the extreme case, nanowire structures. It is estimated that in current technologies, series resistance degrades the saturation current by 1/3 from that of ideal case. This proportion will likely become harder to maintain or worst with scaling.

Metal gate/high-K gate stacks have been implemented in the most recent technology generation in order to allow scaling of the EOT, consistent with the overall transistor scaling while keeping gate leakage currents within tolerable limits. Further scaling of EOT with higher-K materials (K>30) becomes increasingly difficult and has diminishing returns. The reduction or elimination of the SiO_2 interfacial layer has been shown to cause interface states and

degradation of mobility and reliability. Another challenge is growing gate dielectrics on vertical surfaces in multi-gate structures. A fundamental burden placed on the overall gate capacitance is the non-scalable quantum capacitance in series with the gate dielectric capacitance.

Threshold-voltage tuning and control with metal gate/high-K gate stacks has proven to be challenging, especially for low-threshold-voltages as V_{dd} continues to go down. For planar bulk devices, this is mainly because of the difficulties in cost effectively and reliably setting the gate stack's effective work-function at or near the conduction band edge for *n*-MOSFETs and valence band edge for *p*-MOSFETs. This issue will be even more critical in fully depleted channels such as multi-gate and SOI, where the effective work-function needs to be in the bandgap (although at different values for *p*-MOSFETs and *n*-MOSFETs), and where the work-function is especially critical in setting the threshold voltage because of the lack of channel doping as a variable. Furthermore, since multiple threshold voltages are sometimes required, an ability to cost effectively tune the work-function over the bandgap would be very useful.

Enhanced channel-carrier low-field mobility and high-field velocity due to internally applied strain is a major contributor to meeting the MOSFET performance requirements. In inducing adequate strain some current process techniques tend to be less effective with scaling. Also, to apply known techniques derived from planar structure to non-planar structures will be facing additional difficulty and complexity. Moreover, transport enhancement is projected to saturate with strain at some point. (For more detail, see Logic Potential Solutions section.)

[2] Implementation of high-mobility CMOS channel materials—

The basic challenges are similar to that of Si CMOS scaling described above. Following presents additional challenges from these new channel materials.

Growing MOSFET quality oxides on III-V materials has long been an industry goal and struggle. Work on the field has been going on for decades, and success has only started to appear only very recently. Nevertheless, there are still much work to be done in the areas of high-K dielectrics, interface quality, yield, variability, and reliability.

Most III-V materials lack good mobility for *p*-type carriers. In order to provide a CMOS solution, Ge is projected to be a good choice, even though it adds complexity to the whole process (see below). A single channel material for both types of channels would be preferable, and materials other than InGaAs are being researched. Ge CMOS is promising for much higher intrinsic mobility for both *n*- and *p*-type carriers compared to Si, but the *n*-channel implementation has been challenging due to source-drain doping and contact problems.

In order to take advantage of the well-established Si platform, it is anticipated that the new high-mobility materials will be epitaxially grown on Si substrate. The lattice mismatch presents a fundamental challenge in terms of material quality and yield, and a practical challenge in cost.

The reason for the requirement of the high-mobility materials to be grown on Si substrate is not only for the established processing steps, but also for the expectation that Si components will be included in the same chips. Examples of these Si based components are embedded DRAM and nonvolatile memories, active analog devices including power devices, analog passives, and large circuit CMOS blocks that do not require high performance but better yield. Integrating these different materials with different process requirements is a huge challenge. Take as an example to integrate Si CMOS with III-V/Ge CMOS. There would be likely three kinds of high-K dielectrics required. Different kinds of metal gates are also required to provide different work functions to yield the necessary threshold voltages. And all processes have to be compatible with one another in terms of thermal budget.

[3] Scaling of DRAM and SRAM—

For DRAM, a key issue is implementation of high-κ dielectric materials in order to get adequate storage capacitance per cell even as the cell size is shrinking. Also important is controlling the total leakage current, including the dielectric leakage, the storage junction leakage, and the access transistor source/drain subthreshold leakage, in order to preserve adequate retention time. The requirement of low leakage currents causes problems in obtaining the desired access transistor performance. Deploying low sheet resistance materials for word- and bit-lines to ensure acceptable speed for scaled DRAMs and to ensure adequate voltage swing on word-line to maintain margin is critically important**.** The need to increase bit density and to lower production cost is driving toward $4F^2$ type cell, which will require high aspect ratio and non-planar FET structures. Revolutionary solution to have a capacitor-less cell would be highly beneficial.

For SRAM scaling, difficulties include maintaining both acceptable noise margins in the presence of increasing random V_T fluctuations and random telegraph noise, and controlling instability, especially hot-electron instability and negative bias temperature instability (NBTI). There are difficult issues with keeping the leakage current within

tolerable targets, as well as difficult lithography and etch process issues with scaling. Solving these SRAM challenges is critical to system performance, since SRAM is typically used for fast, on-chip memory.

[4] Scaling high-density non-volatile memory (NVM)—

For floating-gate devices there is a fundamental issue of non-scalability of tunnel oxide and interpoly dielectric (IPD), and high (> 0.6) gate coupling ratio (GCR) must be maintained to control the channel and prevent gate electron injection during erasing. For NAND Flash, these requirements can be slightly relaxed because of page operation and error code correction (ECC), but IPD < 10 nm still seems unachievable. This geometric limitation will severely challenge scaling far below 20 nm half-pitch. In addition, fringing-field effect and floating-gate interference, noise margin, and few-electron statistical fluctuation for V_t all impose deep challenges. Since NAND half-pitch has pulled ahead of DRAM and logic, lithography, etching, and other processing advances are also first tested by NAND technology.

Charge-trapping devices help alleviate the floating-gate interference and GCR issues, and the planar structure relieves lithography and etching challenges slightly. Recently, high-K IPD and metal gate for planar floating gate Flash memory have been successfully developed and products with 1/2 pitch as small as 16nm have been introduced. Scaling far below 16 nm is still a difficult challenge, however, because fringing-field effects and few-electron V_t noise margin are still not proven and more important, electric breakdown between adjacent word lines may ultimately restrict word line 1/2 pitch to > 10nm.

Endurance reliability and write/read speed for both devices are still difficult challenges for MLC (multi-level cell) high-density applications.

3-D NAND flash is being developed to build high-density NVM beyond 256 Gb. Cost effective implementation of this new technology with MLC and acceptable reliability performance remains a difficult challenge. Contrary to earlier (2011) projection, actual product introduced in 2013 started with larger cell pitch and high layer numbers. Starting with a large layer number will quickly push the layer numbers in the future nodes to > 100 since each new node needs to double the layers. This will cause additional difficult challenges to processing technology to achieve such structures.

[5] Reliability due to material, process, and structural changes, and novel applications—

In order to successfully scale ICs to meet performance, leakage current, and other requirements, it is expected that numerous major processes and material innovations, such as high-κ gate dielectrics, metal gate electrodes, elevated source/drain, advanced annealing and doping techniques, low-κ materials, etc., are needed. Also, it is projected that new MOSFET structures, starting with ultra-thin body SOI MOSFETs and moving on to ultra-thin body, multi-gate MOSFETs, will need to be implemented. Understanding and modeling the reliability issues for all these innovations so that their reliability can be ensured in a timely manner is expected to be particularly difficult.

The first near-term reliability challenge concerns failure mechanisms associated with the MOS transistor. The failure could be caused by either breakdown of the gate dielectric or threshold voltage change beyond the acceptable limits. The time to a first breakdown event is decreasing with scaling. This first event is often a “soft” breakdown. However, depending on the circuit it may take more than one soft breakdown to produce an IC failure, or the circuit may function for longer time until the initial “soft” breakdown spot has progressed to a “hard” failure. Threshold voltage related failure is primarily associated with the negative bias temperature instability (NBTI) observed in *p*-channel transistors in the inversion state. It has grown in importance as threshold voltages have been scaled down. Burn-in options to enhance reliability off end-products may be impacted, as it may accelerate NBTI shifts. Introduction of high-κ gate dielectric may impact both the insulator failure modes (e.g., breakdown and instability) as well as the transistor failure modes such as hot carrier effects, positive and negative bias temperature instability. The replacement of polysilicon with metal gates also impacts insulator reliability and raises new thermo-mechanical issues. The simultaneous introduction of high-κ and metal gate makes it even more difficult to determine and model reliability mechanisms. To put this change into perspective, even after decades of study, there are still issues with silicon dioxide reliability that need to be resolved.

As mentioned above, the move to copper and low-κ dielectrics has raised issues with electromigration, stress voiding, poorer mechanical strength, interface adhesion, and thermal conductivity and the porosity of low-κ dielectrics. The change from Al to Cu has changed electromigration (from grain boundary to surface diffusion) and stress voiding (from thin lines to vias over wide lines). Reliability in the Cu/low-κ system is very sensitive to interface issues. The poorer mechanical properties of low-κ dielectrics also impact wafer probing and packaging. The poorer thermal conductivity of low-κ dielectrics leads to higher on-chip temperatures and higher localized thermal gradients, which impact reliability. The porosity of low-κ dielectrics can trap and transport process chemicals and moisture, leading to corrosion and other failure mechanisms.

There are additional reliability challenges associated with advanced packaging for higher performance, higher power integrated circuits. Increasing power, increasing pin count, and increasing environmental regulations (e.g., lead-free) all impact package reliability. The interaction between the package and die will increase, especially with the introduction of low-K intermetal dielectrics. The move to multi-chip packaging and/or heterogeneous integration makes reliability even more challenging. As currents increase and the size of balls/bumps decreases, there is an increased risk of failures due to electromigration. Cost cutting forces companies to replace gold bond wires to materials like copper, which poses additional requirements in order to make this as reliable as gold.

ICs are used in a variety of different applications. There are some special applications for which reliability is especially challenging. First, there are the applications in which the environment subjects the ICs to stresses much greater than found in typical consumer or office applications. For example, automotive, military, and aerospace applications subject ICs to extremes in temperature and shock. In addition, aviation and space-based applications also have a more severe radiation environment. Furthermore, applications like base stations require ICs to be continuously on for tens of years at elevated temperatures, which make accelerated testing of limited use. Second, there are important applications (e.g., implantable electronics, safety systems) for which the consequences of an IC failure are much greater than in mainstream IC applications.

At the heart of reliability engineering is the fact that there is a distribution of lifetimes for each failure mechanism. With increasing low failure rate requirements we are more and more interested in the early-time range of the failure time distributions. There has been an increase in process variability with scaling (e.g., distribution of dopant atoms, CMP variations, and line-edge roughness). At the same time the size of a critical defect decreases with scaling. These trends will translate into an increased time spread of the failure distributions and, thus, a decreasing time to first failure. We need to develop reliability engineering software tools (e.g., screens, qualification, and reliability-aware design) that can handle the increase in variability of the device physical properties, and to implement rigorous statistical data analysis to quantify the uncertainties in reliability projections. The use of Weibull and log-normal statistics for analysis of breakdown and electromigration reliability data is well established. However, the shrinking reliability margins require more careful attention to statistical confidence bounds in order to quantify risks. This is complicated by the fact that new failure physics may lead to significant and important deviations from the traditional statistical distributions, making error analysis non-straightforward. Statistical analysis of other reliability data such as BTI and hot carrier degradation is not currently standardized in practice, but may be needed for accurate modeling of circuit failure rate.

2.2 LONG-TERM 2021-2028

[1] Implementation of advanced multi-gate structures—

For the long-term years till the end of current roadmap when the transistor gate length is projected to scale below 10 nm, ultra-thin body multi-gate MOSFETs with lightly doped channels are expected to be utilized to effectively scale the device and control short-channel effects. All other material and process requirements mentioned above, such as high-K gate dielectrics, metal gate electrodes, strained silicon channels, elevated source/drain, etc., are expected to be incorporated. Body thicknesses for both SOI and MG below 2 nm are projected and the impact of quantum confinement and surface scattering effects on such thin devices are not well understood. The ultra-thin body also adds additional constraint on meeting the source/drain parasitic resistance requirements. Finally, for these advanced, highly scaled MOSFETs, quasi-ballistic operation with enhanced thermal carrier velocity and injection at the source end appears to be necessary for high current drive. But strain enhancement on these non-planar devices is more difficult.

[2] Identification and implementation of new memory structures—

Increasing difficulty is expected in scaling DRAMs, especially in continued demand of scaling down the foot-print of the storage capacitor. Thinner dielectric EOT utilizing ultra-high-κ materials and attaining the very low leakage currents and power dissipation will be required. A DRAM replacement solution getting rid of the capacitor all together would be a great benefit. The current 6-transistor SRAM structure is area-consuming, and a challenge is to seek a revolutionary replacement solution which would be highly rewarding.

Dense, fast, and low-power non-volatile memory will become highly desirable. Ultimate density scaling may require 3-D architecture, such as vertically stackable cell arrays in monolithic integration, with acceptable yield and performance. 3-D NAND flash will require > 100 layers of stacked devices and processing technology to achieve such structures and cost effective implementation are challenging. Cost effective implementation of non-charge-storage type of NVM is a difficult challenge, and its success may hinge on finding an effective isolation (selection) device. Non-charge-storage NVM may also need to be stacked into 3-D structures to reach Tb density. Without a built-in

isolation device as flash memory, the stacking of these two-terminal devices is both costly and difficult. Much innovation is needed to continue increasing storage density to 1 Tb and beyond.

See Emerging Research Devices section for more detail.

[3] Reliability of novel devices, structures, and materials—

The long-term reliability difficult challenge concerns novel, disruptive changes in devices, structures, materials, and applications. For example, at some point there will be a need to implement non-copper interconnect (e.g., optical or, carbon nanotube based interconnects), or tunnel-based FETs instead of classical MOSFETs. For such disruptive solutions there is at this moment little, if any, reliability knowledge (as least as far as their application in ICs is concerned). This will require significant efforts to investigate, model (both a statistical model of lifetime distributions and a physical model of how lifetime depends on stress, geometries, and materials), and apply the acquired knowledge (new built-in reliability, designed-in reliability, screens, and tests). It also seems likely that there will be less-than-historic amounts of time and money to develop these new reliability capabilities. Disruptive materials or devices therefore lead to disruption in reliability capabilities and it will take considerable resources to develop those capabilities.

[4] Power Scaling—

It is well known that V_{dd} is more difficult to scale than other parameters, mainly because of the fundamental limit of the subthreshold slope of ~60 mV/decade. This trend will continue and become more severe when it approaches the regime of 0.6 V. This fact along with the continuing increase of current density (per area) causes the dynamic power density (proportional to V_{dd}^2) to climb with scaling (although power per transistor is dropping), soon to an unacceptable level. Alternate high-mobility channel materials can provide some relief in this area by allowing more aggressive V_{dd} scaling. On the other hand, for supply voltages lower than ~0.6 V, the circuit margin due to process variability on the threshold voltage needs to be considered. LP technology is specifically designed to minimize the static power.

For high-performance logic, in the trend of increasing chip complexity and increasing transistor on-current with scaling, chip static power dissipation is expected to become particularly difficult to control while at the same time meeting aggressive targets for performance scaling. Innovations in circuit design and architecture for performance and power management (e.g., utilization of parallelism as an approach to improve circuit/system performance, aggressive use of power down of inactive transistors, etc.), as well as utilization of multiple types of transistors (high performance with high leakage and low performance with low leakage) on chip, are needed to design chips with both the desired performance and power dissipation. A trade-off of speed performance for low off-current, or low standby power, is the goal of LP technology.

[5] Integration for functional diversification—

The performance of a chip or technology not only can be measured in speed, density, power, noise, reliability, etc, but also in functionality. There has been an industry trend to include more and more functions on the same chip. Examples are; sensors, MEMS, photovoltaic, energy scavenging, RF and mm-wave devices, etc. Naturally to integrate variety of different materials is a huge challenge. Similarly, integration of high-mobility channel CMOS on Si-based CMOS logic and memories present many challenges as mentioned before.

To improve density on the chip, the trend of the industry is 3-D integration. The impacts within PIDS' scope are induced stress, higher temperature of operation, parasitic capacitances, interference, isolation requirement, process requirements and their compatibility with one another, and device reliability.

3 LOGIC

3.1 LOGIC TECHNOLOGY REQUIREMENTS

The technology requirements address the MOSFET requirements of both high-performance (HP) and low-power (LP) digital ICs. High-performance logic refers to chips of high complexity, high speed, and relatively high power dissipation, such as microprocessor unit (MPU) chips for desktop PCs, servers, etc. Low-power logic mainly refers to chips for mobile systems, where the allowable power dissipation and hence the allowable off-currents are limited by battery life. The LP technology is similar to the LSTP (low standby power) technology of previous years, and the LOP (low operating power or low dynamic power) technology has been eliminated starting from this year.

The main indicator for low standby power is off-current I_{off}. Other leakage currents going through the gate and from the drain junction are assumed smaller so they do not add to this value significantly, although their impact on

reliability is another consideration. For this reason, I_{off} of the LP technology is much lower than that of the HP technology, implying the on-current being also lower as a consequence. The main indicator for dynamic power is CV^2.

In generating the roadmap projection for logic technology, the guiding metric has been the transistor intrinsic speed, the inverse of *CV/I*. (It should be noted that another transistor delay metric, CV/I_{eff}, where I_{eff} is a modified drain current derived from a linear superposition of currents, [1] has been developed and appears to be somewhat more accurate than the $CV/I_{d,sat}$ metric. We are continuing to use the original metric because it is sufficiently accurate to follow the key scaling trends, and for consistency with previous roadmaps.) Logic scaling is characterized by this *I/CV* speed metric, with certain percentage increase per year. This yearly increase is accomplished with a combination of increase of on-current (while fixing the off-current constant), decrease of capacitance by shortening the gate length, and decrease of supply voltage V_{dd}. For many years, this slope had been larger than 10%/year. Recent surveys and literature indicate that the gate-length scaling has been less aggressive than in the past. Similar trend of less rapid increase in circuit clock frequency had been observed at the same time.

To indicate more accurately the practical device speed in real circuits, a better metric would be the ring-oscillator delay. This delay includes more parasitic effects such as the junction capacitance and the interconnect capacitance. This is also the fastest circuit speed that can be physically measured. On the other hand, we note that *CV/I* and the ring-oscillator delay in a properly scaled technology are tracked proportionally [2], so we continue to use the former as the main speed indicator at the device level.

Eventually scaling of MOSFETs is likely to require alternate channel materials in order to continue to improve speed but with low power at the same time. To attain higher drive currents, materials with lighter effective masses are greatly beneficial in quasi-ballistic transport with enhanced thermal velocity and injection velocity at the source end. In current view the materials of choice seem to be InGaAs for *n*-channel and Ge for *p*-channel. The higher performance will likely focus on consuming less power with similar speed (*I/CV*) compared to the Si counterpart. Such technology is anticipated to be in production in year 2018.

In each of these logic devices, multiple parallel paths in transistor structures are sometimes followed. Planar bulk CMOS is extended as long as possible because of its lowest cost, while more advanced CMOS technologies — ultra-thin body and BOX (UTBB) fully depleted (FD) silicon-on-insulator (SOI) MOSFETs and multi-gate (MG) MOSFETs (FinFETs) are implemented for better electrostatic control for improved short-channel effects, and run in parallel with the planar bulk CMOS for some period (for details see the logic tables). There is always a question for the multi-gate structures, whether they will be on bulk wafers or SOI wafers. It is assumed that their intrinsic DC and AC performances are similar in these two different substrates, so they do not affect the outcome of the performance prediction [3]. The issues there have to do with trade-offs in cost, process complexity, variability, and design layout complexity. Hopefully that choice will become clear in the near future. With scaling, difficulties arise with planar bulk MOSFETs because of high channel doping, inability to adequately control short-channel effects, and other issues (for more detail see Difficult Challenges section, Item 1). The advanced CMOS structures (SOI and MG) can be scaled more effectively, and hence can be utilized later in the Roadmap. In fact, multi-gate MOSFET scaling is barely superior with respect to UTBB FDSOI scaling. The physical reasons behind different electrostatic integrity and comparison of scalability of MG, UTBB FDSOI and Bulk are explained in [2]. As the structure showing the best scalability, the MG MOSFET is projected to be feasible till the end of this roadmap period. For the industry as a whole, multiple paths are likely, as different companies choose different timing in extending planar bulk and then switching to the advanced CMOS technologies, depending on their needs, plans, and technological strengths. The multiple parallel paths in overlapped years are meant to reflect this.

Beyond the multi-gate (FinFET) structure, a natural progression would be the gate-all-around (GAA) nanowire structure. This is the ultimate structure in terms of electrostatic control to scale to the shortest possible effective channel length. To accurately project the device performance, 3-D simulation is necessary and it demands much more effort. We regret that there was not enough time to generate results for this structure, but hope to do that in the next edition.

The transistor structures considered in this chapter are shown in Figure PIDS1, which also represent the structure assumptions for the TCAD simulations. In the bulk device, the source/drain profile has a slope tilted away from the channel as shown. This design has been found to be acceptable to emulate a two-step junction for reasonable short-channel effects and minimum series resistance. (See more discussion on series resistance later.) All effective channel lengths are assumed to be 80% of the gate lengths, so the gate/source and gate/drain overlaps are 10% of gate length at each side. Note that for the MG structure, the figure becomes a top view of a FinFET, so it is considered to be of infinite height. It should be noted that the current unit is per channel width, so for FinFET, it would be per height

(which is channel width) for each gate or each side of the fin. In other words we are speaking here of a normalization per unit length of the inversion width, rather than per footprint.

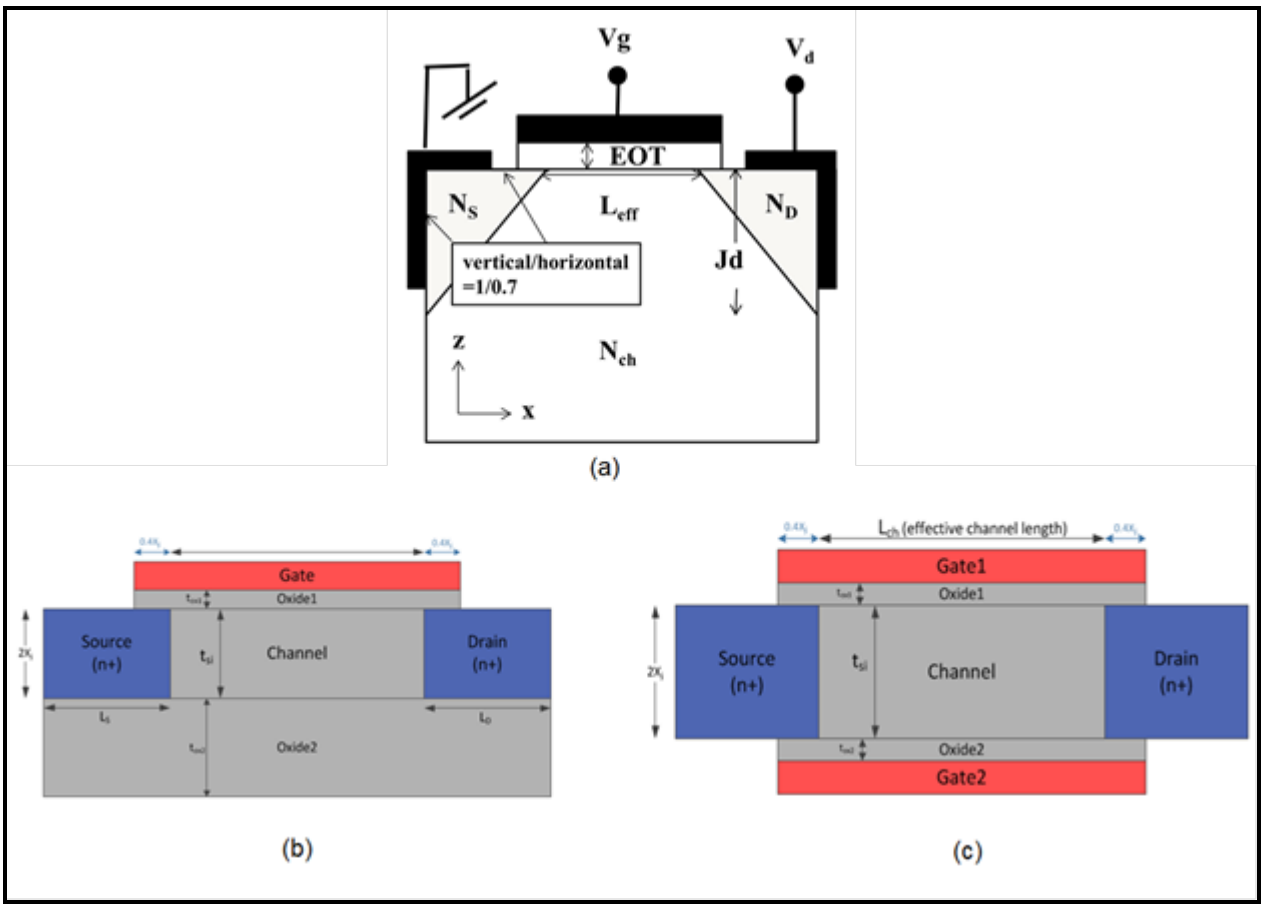


Figure PIDS1 Transistor Structures used for Simulations: (a) Bulk, (b) SOI, (c) MG

All dopings are assumed uniform in order to keep the number of variables under control. It is also impossible to bench-mark reported data without this degree of details given in literature. One then has to interpret the uniform doping as an equivalent doping level. The junction profiles are also assumed to be abrupt or step-like. This is done for the same reasons.

It has been known that even for the same EOT, different K-value corresponds to different physical oxide thickness, and that has effect on the potential distribution. For this reason, starting from this year, the K-value is assumed and entered into the requirement tables. Even though in practice, the total gate oxide probably consists of different regions of compositions. The single K-value can be looked at as the equivalence of the total.

The intrinsic series resistance of a MOSFET is mainly determined by the S/D doping gradient, S/D sheet resistance, and the metal-semiconductor contact resistivity, all are seldom provided in literature for calibration. So the values for series resistance are not calculated from device structures. Rather, they are given here as a tolerable value that degrades the ideal current with zero series resistance by 33%.

The specific set of projected parameter values in each of the tables reflects a particular scaling scenario in which the targeted values for the key outputs are achieved. However, since there are numerous input parameters that can be varied, and the output parameters are complicated functions of these input parameters, other sets of projected parameter values (i.e., different scaling scenarios) may be found to achieve the same target. For example, one technology would scale the EOT more aggressively by introducing higher-K dielectric, while another would achieve equivalent results by optimizing doping or/and strain enhancement. Hence, the scaling scenarios in these tables only

constitute a good guidance for the industry but are not meant to be unique solutions, and there will be considerable variance in the actual paths that the various companies will take.

It should be emphasized that due to the rough assumptions that have been made, it is difficult to achieve absolute accuracy. As mentioned, calibration with the published data cannot be 100% complete due to the lack of details provided. Another factor is the long-range projection of 15 years. We can only hope for or claim accuracy in the order of ±20%. It is our goal that the relative performance and the trend can be of better value than the absolute values themselves.

For generating the entries in the logic technology requirement tables, an MOSFET modeling software MASTAR has been used [4]-[6]. The software contains detailed analytical MOSFET models that have been verified against literature data. It is well suited to efficiently analyzing technology trade-offs, and has been used for the PIDS calculations for many years. The MASTAR program and the specific MASTAR input and output files are available to the public to be downloaded from the ITRS website, with the goal that readers can reproduce the results on their own. MASTAR is a compact-model based software, different from finite-element, numerical TCAD programs. While it has the advantages of simplicity, the inputs and methodology are less fundamental compared to those of TCAD. Transport parameters are assigned as inputs to control the values of mobility and saturation velocity, and the degree of ballistic transport.

As the channel length gets scaled into the sub-10-nm range near the end of the Roadmap, it would require TCAD device simulators to account for quantum effects such as tunneling and quantum confinement. For this edition, for the first time we have been working to develop TCAD tools for ITRS purposes. The philosophy of PIDS has been that the tools and input files are accessible by the general public, so everyone if can reproduce the results, or if interested, can vary the input parameters to see their impacts. This constraint rules out commercial TCAD tools.

We are fortunate to be able to get support from the NanoHub Team of Purdue University. Not only NanoHub is well-known for its variety of tools and lectures, its nature of open-forum is particularly fitting for ITRS' purposes [7]. The goal is to have an ITRS site within the NanoHub, where the simulation tools, input files, as well as instructions will reside [8].

Two TCAD tools have been identified and modified for the purpose of ITRS. Padre [9] will be used to simulate bulk devices whereas NEMO5 for SOI and MG devices [10].

Padre is based on drift-diffusion (DD) model and faces critical challenges since the quantum confinement effects and the ballistic effects are becoming prominent in MOSFETs with gate lengths less than 20 nm [9]. Atomistic full-band simulation can in principle handle the required physics [10]. It is most suitable for ultra-thin bodies such as in SOI, FinFET or nanowire FETs. For bulk MOSFETs, however, such full atomistic quantum approaches require very large computational resources and long run times due to the large device volume. The inclusion of incoherent scattering in full quantum transport in bulk MOSFETs appears to be completely prohibitive [11]. Other simulation methods such as Monte-Carlo, energy valence, hydrodynamic, and density gradient method are also computationally heavy and slow in convergence. Especially the energy balance [12] and the hydrodynamic model [13] can overestimate the current. The goal of modification is to extend the usability of the industrial standard TCAD tools based on the DD model [14] to the next generations of nano-scale bulk MOSFETs for L_g down to 14 nm. The limitation of the current model is that it does not include the source-to-drain tunneling and the variability effects due to random dopant fluctuation, so it should be used with a great care. In the boundary of this limit, through inclusion of quantum corrections, the DD simulator is able to describe the most recent experimental *I-V* characteristics for *n*-type bulk MOSFETs with L_g = 32 nm and 20 nm. The new model is also benchmarked with the MIT virtual source (MVS) model [15] at L_g = 18 nm. Quantum mechanical confinement effects are included by equivalent oxide thickness (EOT) modifications in the same way as in MASTAR. Ballistic resistance and ballistic transport are accounted for by modifications of the longitudinal field-dependent mobility model:

$$v = \frac{\mu_0 E}{1+\left[\left(\frac{\mu_0 E}{v_{sat}}\right)^{\beta}\right]^{\frac{1}{\beta}}} \qquad (1)$$

NEMO5, an atomistic quantum transport simulator based on Quantum Transmitting Boundary Method (QTBM) with the nearest-neighbor sp^3d^5s tight-binding (TB) [16] is used to calculate intrinsic device characteristics in the ballistic regime for both SOI and MG devices. The validity of the TB band structure is confirmed via comparison against first-principle electronic structure calculations in ultra-thin body (UTB) silicon for different SOI thicknesses. To capture the scattering effect, the Lundstrom model is used. After calculation of the device ballistic characteristics, backscattering model [17]-[20] is applied using the following equations [17]:

$$T_c = \frac{\lambda_0}{\lambda_0 + 2l_{kT}} \quad (2)$$

with high V_{DS}:

$$I_{Scattering} = \frac{T_c}{2 - T_c} I_{Ballistic} \quad (3)$$

with low V_{DS}:

$$I_{Scattering} = T_c \times I_{Ballistic} \quad (4)$$

where T_c is the transmission coefficient and l_{kT} is effective on-current channel length, which is the distance between top of the barrier to location of one kT lower [19],[21],[22]. This value is calculated from ballistic potential profile for each bias point. The mean free path (λ) value is required to include scattering effect by backscattering model. This value is extracted from the experimental reported values [23]-[26] for different UTB body thicknesses and different charges under the gate. The approach that we used for calculation of mean free path is as below:

$$\lambda = \frac{V_{ds}}{v_{inj}} \mu \quad (5)$$

where V_{ds} is very low, i.e., 5 mV and v_{inj} is calculated by dividing the current by the charge at top-of-the-barrier. μ is mobility which is dependent on the charge under the gate and device body thickness.

Further assumptions for TCAD simulations are on orientation (100) and without strain. Metal gates are assumed so there is no depletion layer at the gate-insulator interface. The work function of the metal is chosen to adjust the threshold to match the pre-determined subthreshold off-current I_{off}.

Since TCAD is used for the first time here for ITRS, we keep the near-term projections by MASTAR for the sake of continuity, for both Si HP and LP technologies. For III-V/Ge, due to the lack of time, the same results generated by MASTAR from previous edition are reproduced here. It is our goal that we will be able to generate more accurate results by TCAD on these alternate channel materials as well.

Before we move to description of specific assumptions relative to HP and LP transistors, let's comment on three specific features that are related to UTBB FDSOI and MG technologies. Those two technologies continue transistor scaling whenever bulk planar stalls due to its inability to sustain short-channel control.

1. Short-channel control: MG has good electrostatics integrity due to its tall narrow channel that is controlled by a gate from three-sides where this allows relaxing the scaling requirements of fin thickness (i.e. body thickness) compared to UTBB FDSOI. In UTBB FDSOI electrostatic control could be done by using silicon (i.e. body) thickness and BOX thickness [27] where convergent scaling of both silicon thickness and BOX thickness enables electrostatics scaling (DIBL < 100 mV/V) down to Lg beyond 10 nm. Tbox and Tsi scalings are typically kept at compromise between manufacturability and short-channel-effects control.
2. Drive at unit footprint: MG has a better drive at unit footprint (by enabling drive in the third dimension) if fin pitch can be aggressively scaled. This increased drive at unit footprint by scaling the fin pitch comes at a trade-off between increased fringing capacitance between gate and contact and increased series resistance. This compromise in fringe capacitance is reflected in the PIDS tables. In the current PIDS tables we normalize the current to the MG periphery, not to the footprint.
3. Dynamic backgate control to modulate Vt: In UTBB FDSOI the body effect remains efficient and also has a more comfortable range with respect to planar, i.e. it is possible to go up to several times Vdd, backward and forward, thanks to the isolation of the S/D diodes by the BOX from the substrate. The threshold voltage can be modulated up to as much as 100 mV upwards and downwards with BB of -Vdd and +Vdd, respectively [28]. The compromises of this flexibility are the increased subthreshold leakage resulting from the reduced Vth and floorplan requirements to implement back-gate control. In MG (e.g. finFET) technologies the body bias is not efficient due to lack of penetration of the body potential into the Fin. The body of the fin is completely controlled by the gate field that is by the way the reason behind the very good electrostatic integrity of FinFET. Typically multi-Vt in finFET is handled by the gate work function tuning during process or changing gate CD during design-only. We do not consider different Vt options in the PIDS tables since the device parameters consider different Ioff values that already leads up to different Vts.

For the high-performance HP logic technology, the transistor structure assumptions are listed in Table PIDS2a. The scaling difficulties for planar bulk are reflected by the required channel doping which increases sharply with year, to a

very high value of $9x10^{18}$ cm^{-3} in 2017. In the present edition the UTBB SOI roadmap is not conducted beyond 2020, and only the MG structure continues until the end of the roadmap 2028. This will be reconsidered with more simulation results in next ITRS editions.

Table PIDS2a High-performance (HP) Logic Technology Requirements - TCAD

The transistor off-current, I_{off}, is fixed at a value of 100 nA/µm for all years (for HP devices), which has important consequences for the chip static power dissipation. The *n*-channel MOSFET saturation drive current, $I_{d,sat}$, is found to increase only for a few years and then starts to drop. This is drastically different from the previous projections where the current continued to rise with year. The reason for the drop of current is mainly due to V_{dd} scaling and significant source-drain tunneling which comes to the picture for channel lengths below 10 nm. This source-drain tunneling makes the device harder to turn off and increases the subthreshold swing (SS) [27]. The tunneling current requires the threshold voltage to be higher in order to maintain the fixed I_{off} [27], and consequently leads to a reduction in the inversion charge.

An important speed metric for the transistor is the intrinsic speed *I/CV* where *C* includes the gate capacitance plus the gate fringing capacitances. These fringing capacitances have been found to be larger than the intrinsic capacitance over the channel region. As shown in the table, the ratio of total fringing capacitances to the gate capacitance over the channel is assumed to increase with scaling and saturates at 2.0 [30]. Figure PIDS2 captures graphically all the important scaling trends listed in Table PIDS2a. The final plot indicates that the intrinsic speed improves initially with a slope of 4% per year, and then levels off afterwards. The decreased *I/CV* slope compared to previous projections is a consequence of the trend of $I_{d,sat}$.

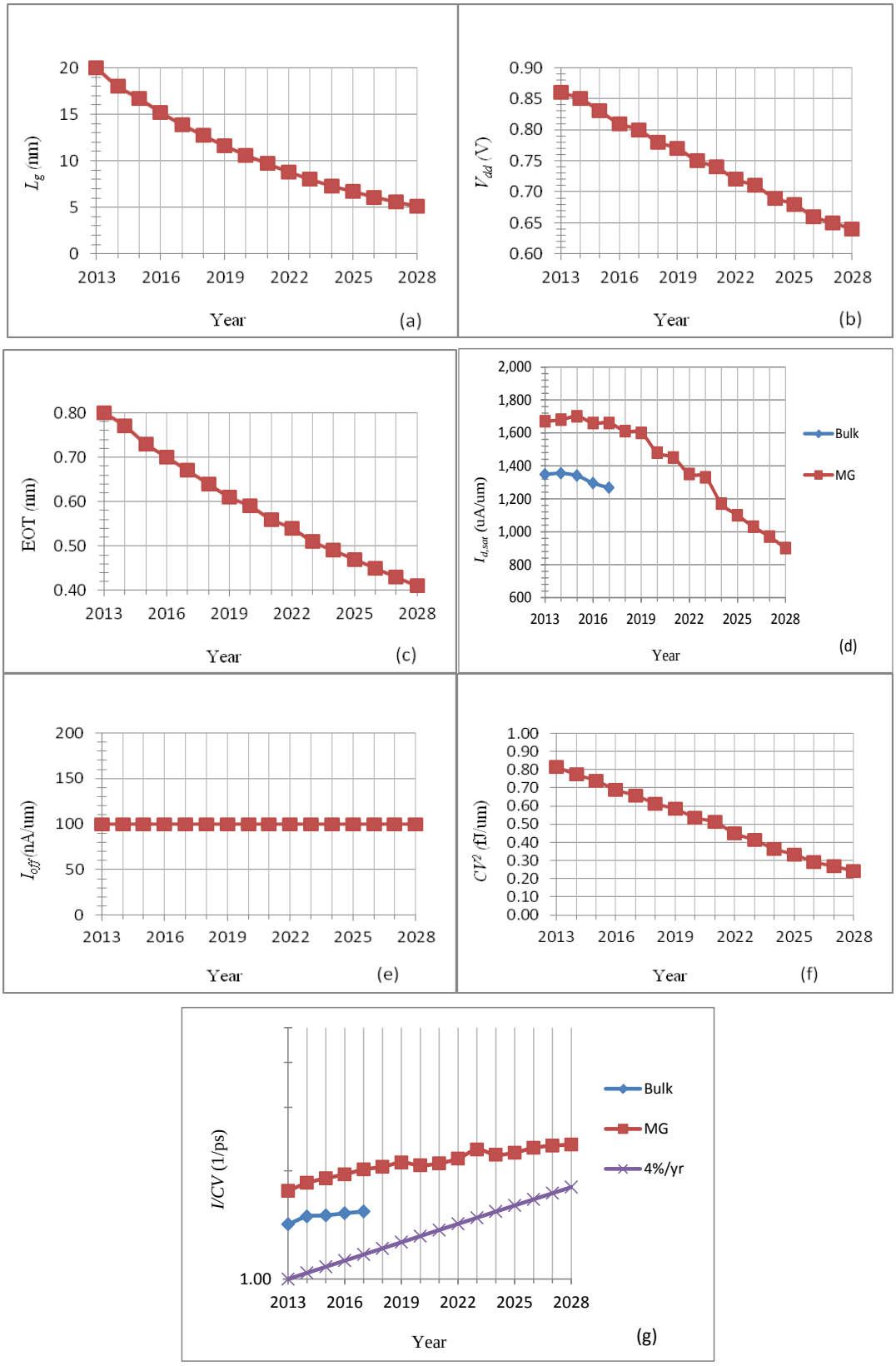


Figure PIDS2 Scaling Trend of Logic HP Transistors. (a) Gate length, (b) Supply voltage, (c) EOT, (d) On-current Id,sat, (e) Off-current Ioff, (f) Dynamic power CV2, and (g) Intrinsic speed (I/CV)

For continuity, HP transistor results generated from MASTAR for the short-term years are listed in Table PIDS2b.

Table PIDS2b High-performance (HP) Logic Technology Requirements - MASTAR

The main feature of LP transistors is low DC power with reasonably reduced speed. The guidelines we have chosen are that the I_{off} is set at 10 pA/um (four decades lower than that of HP) with an on-current at least 35% of the HP transistor. If the on-current drops below that level, I_{off} is allowed to increase until this $I_{d,sat}$ criterion is met. Note that to meet the leakage current requirement, the gate length scaling of LP logic lags behind that of HP logic. As shown in Table PIDS3a, EOT and V_{dd} are kept the same as those of the HP transistors. Figure PIDS3 shows the important scaling trend of LP transistors. Here the intrinsic transistor speed *I/CV* increases only at a slope of ~2% per year. For LP transistors, the criterion of source/drain series resistance is changed. We simply assume they have the same values as the HP transistors because of the available contact technology at the same period.

Table PIDS3a Low Power (LP) Technology Requirements - TCAD

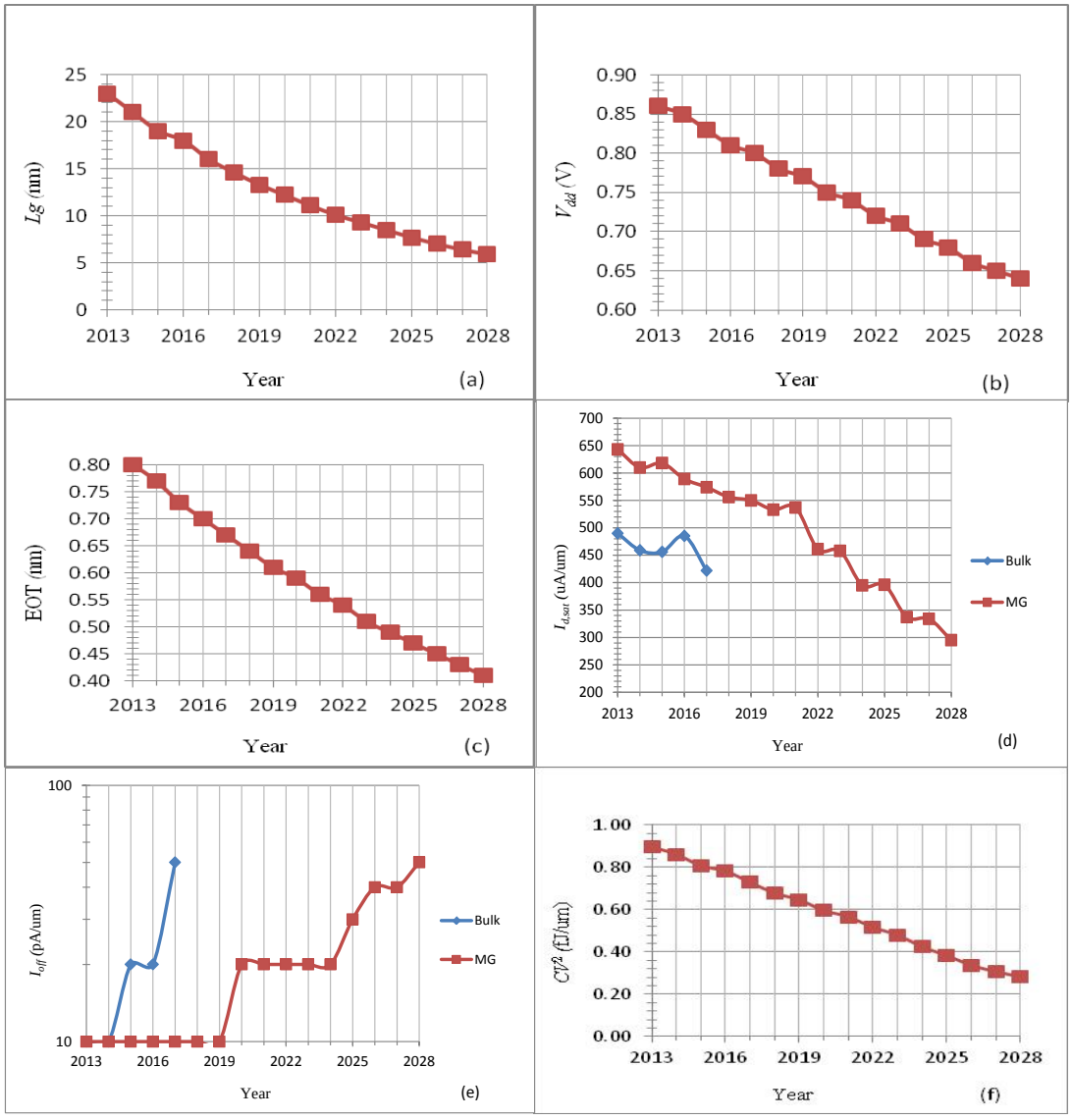


Figure PIDS3 Scaling Trend of Logic LP Transistors. (a) Gate length, (b) Supply voltage, (c) EOT, (d) On-current Id,sat, (e) Off-current Ioff, (f) Dynamic power CV2 and (continued on next page)

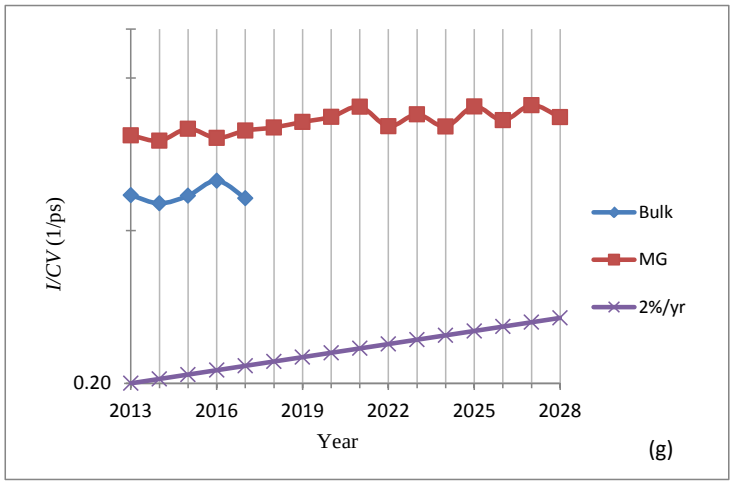


Figure PIDS3 (continued) Scaling Trend of Logic LP Transistors. (g) Intrinsic speed (I/CV)

LP results generated from MASTAR for short-term years are shown in Table PIDS3b for comparison.

Table PIDS3b Low Power (LP) Technology Requirements - MASTAR

Between HP and LP transistors, it is a matter of trade-off between speed (high $I_{d,sat}$) and DC power (low I_{off}), because they are both based on the same channel materials of silicon. It is well known that for alternate channel materials of lighter effective masses, the carrier transport is much improved, leading to higher channel currents. In current view, the leading candidates are InGaAs for *n*-channel and Ge for *p*-channel. Since in reality, power is a more severe limitation in a large IC chip, the III-V/Ge technology is expected to have a much lower V_{dd}, and with a slightly higher speed at the same time. The gate length is estimated to be lagging from that of HP by about one year, since it involves a completely new set of materials. The technology requirements and results are shown in Table PIDS4. These results are reproduced from the last edition, and had been generated by MASTAR. It is our goal that we will also use TCAD to update the results in the next edition. It should be cautioned that comparison of Si HP and LP transistor to III-V/Ge is not very meaningful here as MASTAR has different assumptions such as gate fringing capacitances.

Table PIDS4 III-V/Ge High-performance Logic Technology Requirements - MASTAR

To have an overview of the three device technology options HP, LP, and III-V/Ge, they are summarized in the following table (Table PIDS5) in terms of speed, dynamic power, and static power. Ultimately for the same material system, the trade-off between different logic technologies is speed vs. power which is consisted of static power and dynamic power. Here we only list the ratio in relation to the values in HP. It can be seen that between HP and LP, there is trade-off of speed, static power, and dynamic power because they are both Si-based technologies. Whereas for III-V/Ge, there is a net improvement in both speed and power since it is a completely different material system.

Table PIDS5 Comparison of HP, LP, and III-V/Ge Technologies in terms of speed and power.

	HP	LP	III-V/Ge
Speed (*I/CV*)	1	~0.4	>1
Dynamic power (CV^2)	1	~1	<1
Static power (I_{off})	1	~1x10^{-4}	1

3.2 LOGIC POTENTIAL SOLUTIONS

There is a strong correlation between the challenges indicated by the colors in the technology requirement tables and the potential solutions (Figure PIDS4). In many cases, red coloring (manufacturable solutions are not known) in the technology requirement tables corresponds to the projected year of introduction for a potential solution to the challenge. Another important general point is that each potential solution involves significant technological innovation. The qualification/pre-production interval has been set to two years or more in order to understand and deal with any new and different reliability, yield, and process integration issues associated with these innovative solutions. Many of the potential solutions may be required first for high-performance logic, followed by the low-power option. Finally, the industry faces an overall challenge due to the sheer number of major technological innovations required over the next five years to continue to improve year over year: enhanced mobility from strain, higher K values for gate dielectrics, ultra-thin bodies for fully depleted SOI and multi-gate MOSFETs, and controlling series resistance with smaller dimensions.

After the transitions from planar bulk to SOI to MG, the next natural evolution would be gate-all-around or nanowire transistor structure. This form provides the ultimate electrostatic control of the channel by the gate, and will give the best performance in terms of short-channel effects. By that structure, the shortest effective channel length of a material system can be realized. Carbon nanotube would also fall into this category.

Eventually later in the roadmap, more forward-looking solutions in the utilization of alternate channel materials to further enhance the transport will be adopted. It is anticipated the first solutions would be III-V (for *n*-channel) and Ge (for *p*-channel) combination, still based on MOSFET operation. It is projected the first product will be introduced in 2018. Other possibilities beyond these semiconductors are 2-D crystals. These include grapheme, boron nitride (BN), dichalcogenides such as MoS_2, WS_2, $NbSe_2$, and complex oxides such as $Bi_2Sr_2CaCu_2O_x$.

As scaling continues, the power density of the IC continues to go up with the transistor density, although the power per transistor goes down. An effective solution would be based on transistor actions that do not depend on the Boltzmann distribution which sets a lower limit of subthreshold slope of 60 mV of gate voltage per decade of channel current. One such conduction mechanism is tunneling. A class of transistor based on this effect is called tunneling FET (TFET) [31]. It is basically a *p*-*n* junction placed under an MOS gate. With a proper design of the heterojunction under the gate, ultra-low V_{dd} operation is the goal.

Another means to achieve sharp subthreshold slope is by incorporating ferroelectric gate dielectrics in an MOSFET [32]. When the transistor is biased towards the on-condition, the electric field moves the charges within the ferroelectric gate oxide, and that polarization further reduces the threshold voltage, resulting in a higher gate over-drive, as if a higher gate voltage was applied. This internal gain, sometimes called negative capacitance, creates an effect of deeper subthreshold slope. The goal also is operation with ultra-low V_{dd} and low power.

Finally, beyond the roadmap range of this edition (beyond 2028), MOSFET scaling will likely become ineffective and/or very costly. Completely new, non-CMOS type of logic devices and maybe even new circuit architecture are potential solutions (see Emerging Research Devices section for detailed discussions). Such solutions ideally can be integrated onto the Si-based platform to take advantage of the established processing infrastructure, as well as being able to include Si devices such as memories onto the same chip.

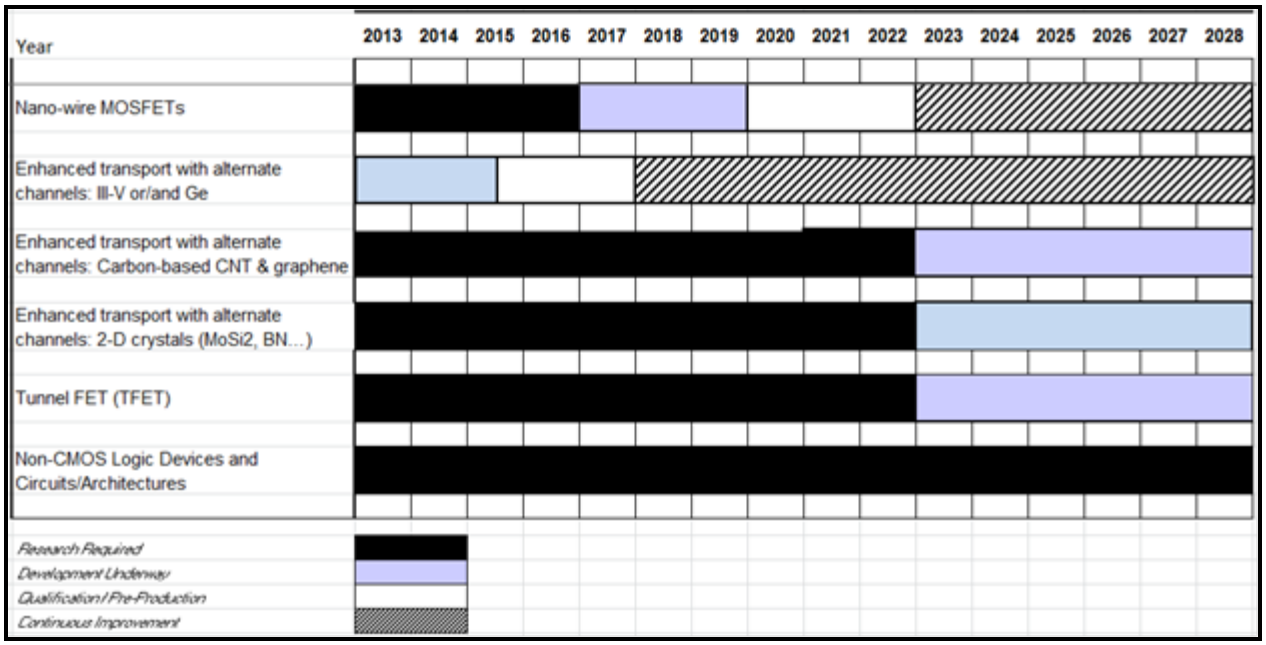


Figure PIDS4 Logic Potential Solutions

4 DRAM

4.1 DRAM TECHNOLOGY REQUIREMENTS

In general, technical requirements for DRAMs become more difficult with scaling (see Table PIDS6). In the past couple of years, DRAM was introduced with many new technologies (e.g. 193 nm argon fluoride (ArF) immersion high-NA lithography with double patterning technology, improved cell FET technology including fin type transistor [33]-[35], buried word line/cell FET technology [36] and so on). Due to new technologies, DRAM will continue to scale with 2-3 year cycle and 20 nm HP (minimum feature size) DRAM will be available by 2017.

Table PIDS6 DRAM Technology Requirements

Of course, there are still plenty of technical challenges and also the issue of process step increase to sustain the cost scaling. Fundamentally, there exist several significant process flow issues from a production standpoint, such as process steps of capacitor formation, or high aspect ratio contact etches requiring photoresists with hard mask pattern transferring layer that can stand up for a prolonged etch time. Furthermore, continuous improvements in lithography/hard mask and etch will be needed. Also lower WL/BL resistance is necessary for getting the same or better performance.

Although 3-D type cell FETs like saddle-fin FETs are introduced and have revolutionized the one transistor-one capacitor (1T-1C) cell, it is getting more difficult to design due to the need to maintain a low level of both subthreshold leakage and junction leakage current to meet the retention time requirements. To optimize these operation windows in future devices, fully depleted type FET device (like a surrounded gate) will be needed to reduce the BL capacitance to get the sense margin. Another challenge is a highly reliable gate insulator. A highly boosted gate voltage is required to drive higher drain current with the relatively high threshold voltage adopted for the cell FET to suppress the subthreshold leakage current. The scaling of the DRAM cell FET dielectric, maximum word-line (WL) level, and the electric field in the cell FET dielectric are critical points for gate insulator reliability concern. To keep the electric field to a sustainable level in the dielectric with scaling, process requirements for DRAMs such as front-

end isolation, recess-FET formation, conformal oxidation process, gate filling process, and damageless recess process are all needed for future high-density DRAMs.

4.2 DRAM POTENTIAL SOLUTIONS

Since the DRAM storage capacitor gets physically smaller with scaling, the EOT must scale down sharply to maintain adequate storage capacitance. To scale the EOT, dielectric materials having high relative dielectric constant (κ) will be needed. Therefore MIM (metal-insulator-metal) capacitors have been adopted using high κ ($ZrO2/Al_2O/ZrO2$) [37] as the capacitor of 40-30's nm half-pitch DRAM. And this material evolution and improvement are continued until 20 nm HP and ultra high-K (perovskite K > 50 ~ 100) material will be released in 2016. Also, the physical thickness of the high-κ insulator should be scaled down to fit the minimum feature size. Due to that, capacitor 3-D structure will be changed from cylinder to pillar shape.

On the other hand, with the scaling of peripheral CMOS devices, a low-temperature process flow is required for process steps after formation of these devices. This is a challenge for DRAM cell processes which are typically constructed after the CMOS devices are formed, and therefore are limited to low-temperature processing. DRAM peripheral device requirement can relax I_{off} but demands more I_{on} of LSTP device. But, in the future, high-κ metal gate will be needed for sustaining the performance.

The other big topic is $4F^2$ cell migration. As the half-pitch scaling become very difficult, it is impossible to sustain the cost trend. The most promising way to keep the cost trend and increasing the total bit output by generation is changing the cell size factor (a) scaling (where a = [DRAM cell size]/[DRAM half pitch]^2). Currently $6F^2$ (a = 6) is the majority. To migrate $6F^2$ to $4F^2$ cell is very challenging. For example, vertical cell transistor must be needed but still a couple of challenges are remaining.

All in all, maintaining sufficient storage capacitance and adequate cell transistor performance are required to keep the retention time characteristic in the future. And their difficult requirements are increasing to continue the scaling of DRAM devices and to obtain the bigger product size (i.e. >16 Gb). In Figure PIDS5 the potential solutions are listed, but many future technologies will be necessary for 30 nm half-pitch or less. And these future technologies are still unknown.

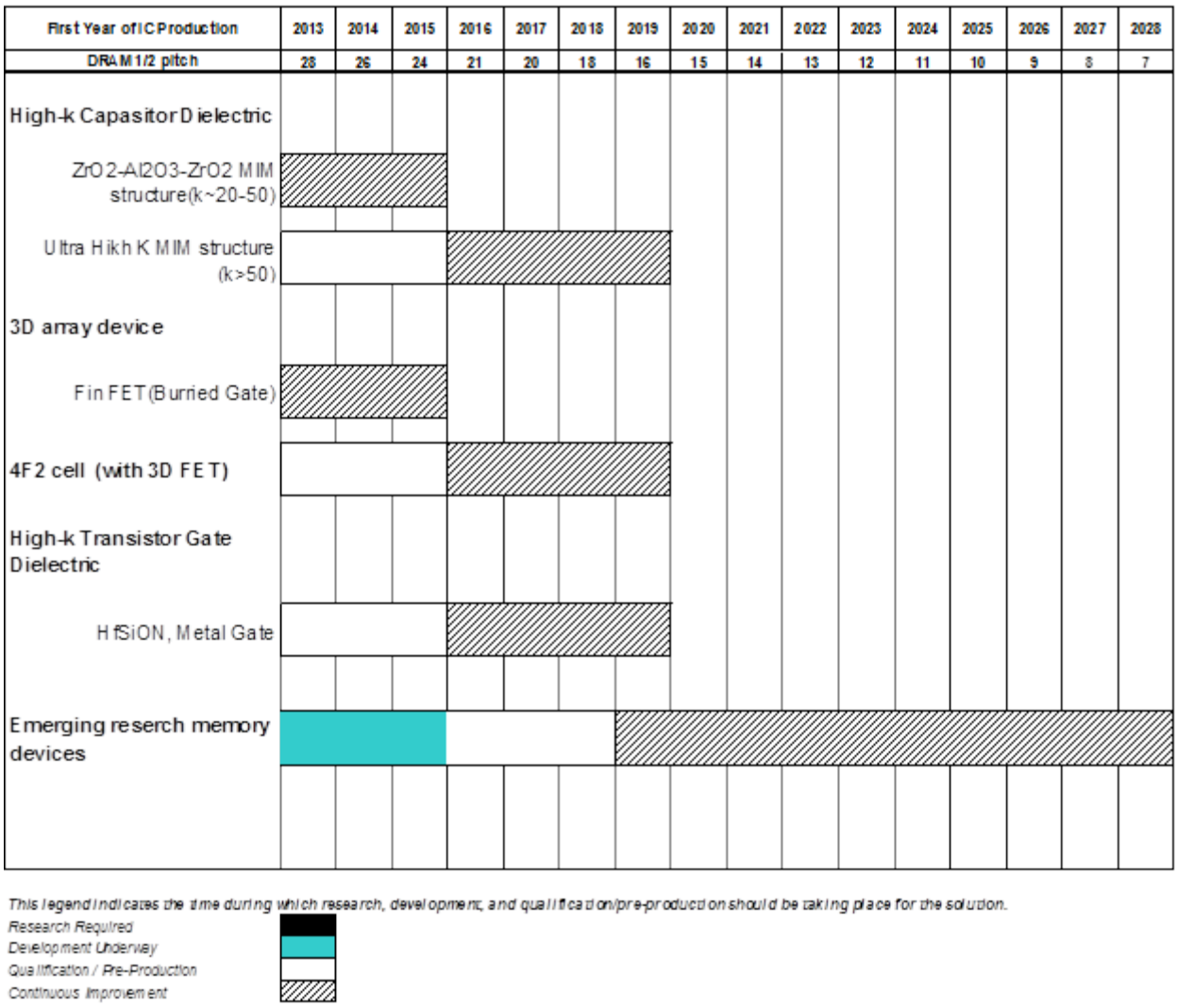


Figure PIDS5 DRAM Potential Solutions

5 NON-VOLATILE MEMORY

5.1 NON-VOLATILE MEMORY TECHNOLOGY REQUIREMENTS

Non-volatile memory consists of several intersecting technologies that share one common trait – non-volatility. The requirements and challenges differ according to the applications, ranging from RFIDs that only require Kb of storage to high-density storage of hundreds of Gb in a chip. The requirements tables are divided into two large categories—Flash memories (NAND Flash and NOR Flash), and non-charge-storage memories. Starting in 2013 resistive memory (ReRAM) (included in Table PIDS7b) is added as a potential solution. Non-volatile memories are essentially ubiquitous, and a lot of applications use embedded memories that typically do not require leading edge technology nodes. The requirements tables only track memory challenges and potential solutions for leading edge applications, thus most of the embedded applications are not included.

Flash memories are based on simple one transistor (1T) cells, where a transistor serves both as the access (or cell selection) device and the storage node. Several non-conventional non-volatile memories that are not based on charge storage (Ferroelectric or FeRAM, Magnetic or MRAM, phase-change or PCRAM, and resistive or ReRAM) form the category of often called "emerging" memories. These memory elements (the storage node) usually have a two-terminal structure (e.g. resistor or capacitor) thus do not serve as the cell selection device. The memory cell must include a separate access device in the form of 1T-1C, 1T-1R, or 1D-1R. A technology may be realized by more than one approach. For example, NOR Flash memories are fabricated using both floating gate device and nitride charge

trapping device. Although each may follow its own scaling trend, however, because they serve the same application market their scaling naturally converges.

Information on each technology is organized into three categories. The requirements tabulation for each technology first treats the issue of packing density. The applicable feature size "F" is identified and the expected area factor "a" is given (cell size in terms of the number of F^2 units required). Second, the tabulation presents a number of parameters important to each specific technology such as gate lengths, write-erase voltage maxima, key material parameters, etc. These parameters have significance because they are important to the scaling model and/or identify key challenge areas. Third, the endurance (erase-write cycle or read-write cycle) ratings and the retention ratings are presented. Endurance and retention are requirements unique to NVM technologies and determine whether the device has adequate utility to be of interest to an end customer.

Table PIDS7a shows technology requirements for NAND Flash and NOR Flash, and PIDS7b non-charge-storage memories for 2013 through 2028. The tables identify both the current CMOS half-pitch and the feature size actually used to form the NVM cells (i.e., the NVM technology "F" in nanometers). Rapid progress in NAND technology in recent years resulted in tighter half-pitches (uncontacted poly half-pitch) for NAND than those for DRAM and CMOS logic devices. In the longer term years, however, there is a significant slowing down of scaling because NAND Flash is facing scaling limitations, including the number of electrons per logic level and breakdown voltage between neighboring word lines. Eventually, NAND scaling falls behind logic devices. Some non-charge based memories seem to promise continued scaling and may eventually overtake NAND. Meanwhile, several approaches have been proposed to stack NAND cells to form 3D NAND arrays as a means to further increase NAND density with lower bit cost.

Table PIDS7a FLASH Technology Requirements

Table PIDS7b Non-charge-based Non-Volatile Memory (NVM) Technology Requirements

5.2 NON-VOLATILE MEMORY POTENTIAL SOLUTIONS

Nonvolatile memory (NVM) technologies combine CMOS peripheral circuitry with a memory array. The memory array generally requires additional, but CMOS compatible, processes to implement the non-volatility. Non-volatile memories are used in a wide range of applications, some standalone and some embedded, with varying requirements that depend on the application. The memory array architecture and signal sensing method also differ for different applications. The technical challenges are difficult, and in some cases fundamental physics limitations may be reached before the end of the current roadmap. For charge storage devices, the number of electrons in the storage node, whether for single level logic cells (SLC) or multi-level logic cells (MLC), needs to be sufficiently high to maintain stable threshold voltage against statistical fluctuation, and cross talk between neighboring bits must be reduced while the spacing between neighbors decreases. Meanwhile, data retention and cycling endurance requirements must be maintained, and in some cases even increased for new applications. Non-charge-storage devices also may face fundamental limitations when the storage volume becomes small such that random thermal noise starts to interfere with signal.

5.2.1 NAND FLASH MEMORY

5.2.1.1 FLOATING GATE NAND FLASH

Floating gate Flash devices achieve non-volatility by storing and sensing the charge stored "in" (on the surface of) a floating gate. The conventional memory transistor vertical stack consists of a refractory polysilicon or metal control gate, an interpoly dielectric (IPD) that usually consists of triple oxide-nitride-oxide (ONO) layers, a polysilicon floating gate, a tunnel dielectric, and the silicon substrate. The tunnel dielectric must be thin enough to allow charge transfer to the floating gate at reasonable voltage levels and thick enough to avoid charge loss when in read or off modes. The gate coupling ratio (GCR), defined as the capacitance ratio of the control gate to floating gate capacitor to the total floating gate capacitance (control gate to floating gate + floating gate to substrate), is a critical parameter for proper function (to ensure sufficient percent of voltage drops across the tunnel oxide during program and erase operations) of the device, and must be ≥ 0.6. In most structures, to achieve a GCR ≥ 0.6, the control gate (word line) needs to wrap around the sidewall of the floating gate to provide extra capacitance.

A NAND Flash cell consists of a single MOS transistor, serving both as the cell selection and as the storage device. The NAND array consists of bit line strings of now 64 devices or more with a string selection device at each end. This architecture requires no direct bit line contact to the cell, thus allows the smallest cell size ($4F^2$, or four features square). During programming or reading, the unselected cells in the selected bit line string must be turned on and serve as "pass" devices, thus the data stored in each device cannot be accessed randomly. Data input/output are structured in "page" mode where a page (on the Word line) is of several KB (now 8KB – 16KB) in size. Both programming and erasing are by Fowler-Nordheim tunneling of electrons into and out of the floating gate through the tunneling oxide. The low Fowler-Nordheim tunneling current allows the simultaneous programming of many bits (page), thus gives high programming throughput, suitable for handling large amount of data. Since devices in the same bit line string serve as pass transistors their leakage current does not seriously affect programming or reading operation (up to a limit), and without the need for hot electrons junctions can be shallow. Thus the scaling of NAND flash is not limited by device punch through and junction breakdown as in NOR flash. Designed to provide storage and access to large quantities of data but not to instantly execute program codes, NAND Flash generally employs error correction code (ECC) algorithms, and is thus more fault tolerant than NOR Flash. Because of fault tolerance thinner tunnel oxide may be used for NAND (than NOR flash) and this helps both scaling and more important, reducing the operation voltage, which is another scaling limiter.

The interpoly dielectric (IPD) thickness must scale with the tunnel dielectric to maintain adequate coupling of applied erase or write pulses to the tunnel dielectric. Because of data retention requirement, both tunnel dielectric and IPD scale slowly, however. In 2012, the most advanced NAND devices are fabricated by both the conventional wrap-around cell and the high-K/metal-gate planar cell with 1/2 pitch of 19nm – 20nm. The IPD for the wrap-around structure is ~ 10nm, compared to ~ 11nm in the 24nm 1/2 pitch device in 2010. It is difficult to achieve the wrap around structure when the bit line spacing becomes 20 nm or less, and slightly larger bit line pitch or slightly narrower floating gate (in the bit line direction) is used to accommodate the IPD yet still leave some room for the control gate to wrap around the floating gate to provide the necessary GCR. Further scaling is difficult for this conventional structure without new innovation.

To maintain a GCR > 0.6 and to avoid floating gate to floating gate cross talk are two difficult challenges when scaling below 20nm. Both can be alleviated by adopting high-K IPD and using a planar structure. Successful implementation of this new innovation in the 20nm and 16nm nodes recently gives hope to scale 2D NAND using a planar cell structure into the ~ 10nm regime. Although high-K also helps to reduce the program/erase voltage, the voltage reduction does not catch up with the rate of 1/2 pitch scaling, thus WL-WL electric field continues to increase and breakdown becomes a serious scaling limitation. Low-K dielectric is already not effective and air gaps between word lines are now adopted to improve the breakdown tolerance. Further scaling, however, still faces this very difficult challenge as the electric field increases at each new node.

Since the tunnel oxide scales very slowly, or not at all, the total EOT of the device is large and fringing field of the scaled device becomes less controlled (by the control gate) thus both degrades the device performance (larger subthreshold swing) and also increases the cell-to-cell interference. The number of storage electrons decreases linearly with the area of the device, in principle, and thus eventually will be too low and will cause unacceptable retention time distribution and severe random (telegraph) noise. Interestingly, when the fringing field dominates the number of electrons required to raise the V_{th} also becomes dominated by the fringing field and no longer decreases linearly with the device area. Thus the very difficult challenge of not having enough storage electrons may be not as formidable as previously feared. However, strong fringing field naturally causes disturb and other interference issues which are also difficult to solve.

(Planar) NAND Flash has now already scaled to 16nm node and further scaling to near 10nm seems possible. Beyond that, WL-WL breakdown, neighboring cell interference and statistical fluctuation of number of storage electrons must be overcome to further scale. 3D NAND and other emerging memories may further extend density.

5.2.1.2 CHARGE TRAPPING NAND FLASH

Currently all NAND products are fabricated with floating gate devices. The difficult challenges of maintaining or increasing the GCR and reducing the neighboring cell cross talk may be reduced by using charge trapping devices, but since rapid progress in planar HK/MG device has already alleviated these issues it is unlikely that 2D charge trapping devices will be adopted. Most 3D NAND devices, however, use charge trapping devices thus their principle and operation are described. No requirements table for charge trapping 2D NAND is prepared since there is no expectation of such products.

Charge trapping devices have only one single gate that controls the MOS device channel directly and thus there is no GCR issue, and the cross talk between thin nitride storage layers is either insignificant or at least much reduced. Nitride trapping devices may be implemented in a number of variations of a basic SONOS type device. SONOS using a simple tunnel oxide, however, is not suitable for NAND application—once electrons are trapped in deep SiN trap levels they are difficult to detrap even under high electric field. In order to erase the device quickly holes in the substrate are injected into the SiN to neutralize the electron charge. Since the hole barrier for SiO_2 is high (~4.1 eV), hole injection efficiency is poor and sufficient hole current is only achievable by using very thin tunnel oxide (~ 2 nm). Such thin tunnel oxide, however, results in poor data retention because direct hole tunneling from the substrate under the weak built-in field caused by storage electrons cannot be stopped. (The rate of direct tunneling is a strong function of the barrier thickness but only weakly depends on the electric field, thus the weak built-in field by charge storage is sufficient to cause direct hole tunneling from the substrate which ruins the data retention.)

Several variations of SONOS have been proposed. Tunnel dielectric engineering concepts are used to modify the tunneling barrier properties to create "variable thickness" tunnel dielectric. For example, triple ultra-thin (1–2 nm) layers of ONO are introduced to replace the single oxide (BE-SONOS) [38]. Under high electric field, the upper two layers of oxide and nitride are offset above the Si valence band, and substrate holes readily tunnel through the bottom thin oxide and inject into the thick SiN trapping layer above. In data storage mode, the weak electric field does not offset the triple layer and both electrons in the SiN and holes in the substrate are blocked by the total thickness of the triple layer. In MANOS (metal-Al_2O_3-nitride-oxide-Si) [39], a high-κ blocking dielectric and a high work function metal gate are combined to both prevent gate injection during erase operation, and to boost the electric field at tunnel oxide. A thicker (3–4 nm) tunnel oxide may be used to prevent substrate hole direct tunneling during retention mode.

Although charge trapping NAND can help the GCR and FG cross talk issues and thus promises scaling below 20nm it does not help the fundamental limitations such as word line breakdown and too few electrons. Therefore, in the roadmap trend it occupies a transition role between planar FG and 3D NAND. When charge trapping devices are used to build 3D NAND, the larger device size naturally solves the electron number and the word line breakdown issues.

5.2.1.3 NON-PLANAR AND MULTI-GATE DEVICES FOR NAND

Non-planar and multi-gate devices such as FinFET, double gate and surround-gate devices provide better channel control and allow further scaling of both floating gate and nitride trapping devices. However, the vertical structure also presents new challenges. For example, the space between fins must be sufficiently wide to allow room for tunnel oxide, floating gate and IPD (for floating gate device) and may forbid scaling beyond 20 nm if innovative solutions are not found. These are not included in the requirements tables.

However, these devices are used universally in 3D NAND because all 3D architectures naturally make these multi-gate devices the easiest to fabricate. In fact, it is much harder to use conventional single-gate device in 3D NAND.

5.2.1.4 3D NAND ARRAYS

When the number of stored electrons reaches statistical limits, even if devices can be further scaled and smaller cells achieved, the threshold voltage distribution of all devices in the memory array will become uncontrollable and logic states unpredictable. Thus memory density cannot be increased indefinitely by continued scaling of charge-based devices. However, density increase may continue by stacking memory layers vertically. Successful stacking of memory arrays vertically has been demonstrated in recent years. One approach uses single crystal Si by lateral epitaxial growth [40]. Another uses polycrystalline Si thin-film transistor (TFT) device [41]. The processing temperature and thermal budget must be such that the layers fabricated earlier are not degraded by the additional thermal processes. This imposes a significant challenge to either achieve identical devices in different layers that experience different thermal processes, or design circuits that can handle devices that are slightly different in each layer. Technical challenges aside, the economy of stacking complete devices is also questionable. As depicted in Figure PIDS6, the cost per bit starts to rise after stacking several layers of devices. Furthermore, the decrease in array efficiency due to increased interconnection and yield loss from complex processing may further reduce the cost-per-bit benefit of this type of 3D stacking.

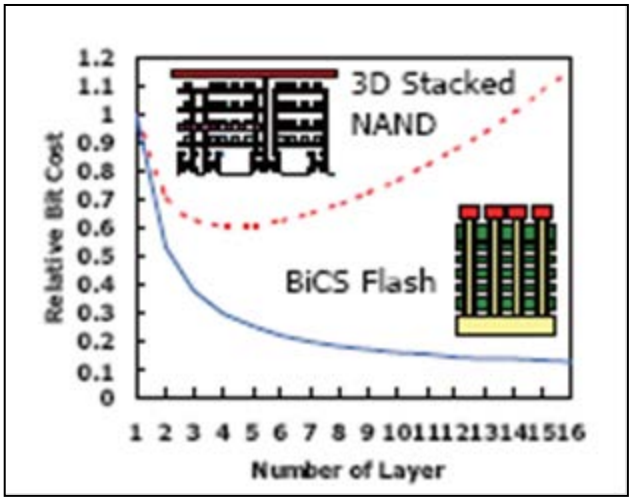


Figure PIDS6 Comparison of Bit Cost between Stacking of Layers of Completed NAND Devices and Making all Devices in Every Layer at Once [42]

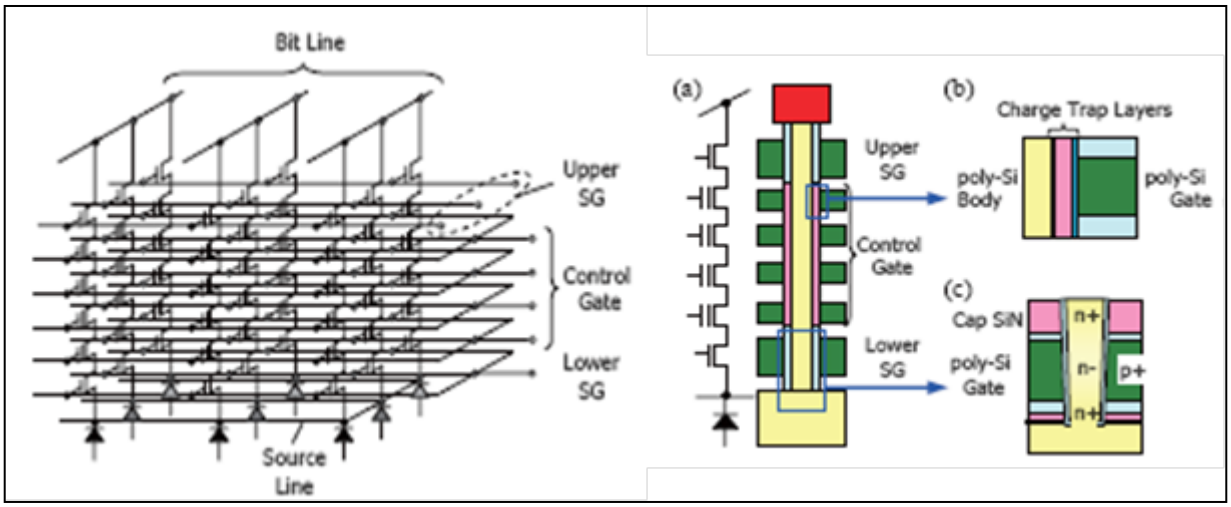


Figure PIDS7 (left) A 3D NAND Array based on a Vertical Channel Architecture [42]. (right) BiCS (Bit Cost Scalable) – a 3D NAND structure using a punch and plug process [42].

In 2007, a "punch and plug" approach is proposed to fabricate the bit line string vertically to simplify the processing steps dramatically [42]. This approach makes 3D stacked devices in a few steps and not through repetitive processing, thus promises a new low cost scaling path to NAND flash. Figure PIDS7 illustrates one such approach. Originally coined BiCS, or Bit Cost Scalable, this architecture turns the NAND string by 90 degrees from a horizontal position to vertical. The word line (WL) remains in the horizontal planes. As depicted in Figure PIDS6, this type of 3D approach is much more economical than the stacking of complete devices, and the cost benefit does not saturate up to quite high number of layers.

Various architectures for low cost 3D NAND have been proposed since BiCS, all employing the same principle of making all devices in a few simple operations [43]-[47]. These approaches may be put into three large categories: vertical channel (Figures PIDS7-PIDS9), vertical gate (Figures PIDS10-PIDS13), and floating gate (Figure PIDS14). In August of 2013 the first 3D NAND product using one of low cost approaches is introduced. All major NAND Flash suppliers have also announced plans to introduce 3D NAND products using various 3D architectures.

The basic architecture for vertical channel approaches is shown in Figure PIDS7 "left," and may be achieved by a number of different structures – BiCS (Bit Cost Scalable, Figure PIDS7 "right," [42]), P-BiCS (Pipe-shaped BiCS, Figure PIDS8 "left," [48]), and TCAT (Terabit Cell Array Transistor, Figure PIDS8 "right," [43]). BiCS is the original punch-and-plug proposal. Because of the difficulty in preserving the tunneling oxide integration when opening the channel contact, an improved version called piped-shaped BiCS is introduced, which eliminates the need for such etching. TCAT adopts a gate-last approach and thus is more suitable for devices using high-K/metal-gate for faster program/erase. VSAT (Vertical Stacking Array Transistor, Figure PIDS9 [44]) has a different architecture. It resembles a folded-up 2D NAND string, as shown in Figure PIDS9. All structures share a common feature that the transistor channel in the array is in the vertical direction. Their detailed working mechanisms can be found in the cited references.

The vertical gate architecture is shown in Figure PIDS10. The structure resembles putting blades of 2D NAND arrays vertically side by side. The three VG approaches shown in Figure PIDS11, Figure PIDS12, and Figure PIDS13 differ in decoding methods.

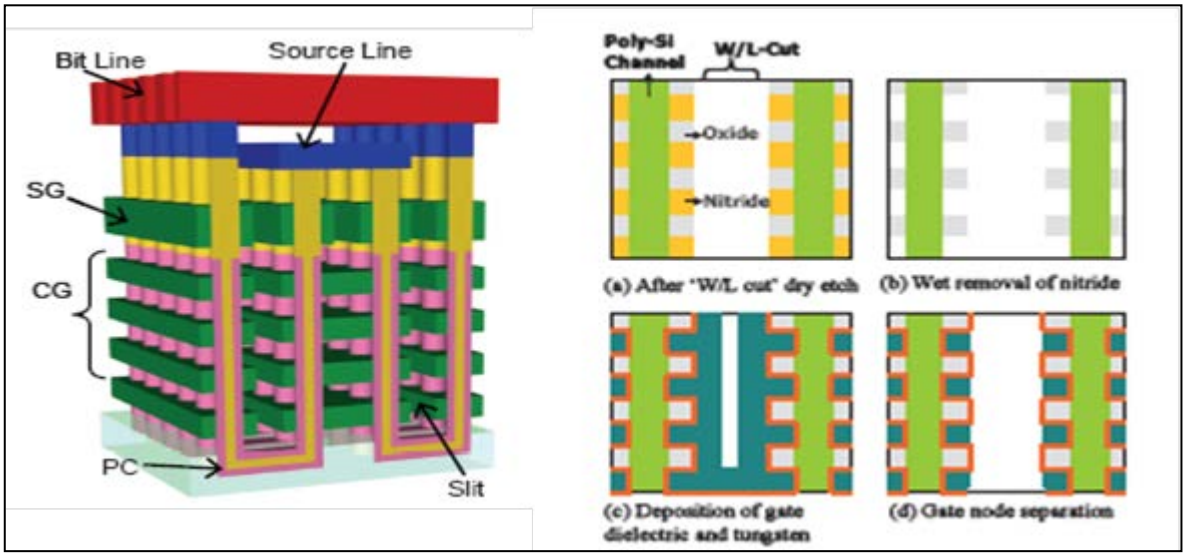


Figure PIDS8 (left) P-BiCS (Pipe-shaped BiCS) – An advanced form of BiCS 3D NAND array [48]. (right) TCAT (Terabit Array Transistor) – A gate last 3D NAND array [43].

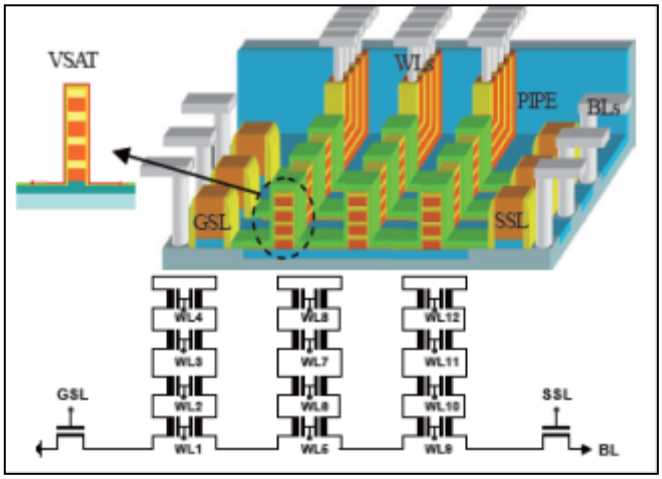


Figure PIDS9 VSAT (Vertical Stacking of Array Transistors) – Equivalent to folding up the horizontal bitline string vertically [44].

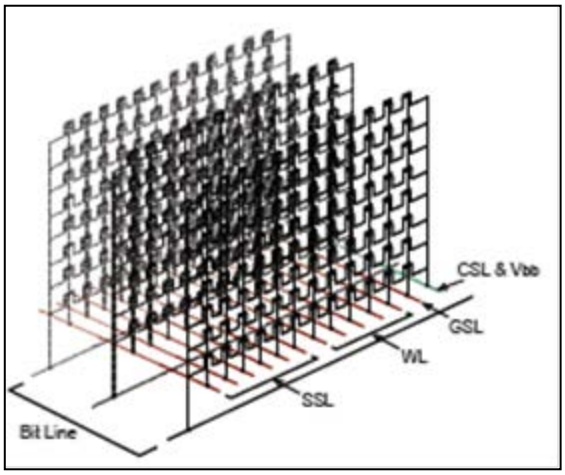
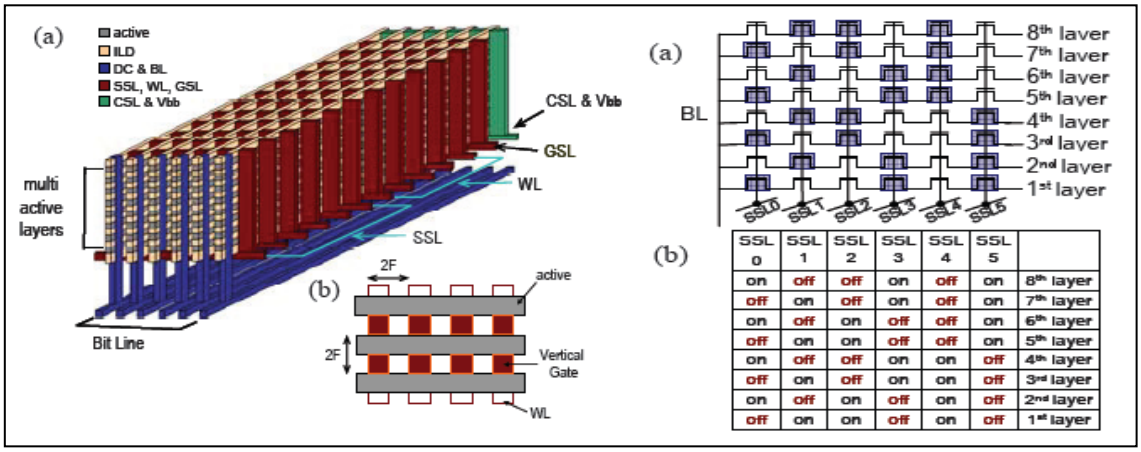


Figure PIDS10 (a) Vertical Gate 3D NAND Architecture. The bitline strings are in the horizontal direction as in the conventional 2D NAND. Each vertical "plane" of NAND devices is reminiscent to a 2D array [45].



SSL 0	SSL 1	SSL 2	SSL 3	SSL 4	SSL 5	
on	off	off	on	off	on	8th layer
off	on	off	on	off	on	7th layer
on	off	on	off	off	on	6th layer
off	on	on	off	off	on	5th layer
on	off	off	on	on	off	4th layer
off	on	off	on	on	off	3rd layer
on	off	on	off	on	off	2nd layer
off	on	on	off	on	off	1st layer

Figure PIDS11 A Vertical Gate 3D NAND Array with Decoding Method [45]

Although most 3D NAND structures use charge trapping (CT) devices, it is possible to construct a 3D NAND array using a floating gate device. Figure PIDS14 shows such a structure using a surround gate floating gate device [49]. Please refer to the cited reference for the construction process and other details.

Even though charge trapping devices do not need the gate coupling ratio to operate and thus may be planar, but once put in a 3D structure the geometric limitation that floating gate devices suffer from (filling IPD and poly WL between adjacent devices) now also applies. Instead of FG to FG cross talk, 3D NAND is subject to Z-direction interference [50]. The implication to cell size and scalability vary depending on the various 3D architectures. In general, vertical channel architectures have stronger geometric limitations, thus need more layers to achieve high density than the horizontal channel approaches, but are easier to fabricate. Therefore, the requirements table does not forecast a unique "node" or 1/2 pitch for all 3D structures. Various 3D approaches may be fabricated at different 1/2 pitches and with different layer numbers to achieve the same packing density. Thus the requirements table allows the choice of 1/2-pitch/layer-number that is suitable to a particular architecture.

3D structures achieve high density by increasing the layers and thus circumvent the few-electrons and word line breakdown limitations, thus the 1/2 pitch is not aggressively scaled. However, 3D structures have unique overhead costs that affect the array efficiency and in addition each layer may need to be contacted separately and that may incur additional processing cost. These may add substantially to the bit cost. Figure PIDS15 shows two schemes that may reduce the number of masks to make contacts [42], [43]. Even in the best case, however, there is a considerable overhead cost for making the 3D structure. If the 1/2 pitch for 3D is substantially relaxed compared to 2D NAND then the number of layers must be high enough to ensure high density and low bit cost. This is a trade-off that each 3D architecture will differ.

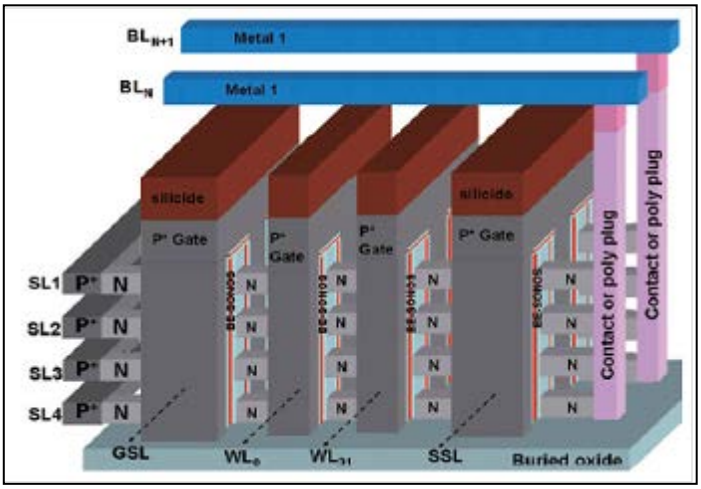


Figure PIDS12 Schematic Diagram of the PN Diode Decoded Vertical Gate (VG) 3D NAND Architecture. PN diodes are formed self-aligned at the source side of the VG NAND. Source lines (SL) of each memory layer are separately decoded, while WL, Bit line (BL), SSL and GSL are common vertically for the multi-layer stacks. Note that there is only one SSL and one GSL in one block [47].

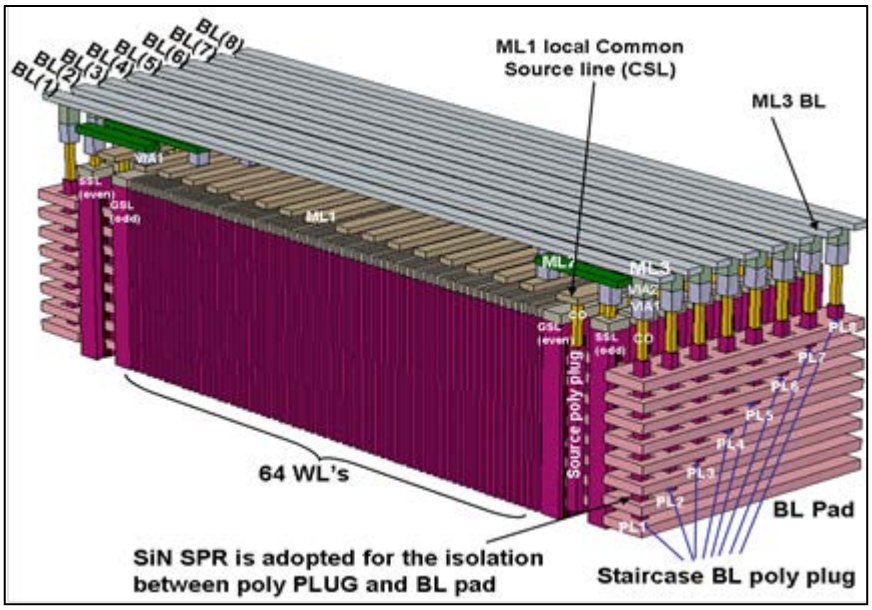


Figure PIDS13 Schematic Diagram of Island Gate SSL Decoded Vertical Gate 3D NAND. Each bit line is decoded by its own SSL, which is contacted through staircase contacts independently [47].

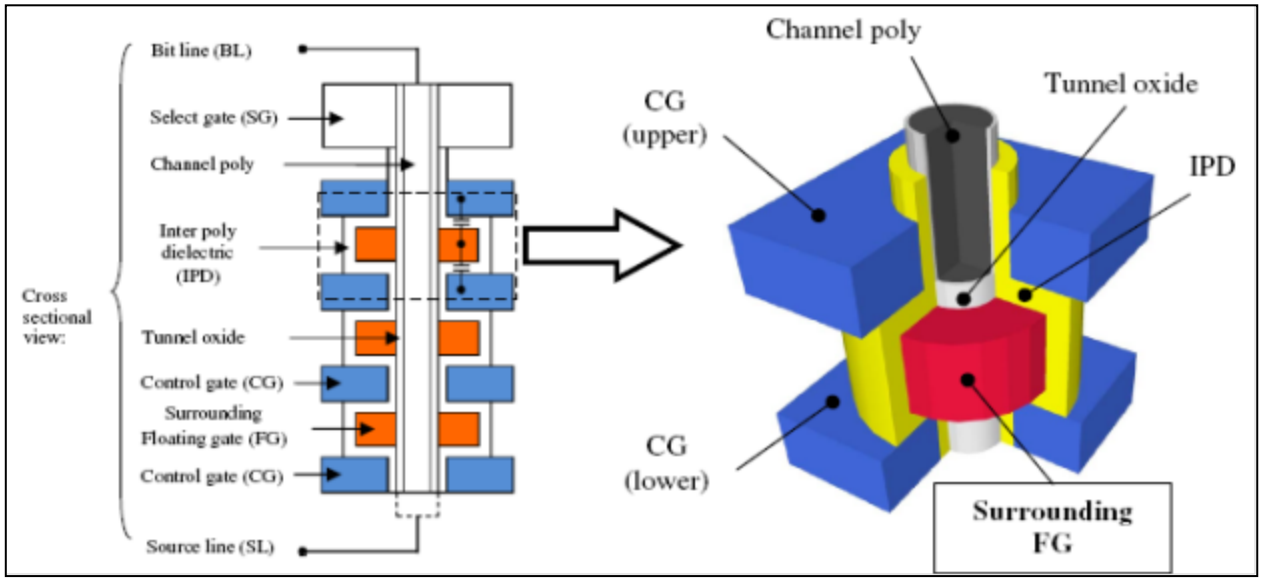


Figure PIDS14 A Surround Gate Floating Gate 3D NAND Structure

Since 3D NAND arrays use polysilicon TFT devices the Si area under the array is essentially unused. It is possible to design the circuits in such a way to put some peripheral circuits under the array, thus improve the array efficiency. The related interconnect layers then also need to be put under the array, and the TFT device fabrication is subject to some thermal restrictions. The cost-benefit of this approach must be analyzed since this peripheral-under-array approach has more complex processes.

Finally, it is important to clarify that 3D NAND is different from the often mentioned 3D integration of different chips using through silicon vias (TSV). It is also different from the common practice of stacking bare dies in one package to reduce the form factor. 3D NAND uses innovative structures and processes to make all layers of NAND devices at once, using few steps of lithography and etching. The stacking of NAND devices in 3D NAND is a description of the final structure, not to be confused with an action of stacking physical chips.

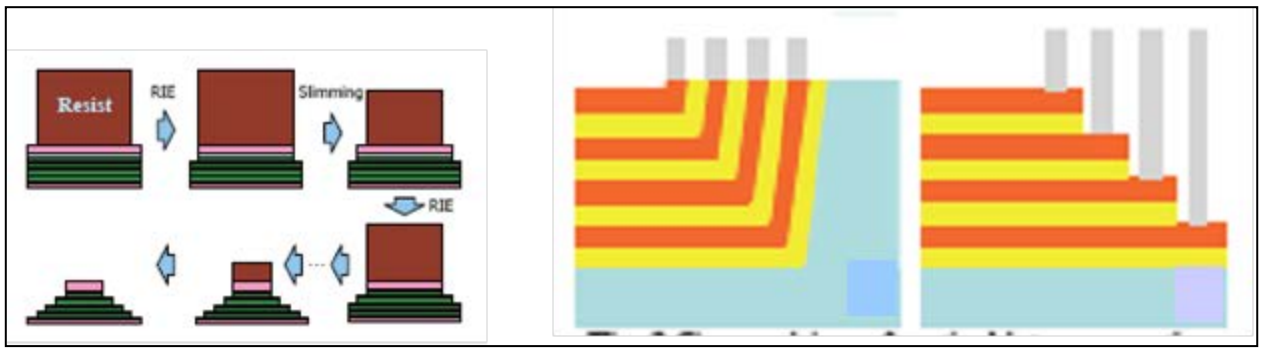


Figure PIDS15 (left) Scheme to make staircase landing pads for all layers by trimming one single layer of photoresist [42]. (right) A scheme to make contacts using tapered deposition and surface contact. Left: surface contacts are made in one operation. Right: conventional staircase contacts [44].

5.2.1.5 2D NAND TO 3D NAND TRANSITION

3D NAND is a logical extension of equivalent scaling after 2D NAND reaches its scaling limit, somewhere near 10nm node. This path, however, might not be followed by every device manufacturer for various reasons.

First, it is difficult to scale conventional FG NAND Flash to below 20nm node and still maintain 4F2 cell size. The control gate must wrap around the floating gate to produce enough capacitance for GCR, but there is not enough room between adjacent FG to accommodate twice the thickness of ONO IPD. Therefore, at below 20nm node either a planar structure using high-K IPD to get the GCR is adopted, or the cell starts to grow and density suffers.

Second, manufacturing of 2D NAND below 20nm nodes requires intensive investment in manufacturing tools for quadruple patterning. Much of this investment is not used when switching to 3D NAND later on since the pitch for 3D is considerably larger. Meanwhile, DRAM technology 1/2 pitch seems also stuck at > 20nm, thus investment in fine line patterning capacity may only have limited returns.

Third, even as 2D NAND can scale below 20nm node its reliability continues to suffer, most notably in write endurance. Since the tunnel oxide and IPD scaling falls far behind pitch scaling, NAND device suffers more and more from short channel effect, poor subthreshold swing, etc. Especially, when the number of electrons diminishes and cross talk increases the margin for fault tolerance becomes worse at each node. Meanwhile, important applications such as solid state drive (SSD) require better quality devices. 3D NAND, being made with longer channel devices, promises better performance.

Therefore, the choice between further scaling 2D and early introduction of 3D is not straightforward. It likely will be decided by each NAND manufacturer based on the circumstances best suited.

In the 2012 roadmap, introduction of 3D NAND was forecasted around 2016. In 2013, already some manufacturers have announced 3D introduction in 2014 (some in the end of 2013). Yet this does not imply that 3D will start to replace 2D in 2014. Rather, this reflects a divergence of product strategy from various manufacturers, probably due to the above reasons.

Thus the 2013 roadmap has pulled in the introduction of 3D NAND to 2014, but with a relatively large 1/2 pitch and relatively large number of 3D layers. In later years, the 1/2 pitch may be shrunk slightly and higher density parts may be made with the same, or even fewer layers.

5.2.2 NOR FLASH

5.2.2.1 FLOATING GATE NOR FLASH

A NOR Flash cell consists of a single MOS transistor serving both as the cell isolation device and the storage node. The threshold voltage of the transistor is modulated by charge stored in the floating gate and is used as an indication of the storage status. The storage cell may store single level logic (SLC, actually means bi-level logic 1 and 0) or multiple logic levels (MLC, e.g., (11), (10), (00), and (01)). The memory array is an X-Y cross wire structure, thus allowing random access of data. Programming is by channel hot electron or other variations of hot electron generation, and erasing is by Fowler-Nordheim tunneling of electrons out of the floating gate. The generation of hot electrons requires high lateral electric field under the device and is provided by a steep junction profile. This in turn causes short channel effect and leakage current that produce program disturb. Halo implants are used to control device leakage, and this subsequently reduces the junction breakdown voltage and limits the scaling capability.

The tunnel oxide thickness for the floating gate device poses a great scaling challenge because leakage through oxide thinner than about 8nm destroys retention, and there is no currently recognized solution. The short channel effect caused by thick tunnel oxide and the conflict between hot carrier generation and junction breakdown severely limit the outlook of NOR flash scaling below 32nm half-pitch.

Challenges in technology scaling, however, are not the reason why NOR Flash has essentially stopped evolving. At >10F2 cell size and technology nodes 2+ generations behind NAND, NOR Flash market in high density application, such as in feature phones, have been steadily eroded by NAND. In addition, some emerging memory technologies, such as phase change memory (PCM) can provide performance similar (and better) at somewhat lower cost. These market forces have considerable impact on FG NOR Flash, since some major suppliers have already switched to, or augmented by, PCM, which is more scalable and has a longer (and brighter) technology outlook.

However, NOR flash of all densities are widely used in numerous applications and thus even with the erosion in high end cell phone application the total NOR flash market seems still bright, especially in lower density serial I/O devices. The requirements table remains unchanged from 2011 projection, with current node at 45nm and slowly evolves to about 32nm. Beyond that the technology challenges are steep, but more important, alternative, more scalable NVM technologies (e.g. phase change memory) may prove to be more attractive.

5.2.2.2 CHARGE TRAPPING (CT) NOR FLASH

The threshold voltage of a device may also be affected by charges stored in a charge trapping layer, such as SiN. Charge trapping devices using a SiN as the trapping layer are usually called SONOS, since the device has a SONOS stack—a Si (polycide) gate, a blocking oxide, a nitride storage layer, and a tunnel oxide. The prevailing SONOS device using a relatively thick tunnel oxide in a NOR architecture is commonly known as NROM [51]. NROM uses channel hot electron for programming, and band-to-band tunneling of hot hole for erasing. Since charges injected into the nitride storage layer are well localized near the junctions two bits of information can be stored in the same device. The threshold voltage of the device can be read out by shielding the drain side bit with a drain bias and "reverse read" the source side information.

NROM NOR array can be implemented in a virtual ground architecture for which buried diffusion serves as the bit line and the device channel lies along the word line (polycide) direction. This structure requires neither bit line contact nor STI in the cell, thus offering a substantially smaller cell than the conventional floating gate NOR array. The cross talk between the two storage nodes in the same device cannot be completely eliminated. This so-called "second bit effect" restricts the threshold voltage window each storage node can carry, and the implementation of MLC in NROM poses a higher level of challenge than for floating gate devices. However, NROM is intrinsically 2-bit/cell and a 4-level MLC implementation results in 4-bit/cell, compared to 16-level MLC required for floating gate device for the same density. The virtual ground array offers a factor of 1.5× to 2× density advantage over conventional NOR architecture using the same design rules, and the single poly process reduces the mask layers.

Charge trapping devices do not have the gate coupling ratio issue floating gate devices face; however, the scaling challenges are otherwise quite similar. The virtual ground array and 2-bit/cell operation are sensitive to device leakage and the use of hot carriers for programming and especially the hot hole erasing increases the vulnerability to reliability failures. The scaling limitation is similar to floating gate NOR - leakage from short channel effect and junction breakdown. Without the severe limitation of tunnel oxide thickness its intrinsic scalability may be better, but the hot hole damage and the difficulty in virtual ground array largely offset this advantage. Therefore, the scaling trend in the requirements table stays the same as floating gate NOR flash.

5.2.3 NON-CHARGE-BASED NON-VOLATILE MEMORIES

Since the ultimate scaling limitation for charge storage devices is too few electrons, devices that provide memory states without electric charges are promising to scale further. Several non-charge-storage memories have been extensively studied and some commercialized, and each has its own merits and unique challenges. Some of these are uniquely suited for special applications and may follow a scaling path independent of NOR and NAND Flash. Some may eventually replace NOR or NAND flash. Logic states that do not depend on charge storage eventually also run into fundamental physics limits. For example, small storage volume may be vulnerable to random thermal noise, such as the case of super-paramagnetism limitation for MRAM.

One disadvantage of this category of devices is that the storage element itself cannot also serve as the memory selection (access) device because they are mostly two-terminal devices. Even if the On/Off ratio is high a memory cannot be constructed entirely by two terminal devices since the devices in the "On" state would form leakage paths. Therefore, these devices use 1T-1C (FeRAM), 1T-1R (MRAM, PCRAM and ReRAM) or 1D-1R (PCRAM and ReRAM) structures. It is thus challenging to achieve small ($4F^2$) cell size without innovative access device. In addition, because of the more complex cell structure that must include a separate access device, it is more difficult to design 3D arrays that can be fabricated using just a few additional masks like those proposed for 3D NAND.

5.2.3.1 FeRAM

FeRAM devices achieve non-volatility by switching and sensing the polarization state of a ferroelectric capacitor. To read the memory state the hysteresis loop of the ferroelectric capacitor must be traced and the stored datum is destroyed and must be written back after reading (destructive read, like DRAM). Because of this "destructive read," it is a challenge to find ferroelectric and electrode materials that provide both adequate change in polarization and the necessary stability over extended operating cycles. The ferroelectric materials are foreign to the normal complement of CMOS fabrication materials, and can be degraded by conventional CMOS processing conditions. Thus the ferroelectric materials, buffer materials, and process conditions are still being refined. So far the most advanced FeRAM [52] is substantially less dense than NOR and NAND Flash, fabricated at least one technology generation behind NOR and NAND Flash, and not capable of MLC. Thus the hope for near term replacement of NOR or NAND Flash has faded. However, FeRAM is fast, low power, and low voltage and thus is suitable for RFID, smart card, ID card, and other embedded applications. In order to achieve density goals with further scaling, the basic geometry of the

cell must be modified while maintaining the desired isolation. Recent progress in electrode materials shows promise to thin down the ferroelectric capacitor and extends the viability of 2D stacked capacitor through most of the near term years. Beyond this the need for 3D capacitor still poses steep challenges.

5.2.3.2 MRAM

MRAM devices employ a magnetic tunnel junction (MTJ) as the memory element. An MTJ cell consists of two ferromagnetic materials separated by a thin insulating layer that acts as a tunnel barrier. When the magnetic moment of one layer is switched to align with the other layer (or to oppose the direction of the other layer) the effective resistance for current flow through the MTJ changes. The magnitude of the tunneling current can be read to indicate whether a ONE or a ZERO is stored. Field switching MRAM probably is the closest to an ideal "universal memory" since it is non-volatile and fast and can be cycled indefinitely, thus may be used as NVM as well as SRAM and DRAM. However, producing magnetic field in an IC circuit is both difficult and inefficient. Nevertheless, field switching MTJ MRAM has successfully been made into products. In the near term, the challenge will be the achievement of adequate magnetic intensity H fields to accomplish switching in scaled cells, where electromigration limits the current density that can be used. Therefore, it is expected that field switch MTJ MRAM is unlikely to scale beyond 65nm node, and this is reflected in the requirements table (PIDS 7b).

Recent advances in "spin-torque transfer (STT)" (also referred to as spin-transfer torque) approach where a spin-polarized current transfers its angular momentum to the free magnetic layer and thus reverses its polarity without resorting to an external magnetic field offer a new potential solution [53].During the spin transfer process, substantial current passes through the MTJ tunnel layer and this stress may reduce the writing endurance. Upon further scaling the stability of the storage element is subject to thermal noise, thus perpendicular magnetization materials are projected to be needed at 32nm and below. New materials for perpendicular magnetization are still being researched, and are discussed in the ERM chapter. Perpendicular magnetization has been recently demonstrated [54].

With rapid progress of NAND Flash and the recent introduction of 3D NAND that promises to continue the equivalent scaling, the hope of STT-MRAM to replace NAND seems remote. However, its SRAM-like performance and much smaller footprint than the conventional 6T-SRAM have gained much interest in that application, especially in mobile devices which do not require high cycling endurance as in computation.

5.2.3.1 PCRAM

PCRAM devices use the resistivity difference between the amorphous and the crystalline states of chalcogenide glass (the most commonly used compound is $Ge_2Sb_2Te_5$, or GST) to store the logic ONE and logic ZERO levels. The device consists of a top electrode, the chalcogenide phase change layer, and a bottom electrode. The leakage path is cut off by an access transistor (or diode) in series with the phase change element. The phase change write operation consists of: (1) RESET, for which the chalcogenide glass is momentarily melted by a short electric pulse and then quickly quenched into amorphous solid with high resistivity, and (2) SET, for which a lower amplitude but longer pulse (usually >100ns) anneals the amorphous phase into low resistance crystalline state. The 1T-1R (or 1D-1R) cell is larger or smaller than NOR Flash, depending on whether MOSFET or BJT (or diode) is used, and the device may be programmed to any final state without erasing the previous state, thus provides substantially faster programming throughput. The simple resistor structure and the low voltage operation also make PCRAM attractive for embedded NVM applications. The major challenges for PCRAM are the high current (fraction of mA) required to reset the phase change element, and the relatively long set time. Since the volume of phase change material decreases rapidly with each technology generation, there is hope both above issues become easier with scaling. Interaction of phase change material with electrodes may pose long-term reliability issues and limit the cycling endurance and is a major challenge for DRAM-like applications. Because PCRAM does not need to operate in page mode (no need to erase) it is a true random access, bit alterable memory like DRAM.

The scalability of PCRAM device to < 5nm has been recently demonstrated using carbon nanotubes as electrodes [55]-[56], and the reset current followed the extrapolation line from larger devices. In at least one case, cycling endurance of 1E11 was demonstrated [57].

Phase change memory has been used in feature phones to replace NOR Flash since 2011, and has been in volume production at about 45nm node since 2012. NOR Flash, however, is unlikely the ultimate application for PCRAM. The performance advantage and scalability of PCRAM make it a better candidate for two important applications. One is storage class memory (SCM), which requires high-density, fast read/write and high endurance, which PCRAM is one of a few candidates that can satisfy. (Another candidate is ReRAM.) The other is a complementary high-density memory to DRAM. The limited cycling endurance and the smaller bandwidth (due to high current required for writing)

make PCRAM unsuitable to replace DRAM. However, it is otherwise similar to DRAM and its scalability may make it less expensive than DRAM in the future. And since PCRAM is nonvolatile it saves both the refreshing power, and more important, the dead time for refreshing which becomes increasingly a problem for DRAM. Therefore, a hybrid memory using small amount of DRAM and mostly PCRAM can be a low cost solution for high performance memory.

The evolution of PCRAM has followed the 2011 projection only moderately. 45nm PCRAM indeed went into production, but the subsequent higher-density (smaller F) products have not been introduced yet. However, the intermediate and long term projections for its applications are still the same. One significant change in the requirements table in 2013 is to expand the retention time range from 10 years to 0.83 to 10 years. The much shorter retention time reflects SCM type of applications that do not need long retention time.

5.2.3.4 Resistive memory - ReRAM

Beyond FeRAM, MRAM and PCRAM a large category of two-terminal resistive devices are being studied for memory applications. Many of these resistive memories are still in research stage and are discussed in more detail in the ERD/ERM chapters. Resistive memories are included in the PIDS chapter as a potential solution starting in 2013 because of their promise to scale below 10nm and the focused R&D efforts in many industrial labs make this technology widely considered a potential successor to NAND (including 3D NAND).

Although all resistive memories share a common trait of switching between (among) two or more resistive states they fall into many categories in the ERD/ERM chapter, based on their resistive switching mechanism and characteristics. (Please see ERD/ERM chapters for details.) In PIDS only two categories are discussed based on their potential applications, thus the descriptions are only functional in nature and type inclusions are not exhaustive. Materials and/or structures in other ERD/ERM categories may overlap with PIDS application-sorted categories.

Applications:

1. High-density non-volatile memory

A. High-density storage

Current roadmap forecasts NAND Flash memory will continue dominate high-density storage in the short and intermediate terms. 2D NAND may scale to at least ~ 15nm (1Ynm) node and quite likely to nearly 10nm (1Znm) node. 2D NAND scaling is ultimately limited by the small number of storage electrons per device (and even smaller per logic level) and the electric breakdown between neighboring word lines. (At 10nm node, the 10V across two neighboring word lines cause an electric field of approximately 10 MV/cm.) 3D NAND may start around 2014 in parallel with the continuing scaling of 2D NAND, and then continue the equivalent scaling by increasing the number of 3D layers. 3D NAND scaling is not limited by the number of storage electrons, nor the breakdown between adjacent word lines, thus seems unlimited. However, since the number of 3D layers must double at every node the difficulty in building the structure may eventually impose a practical limitation.

Assuming the next generation lithography will become available then an alternative scaling path for nonvolatile memory is to continue the pitch scaling instead of building up vertically. For example, if 2D NAND can continue to scale down to 5nm node then a 3D NAND at 20nm 1/2 pitch would require at least 16 layers to match the same bit cost. It is widely recognized that 2D NAND is unlikely to scale far below 10nm node but emerging memories such as phase change memory and resistive memories have demonstrated some potential to scale below 10nm. In addition, there are possibilities that 3D layers of PCRAM and/or ReRAM may be constructed. Thus, under optimistic scenarios 2D, and even 3D, PCRAM and ReRAM may continue the density scaling beyond 2D and 3D NAND capabilities. For example, in a very optimistic scenario, an MLC 4-layer 4nm 1/2 pitch cross-point ReRAM would provide an array cell density of > 12 Tb/cm2. 2D MLC NAND at 8nm (current minimum 1/2 pitch at end of the roadmap) can only provide about 1Tb/cm2, and 3D NAND with 20nm 1/2 pitch will need > 100 layers to reach the same (12 Tb/cm2) density. Therefore, the power of scaling, especially combined with (limited) capability to also build 3D structures, should not be underestimated.

Time of insertion is still quite uncertain since it depends on a number of factors. 2D NAND should stay dominant for as far as it can scale not only because it is a well-established technology but also since it has a very simple structure, requiring only one transistor and no contact within the cell. ReRAM cell needs an isolation device (cell selection device) to cut off sneak leakage paths as well as a contact in the cell since the array is DRAM like. In addition, many ReRAM's need bipolar operation and for those isolation cannot be done by a simple diode. Despite these challenges promising solutions have been proposed. Since 3D NAND will start (2014 – 2016) with 1/2 pitch well above 2D NAND, thus must be with quite high layer numbers (e.g. 16 layers or more) there is a possibility that 2D ReRAM can compete with 3D NAND with insertion time as early as 2016. Much will depend on how soon and how well the

ReRAM challenges are resolved, the maturity of next generation lithography, how far 2D NAND scales and how quickly 3D NAND can increase its layer number.

B. Storage Class Memory (SCM)

Instead of minimizing the storage cost, it may also be possible to exploit the fast-switching characteristics of ReRAM and its random access structure to improve the I/O bandwidth in a system. The function is now commonly referred to as storage class memory (SCM). The density requirements vary according to the function the SCM serves. For DRAM-like function the speed and endurance should be closer to DRAM and density needs not to be very high (SCM-M). For storage-like function the speed can be relaxed but the density should be comparable to SSD (SCM-S). SCM, no matter –M type or –S type, serves an intermediate role between DRAM and storage (SSD), providing fast and random access, high I/O bandwidth, and at least partial non-volatility to reduce refreshing power and refreshing dead time.

Note that phase change memory (PCRAM) and ReRAM share many common characteristics and their application space also overlap considerably. Both show potential to scale below 10nm. PCRAM is a unipolar device with relatively high switching current and moderate switching speed (tens of nanoseconds) that limits its writing bandwidth. ReRAM seems to promise higher bandwidth but the need for bipolar cell selection device imposes a severe challenge.

Insertion time for SCM application is also uncertain since it depends on the maturity of ReRAM and high-density PCRAM. This application, however, is a pent-up demand for next generation system need (servers and even mobile) to improve I/O speed (thus, system performance) and to reduce power that no other memory can satisfy.

2. Embedded NVM/Programmable Logic

There are currently many logic processes compatible embedded OTP/MTP's available, which satisfy many embedded NVM needs. Embedded Flash memory, which is important for microcontroller and automotive applications, however, still requires many additional masks and also faces scaling issues since it is based on NOR Flash. Phase change memory (PCRAM) and resistive memory, because of their simple (back end of line) MIM structure and low voltage operation are promising solutions. Phase change memory currently still has some operating temperature restrictions because of the low crystallization temperature of the GST-225 (GeSbTe-2:2:5) material that is most commonly used. Some resistive memory materials have shown higher temperature tolerance thus may be suitable for automotive and industrial applications.

The insertion of embedded NVM could be earlier than that for high-density storage because it does not face a strong incumbent (NAND). Automotive and industrial applications, however, require rigid reliability criteria that are often challenging for new technologies. Large volume consumer electronics could be the early implementations.

Some ReRAM's (especially the conduction bridge type, CBRAM) may be suitable for programmable logic application because of its good on/off ratio and logic process compatibility.

Technology:

1. Conduction bridge RAM (CBRAM)

Probably the most easily understood resistive memory is the conductive bridge RAM, for which the forming and destructing of a metallic bridge are induced by applying (and reversing) a small voltage across the two electrodes that encompass a solid electrolyte. The electrochemical process is similar to electroplating and electro-polishing where metal atoms are added or removed from the surface of an electrode. One or more metallic filaments can be grown from the cathode when a positive voltage is applied to the anode. Eventually, one filament would dominate since the tallest filament would experience the highest electric field which in turn would make it grow even faster.

There are several merits and drawbacks for the conducting bridge RAM (CBRAM). The switching mechanism is well understood and since it is an atomic process there seems no physical limitation to how small it can scale. The on/off ratio can be very high. However, there are some trade-offs between stability and programming power. There is a tendency of self-dissolution of the filament back into the electrolyte. If the filament is very thin it easily self-dissolves. To build more robust filament requires larger power and longer time. Once a metallic filament is formed it becomes the dominant current path and disrupts the electrolytic process, thus a filament does not just grow from thin to thick. The details of filament growth are still not well understood. Thus, although in principle only very small current is needed to build a filament, leakage current dominates and the electrolytic (ionic) current is only a small percentage of the total power in switching. Various buffer layers have been proposed to alleviate the self-dissolving of filament, but stability of very thin filament for extremely scaled (< 10nm) device has not been demonstrated yet.

2. Transition metal oxide (TMO)

Virtually every transition metal oxide exhibits some degree of R-V hysteresis which may be used to code different logic states thus serve as memory device. The most widely published TMO devices tend to use materials already familiar in the CMOS processing – TiOx, WOx, CuOx, NiOx, TaOx and HfOx. Since ReRAM is a backend of line (BEOL) process, the use of these familiar materials poses less contamination concern making potential insertion more likely.

Like CBRAM, most reported TMO ReRAM use bipolar operation – one polarity of pulse to switch the resistance from low to high (RESET) and the opposite polarity to switching from high to low (SET). The mechanism for resistance switching has not reached consensus yet, although forming and disrupting of conductive filament path through electrochemical migration of oxygen vacancy has received considerable support recently. In this theory, charged oxygen vacancies near the TMO/electrode interface are driven by the applied electric field toward or away from the interface, thus change the TMO composition to become more or less conducting. This is similar to the CBRAM mechanism except that the moving ion is oxygen (vacancy) instead of metal. TMO switching is also very fast, within a few nanoseconds, thus is essentially a diffusion-free process. Unlike the CBRAM for which a filament is built or disrupted the TMO activities are all very close to the interface. The electrode material has also been shown to greatly affect the TMO behavior. However, some unipolar operation characteristics obviously cannot be explained by electrochemical mechanisms. Mott insulator-metal transition has been proposed, in which a current triggered Mott transition is suggested for the insulator-metal transition.

Depending on the TMO materials and processes a wide range of performance has been reported. If it is possible to combine all the best merits of reported performance a ReRAM could have nanosecond switching time, high temperature tolerance to > 250C, and cycling endurance > 1E9. From practical point of view, it is reasonable to expect performance that is comparable to NOR Flash and PCRAM.

Many ReRAM reported the need of a "forming" step at the very beginning before normal switching occurs. The forming step requires higher voltage and seems to breakdown a non-conducting oxide. The details are still not clear. The near avalanche condition at the forming step, however, draws considerable current thus is a power/speed concern. There are also reports of forming-free TMO devices/processes.

Individual cells of < 10nm has been reported, and there has also been a number of reports on ReRAM array operation. TMO may be implemented with only one extra mask, using all CMOS compatible processes and materials.

Unipolar operation (i.e. use only one voltage polarity) has been demonstrated for some TMO devices. Unipolar operation is inconsistent with the prevailing electrochemical mechanism. Since unipolar operation all reported quite high current for the SET operation (HRS to LRS) some have proposed that it involves thermal processes. For some unipolar operation Mott insulator-metal transition has been proposed. When a voltage is applied the TMO (e.g. NiO) adjacent to the anode becomes deficient in electron and the originally metallic NiO goes through a Mott insulating transition to become an insulator. Later, when some higher voltage is applied, tunneling current becomes strong enough that the injected electrons overcompensate the electron deficiency in the NiO near the anode, resulting in an electron-rich state that triggers the Mott insulator-to-metal transition [58].

To summarize, TMO is a general category that contains many different materials that show considerably different switching characteristics that may be caused by quite different mechanisms.

Challenges

Although a number of successful array operations have been reported and a number of product development publicity announced, CBRAM and TMO based ReRAM have just started to introduce products, and no high-density product has been introduced yet. Array uniformity when scaling to smaller sizes and reliability are probably the most serious concerns.

For high-density applications one difficult challenge is how to achieve small cell size and what isolation device to use. For bipolar operation (which is the most common mode) the lack of a compact isolation (cell selection) device makes it hard to achieve cross-point, 4F2 cell size. Using a conventional MOSFET device easily expands the cell size to 8F2 (DRAM) or 10F2 (NOR Flash). Recently, a mixed ionic/electronic conducting device (MIEC) was introduced [59] to provide bipolar isolation with a Zener-diode type of I-V curve. The reliability of this new device still needs to be proven.

Although unipolar operation may use a p-n junction diode as a cell selection device, the V_{th} of the diode poses a problem. Both writing and reading need to go through the diode, and it is difficult to read at low voltage. The ratio of HRS and LRS must be high enough to not get buried below the diode resistance.

In order to provide fast read bandwidth, the read current has to be in reasonable range (100nA – 10uA). Similarly, programming current should be low, almost in the same range. This can introduce severe read disturb if the switching mechanism is indeed electrochemical.

Probably the most difficult challenge is competing with 2D and 3D NAND. As can be seen in the requirements table, the packing density (bits/cm2) of 2D ReRAM is substantially lower than 3D NAND from 2018 to 2020, and remains below 3D NAND even when the 1/2 pitch becomes substantially smaller. (But we need to keep in mind that the success of 3D NAND with > 100 layers is also not guaranteed.) If 3D ReRAM can be made (which is also uncertain) then the packing density can surpass 3D NAND substantially.

Challenges to 3D ReRAM

Unlike a transistor, a resistor has only two terminals. Although not impossible but it is intrinsically difficult to decode a 3D array of 2-terminal devices since the devices do not have three terminals that may conveniently correspond to (x, y, z) coordinates.

One innovative approach is to use a transistor in the bottom x-y plane to serve as an isolation device that controls a vertical string of 2-terminal ReRAM devices. Each ReRAM device has its own isolation device (cell selecting device) to cut off leakage paths to other nearby devices. However, even though the bottom transistor provides x-y decoding for the string of ReRAM cells it does not really provide isolation in other memory planes. As shown in Figure PIDS16, the word line in each z-plane is not a line but an entire plane [60]. The center pillar serves as one conductor (or electrode) and the ReRAM cell is lateral (in the x-y plane). The ring-shaped isolation (cell selecting) device must have large enough ON/OFF ratio to cut off the numerous leakage paths in the z-plane. Leakage in one cell not only disables the particular cell but also provides a leakage path to other planes (through the center pillar). Thus, the higher the number of layers and the larger the array block (2D), the higher ON/OFF ratio is needed.

Note that a diode-type device is unsuitable as the isolation device in the 3D array, not just because of the requirement for bipolar operation. A diode relies on a depletion region to sustain the reverse bias, and the depletion region is normally tens of nanometers thick. In extreme scaling of < 10nm 1/2 pitch there is no room for such large depletion region between neighbors. Thus, a compact (< 2nm) selecting device with high ON/OFF ratio and very high endurance remains one of the most difficult challenges for 3D ReRAM. Note also that the selecting (isolation) device gets cycled (turned on and off) during read operation, thus must withstand very high cycling endurance.

If unipolar operation of ReRAM is achieved then a diode may be used to serve as the cell selection device by putting the diode in the vertical direction, in series with the storage element, in a WL/BL cross point array. This type of repetitive stacking does not enjoy the cut-and-plug type of cost saving as in 3D NAND. However, according to Figure PIDS6, this type of stacking can cut the bit cost by ~2X with 4-6 stacking layers. The bit cost increases when stacking more layers since the processing cost and array overhead offset the benefit of increased packing density. A 2X decrease in bit cost can bring the cost to comparable and even lower than 3D NAND when ReRAM 1/2 pitch scales below 8nm.

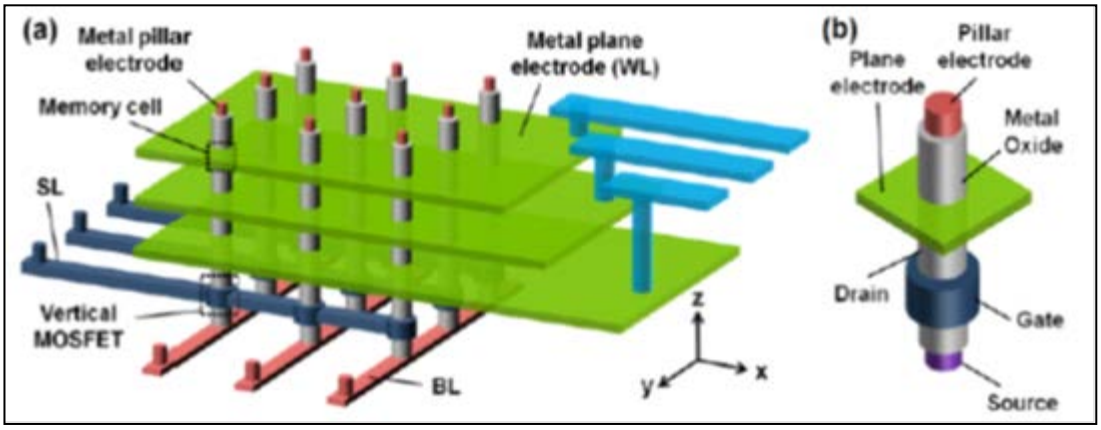


Figure PIDS16 Schematic view of (a) 3D cross-point architecture using a vertical RRAM cell and (b) a vertical MOSFET transistor as the bit-line selector to enable the random access capability of individual cells in the array [60].

Summary

Resistive memories change the MIM resistor conductivity by atomic processes, thus are not limited by the number of storage electrons. In principle, it should eventually also be limited by the number of atoms that provide the electrical characteristics. There is not enough understanding of the atomic details to project when this will limit the scaling of ReRAM. In the device level, < 10nm ReRAM has been reported. In the array level, 20nm 1Gb 2-layer 3D ReRAM has been published. At least one company has announced the introduction of products using ReRAM as an embedded memory. However, high-density ReRAM still must overcome several difficult challenges to be cost competitive to NAND.

Recent progress in 2D NAND to scale below 20nm (and promise to scale into 1Znm) and the introduction of 3D NAND have further compressed the space for ReRAM. Extreme scalability below 10nm and high ON/OFF ratio, bipolar, compact and high-endurance cell selection device are key challenges for high-density ReRAM.

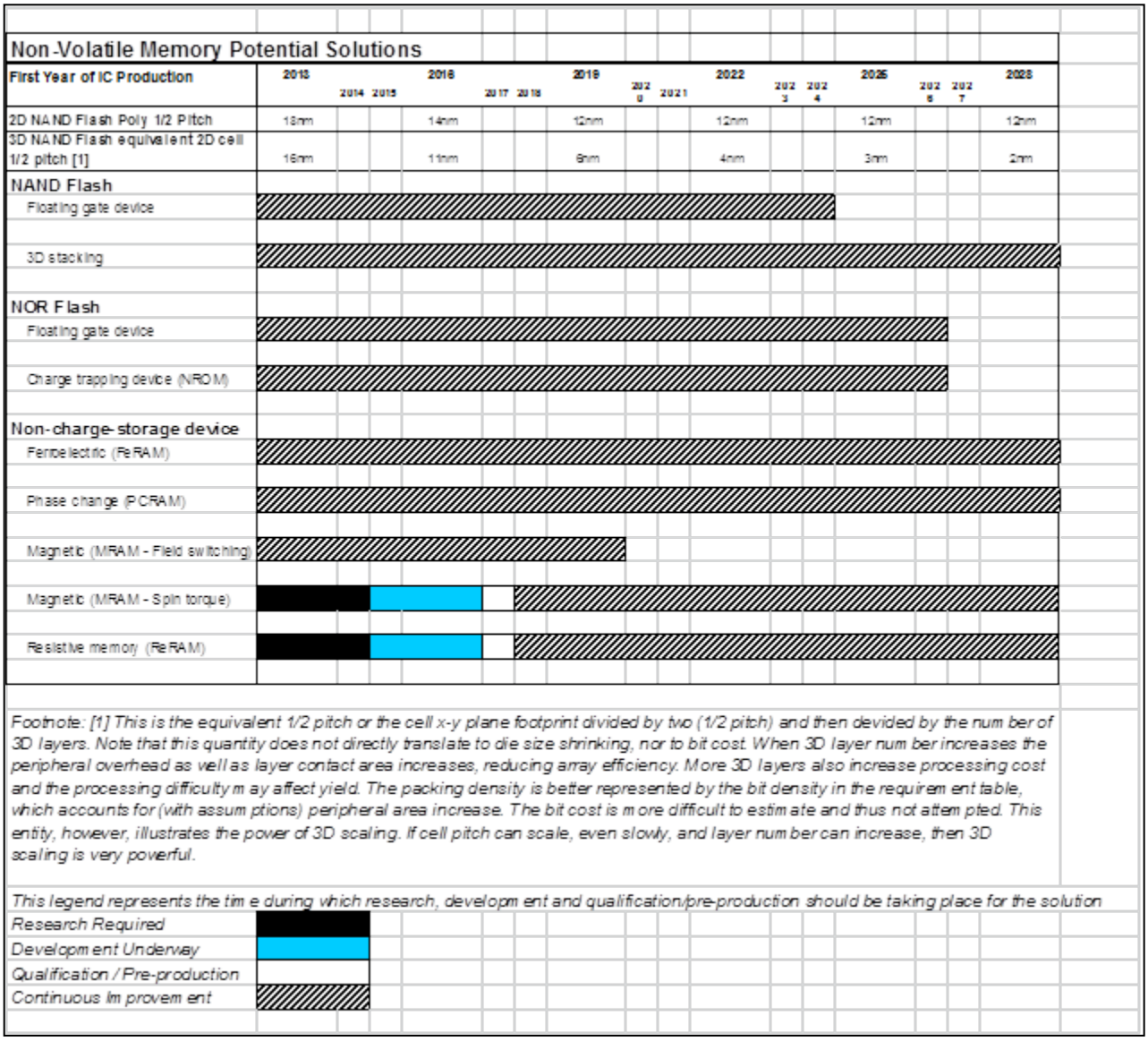


Figure PIDS17 Non-volatile Memory Solutions

6 RELIABILITY

Reliability is an important requirement for almost all users of integrated circuits. The challenge of realizing the required levels of reliability is increasing due to (1) scaling, the introduction of (2) new materials and devices, (3) more demanding mission profiles (higher temperatures, extreme lifetimes, high currents), and (4) increasing constraints of time and money.

1. Scaling produces ICs with more transistors and more interconnections, both on-chip and in the package. This leads to an increasing number of potential failure sites. Failure mechanisms are also impacted by scaling. For example, the time dependent dielectric breakdown (TDDB) of silicon oxy-nitride gate insulators has changed from electric-field-driven to voltage-driven as the insulator thickness has been scaled below 5 nm. In addition, negative bias temperature instability (NBTI) in p-channel devices, which used to be a minor effect when threshold voltages were larger, is now a great concern at the smaller threshold voltages of state-of-the-art devices. When the size of the transistor becomes comparable to or smaller than the values of the fundamental parameters such as mean-free-path of phonons and electrons, and de Broglie wavelength, familiar degradation mechanism may change and new ones may appear. For example, random telegraph noise

(RTN) has emerged as a serious reliability issue due to the smallness of the transistor. Another scaling induced new issue is gate to contact breakdown.

Increase in variability is expected as a result of scaling. Reliability mechanisms that are sensitive to device parameters will couple with the variability and be magnified, making reliability projection with limited number of measurements extremely difficult. For example, Vth variation can change the tunneling current and therefore affect all modes that are sensitive to gate current (BTI, TDDB, etc…). Channel length variation can affect lateral field dependent degradation such as HCI. Both channel length and gate overdrive affect the magnitude of RTN, and therefore their variation will also have effects. Large initial variation will make the observation of variability amplified degradation spread difficult without greatly increase the number of device tested.

Scaling may also lead to an effective increase of the stress factors. First, the current density is increasing and this increase impacts interconnect reliability. Second, voltages are often scaled down more slowly than dimensions, leading to increased electric fields that impact insulator reliability. Third, scaling has led to increasing power dissipation that result in higher chip temperatures, larger temperature cycles, and increased thermal gradients, all of which impact multiple failure mechanisms. The temperature effects are further aggravated by the reduced thermal conductivity that accompanies the reduction in the dielectric constant of the dielectrics between metal lines.

2. There are even more profound reliability challenges associated with revolutionary changes associated with new materials and new devices. Recognized failure mechanisms can change. New materials, such as high/low-K dielectrics or metal gates, and new device architectures, such as multi gate or FINFETs, can introduce new failure mechanisms or change the behavior of well-known failure mechanisms such as TDDB or BTI. Reliability evaluation is further complicated by the interaction between the materials in the gate stack being strongly affected by the process details (deposition techniques, thermal budget, etc.). Such complex multi-component gate stack structures may give rise to novel process-specific degradation mechanisms, both intrinsic and extrinsic. For example, with the transition from oxynitride/poly-Si gates to high-k/metal gates, positive bias temperature instability (PBTI) in n-channel devices appears presenting a more serious problem to the device stability. In addition, the nature of TDDB changes from progressive or multiple breakdowns, observed in poly-Si gate MOSFETs, to a more abrupt breakdown. The poor mechanical and thermal properties of low-k intermetal dielectrics can lead to mechanical failure mechanisms not seen in silicon dioxide intermetal dielectrics.

One of the routes to continue to the increase functionality of an IC is to integrate sensors and actuators on top of the CMOS platform. Such kind of "more than Moore" approach will greatly increase the complexity of reliability assurance. It is highly likely that such technology will come on line before the end of the road map and we must prepare for it. The likelihood that each sensor/actuator brings along a unique set of reliability problem is high and will present a whole new challenge to the reliability effort.

3. Mission profiles tend to be stretched further. For instance in sensor applications in automotive where temperatures exceeding 200C will be required, and in applications like base stations and solar cells, where (almost) continuous use during tens of years is required

4. Almost needless to say, but the ever increasing constraints of time and money in combination with possible major technology changes poses a real challenge for reliability engineering to keep in sync. Moreover the speed of introduction of these new materials and devices is exceeding our capability to build up learning on new failure mechanisms and physics, whereas the failure rate requirements are become more and more demanding. The impact of an unrecognized failure mechanism that makes it into end products would be significant.

These reliability challenges will be exacerbated by the need to introduce multiple major technology changes in a brief period of time. Interactions between changes can increase the difficulty of understanding and controlling failure modes. Furthermore, having to deal simultaneously with several major issues will tax limited reliability resources.

6.1 TOP RELIABILITY CHALLENGES

Table PIDS8 indicates the top near-term reliability challenges. It expands on the PIDS overall Difficult Challenge, “Timely assurance for the reliability of multiple and rapid material, process, and structural changes,” described at the beginning of this chapter.

The first near-term reliability challenge concerns failure mechanisms associated with the MOS transistor. The failure could be caused by either breakdown of the gate dielectric or threshold voltage change beyond the acceptable limits. The time to a first breakdown event is decreasing with scaling. This first event is often a “soft” breakdown. However,

depending on the circuit it may take more than one soft breakdown to produce an IC failure, or the circuit may function for longer time until the initial "soft" breakdown spot has progressed to a "hard" failure. Threshold voltage related failure is primarily associated with the negative bias temperature instability observed in p channel transistors in the inversion state. It has grown in importance as threshold voltages have been scaled down and as silicon oxy-nitride has replaced silicon dioxide as the gate insulator. Burn-in options to enhance reliability offend-products may be impacted, as it may accelerate NBTI shifts. Introduction of high-κ gate dielectric may impact both the insulator failure modes (e.g., breakdown and instability) as well as the transistor failure modes such as hot carrier effects, positive and negative bias temperature instability. The replacement of polysilicon with metal gates also impacts insulator reliability and raises new thermo-mechanical issues. The simultaneous introduction of high-κ and metal gate makes it even more difficult to determine reliability mechanisms. To put this change into perspective, even after decades of study, there are still issues with silicon dioxide reliability that need to be resolved.

As mentioned above, the move to copper and low κ has impacted front end reliability due to poorer thermal conductivity of low-κ dielectrics, leading to higher on-chip temperatures and higher localized thermal gradients.

ICs are used in a variety of different applications. There are some special applications for which reliability is especially challenging. First, there are the applications in which the environment subjects the ICs to stresses much greater than found in typical consumer or office applications. For example, automotive, military, and aerospace applications subject ICs to extremes in temperature and shock. In addition, aviation and space-based applications also have a more severe radiation environment. Furthermore, applications like base stations require IC's to be continuously on for tens of years at elevated temperatures, which makes accelerated testing of limited use. Second, there are important applications (e.g., implantable electronics, safety systems) for which the consequences of an IC failure are much greater than in mainstream IC applications. In general scaled-down ICs are less "robust" and this makes it harder to meet the reliability requirements of these special applications.

At the heart of reliability engineering is the fact that there is a distribution of lifetimes for each failure mechanism. With low failure rate requirements we are interested in the early-time range of the failure time distributions. There has been an increase in process variability with scaling (e.g., distribution of dopant atoms, CMP variations, and line-edge roughness). At the same time the size of a critical defect decreases with scaling. These trends will translate into an increased time spread of the failure distributions and, thus, a decreasing time to first failure. We need to develop reliability engineering software tools (e.g., screens, qualification, and reliability-aware design) that can handle the increase in variability of the device physical properties, and to implement rigorous statistical data analysis to quantify the uncertainties in reliability projections. The use of Weibull and log-normal statistics for analysis of breakdown reliability data is well established, however, the shrinking reliability margins require a more careful attention to statistical confidence bounds in order to quantify risk. This is complicated by the fact that new failure physics may lead to significant and important deviations from the traditional statistical distributions, making error analysis non-straightforward. Statistical analysis of other reliability data such as BTI and hot carrier degradation is not currently standardized in practice, but may be needed for accurate modeling of circuit failure rate.

The single long-term Reliability Difficult Challenge concerns novel, disruptive changes in devices, structures, materials, and applications. For example, at some point there will be a need to implement non-Copper interconnect (e.g., optical or, Carbon nanotube based interconnects), or tunnel-based FET's instead of classical MOSFET's. For such disruptive solutions there is at this moment little, if any, reliability knowledge (as least as far as their application in ICs is concerned). This will require significant efforts to investigate, model (both a statistical model of lifetime distributions and a physical model of how lifetime depends on stress, geometries, and materials), and apply the acquired knowledge (new building-in reliability, designing-in reliability, screens, and tests). It also seems likely that there will be less-than-historic amounts of time and money to develop these new reliability capabilities. Disruptive material or devices therefore lead to disruption in reliability capabilities and it will take considerable resources to develop those capabilities.

Table PIDS8 Reliability Challenges

Near-Term 2013-2020	***Summary of issues***
Reliability due to material, process, and structural changes, and novel applications.	TDDB, NBTI, PBTI, HCI, RTN in scaled and non-planar devices. Gate to contact breakdown. Increasing statistical variation of intrinsic failure mechanisms in scaled and non-planar devices. 3D interconnect reliability challenges. Reduced reliability margins drive need for improved understanding of reliability at circuit level. Reliability of embedded electronics in extreme or critical environments (medical, automotive, grid...).
Long-Term 2021-2028	***Summary of issues***
Reliability of novel devices, structures, and materials.	Understand and control the failure mechanisms associated with new materials and structures for both transistor and interconnect. Shift to system level reliability perspective with unreliable devices. Muon induced soft error rate.

6.2 RELIABILITY REQUIREMENTS

Reliability requirements are highly application dependent. For most customers, current overall chip reliability levels (including packaging reliability) need to be maintained over the next fifteen years in spite of the reliability risk inherent in massive technology changes. However, there are also niche markets that require reliability levels to improve. Applications that require higher reliability levels, harsher environments, and/or longer lifetimes are more difficult than the mainstream office and mobile applications. Note that a constant overall chip reliability levels requires a continuous improvement in the reliability per transistor because of scaling. Meeting reliability specifications is a critical customer requirement and failure to meet reliability requirements can be catastrophic.

These customer requirements flow down into requirements for manufacturers that rely on an in-depth knowledge of the physics of all the relevant failure modes and a powerful reliability engineering capability in design-for-reliability, building-in-reliability, and reliability qualification, defect screening and safe-launch methodologies to meet them. There are some significant gaps in these capabilities today. Furthermore, these gaps will become even larger with the introduction of new materials and new device structures. Inadequate reliability tools lead to unnecessary performance penalties and/or unnecessary risks.

Reliability qualification always involves some risk. There is a risk of qualifying a technology that does not, in fact, meet reliability requirements or a risk of rejecting a technology that does, in fact, meet requirements. At any point in time a qualification can be attempted on a new technology. However, the risk associated with that qualification can be large. The level of risk is directly related to the quality of the reliability physics and reliability engineering knowledge base and capabilities. To mitigate this risk, the concept of robustness validation needs to be exploited further. The combination of thorough failure mode knowledge, modeling and mission profile assessment is meant to minimize the probability of releasing technologies that have inherent wear-out issues, when properly employed it will lead to shorter qualification times, and lower risks.

The other challenge is that already in the product development and qualification phase, a low PPM level in the early part of the bathtub curve needs to be guaranteed. Samples sizes typically used for qualifications will never be able to supply enough statistics to support such guarantee.

The color-coding of the Reliability technology requirements is meant to represent the reliability risk associated with incomplete knowledge and tools for new materials and devices. The assumption is that there is no problem for the current year (2013) and that the next year is largely ready). The progression from yellow to striped indicates a growing reliability risk. The requirements first turn to yellow (Manufacturing Solutions are Known) in 2014 indicating a relatively small risk associated with scaling, increased power. The requirements then turn to striped (Interim Solutions

Known) in 2015 for nBTI and 2017 for pBTI. This date is approximate. The calculated Vmax is extracted from one published paper [61]. The result may or may not represent intrinsic values. Technology and therefore the gate stack quality vary from company to company, so would the Vmax. Thus when the calculated Vmax is smaller than the Vdd, the color is not red which represents no known solution to the problem. Only when the difference is significant red color is used. The risk assessment is, naturally, not very reliable for there are a number of known reliability issues that are still poorly understood. A case in point is the strong acceleration of NBTI in the presence of a drain bias, particularly for highly scaled devices. The assessment of moderate risk is a reflection of the awareness level of the problems. Solving these problems requires considerable effort and resources.

Also not included is the point in time where novel devices or materials are introduced (e.g., optical interconnect or a non-CMOS transistor or memory). As mentioned above these changes present a considerable reliability risk and require a considerable lead time to develop the needed capabilities in reliability physics and reliability engineering. Since we do not know exactly what these disruptive technologies will be and when they will be introduced, we have no way of knowing in advance the reliability risk. Solid red reflects the combination of increase variability, unknown reliability behavior from new materials and new structures, and the interaction between them. It signifies the greatly increased unknown rather than known issues that do not have known solution. The poorer the quality of our reliability knowledge is, the greater the reliability risks.

Table PIDS9 Reliability Technology Requirements

6.3 RELIABILITY POTENTIAL SOLUTIONS

The most effective way to meet requirements is to have complete built-in-reliability and design-for-reliability solutions available at the start of the development of each new technology generation. This would enable finding the optimum reliability/performance/power choice and would enable designing a manufacturing process that can consistently have adequate reliability. Unfortunately, there are serious gaps in these capabilities today and these gaps are likely to grow even larger in the future. The penalty will be an increasing risk of reliability problems and a reduced ability to push performance, cost and time-to-market.

It is commonly thought that the ultimate nanoscale device will have high degree of variation and high percentage of non-functional devices right from the start. This is viewed as an intrinsic nature of devices at the molecular scale. As a result it will not be possible any longer for designer to take into account a ‘worst case’ design window, because this would jeopardize the performance of the circuits too much. To deal with it, a complete paradigm change in circuit and system design will therefore be needed. While we are not there yet, the increase in variability is clearly already a reliability problem that is taxing the ability of most manufacturers. This is because variability degrades the accuracy of lifetime projection, forcing a dramatic increase in the number of devices tested. The coupling between variability and reliability is squeezing out the benefit of scaling. At some point, perhaps before the end of the roadmap, the cost of ensuring each and every one of the transistors in a large integrated circuit to function within specification may become too high to be practical. As a result, the fundamental philosophy of how to achieve product reliability may need to be changed. This concept is known as resilience, the ability to cope with stress and catastrophe. One potential solution would be to integrate so-called knobs and monitors in the circuits that are sensing circuit parts that are running out of performance and then during runtime can change the biasing of the circuits. Such solutions needs to be further explored and developed. Ultimately, circuits that can dynamically reconfigure itself to avoid failing and failed devices (or to change/improve functionality) will be needed.

Growing complexity of a reliability assessment due to proliferation of new materials, gate stack compositions tuned to a variety of specific applications, as well as shorter cycle for process development, may be alleviated to some degree by greater use of the physics-based microscopic reliability models, which are linked to material structure simulations and consider degradation processes on atomic level. Such models, a need for which is slowly getting wider recognition, will reduce our reliance on statistical approach, which is both expensive and time consuming, as discussed above. These models can provide additional advantage due to the fact that they can be incorporated in compact modeling tools with a relative ease and required only a limited calibration prior to being applied to a specific product.

Some small changes may already be underway quietly. A first step may be simply to fine-tune the reliability requirements to trim out the excess margin. Perhaps even have product specific reliability specifications. More sophisticated approaches involve fault-tolerant design, fault-tolerant architecture, and fault-tolerant systems. Research in this direction has increased substantially. However, the gap between device reliability and system reliability is very large. There is a strong need for device reliability investigation to address the impact on circuits. Recent increase in

using circuits such as SRAM and ring oscillator to look at many of the known device reliability issue is a good sign, as it addresses both the issues of circuit sensitivity as well as variability. More device reliability research is needed to address the circuit and perhaps system aspects. For example, most of the device reliability studies are based on quasi-DC measurements. There is no substantial research on the impact of degradation on devices at circuit operation speed. This gap in measurement speed make modeling the impact of device degradation on circuit performance difficult and risky.

In the meantime, we must meet the conventional reliability requirements. That means an in-depth understanding of the physics of each failure mechanism and the development of powerful and practical reliability engineering tools. Historically, it has taken many years (typically a decade) before the start of production for a new technology generation to develop the needed capabilities (R&D is conducted on characterizing failure modes, deriving validated, predictive models and developing design for reliability and reliability TCAD tools.) The ability to qualify technologies has improved, but there still are significant gaps.

There is a limit to how fast reliability capabilities can be developed, especially for major technology discontinuities such as alternate gate insulators or non-traditional devices. An eleventh-hour "sprint" to try and qualify major technology shifts will be highly problematical without the pre-existing and adequate reliability knowledge base.

For the reliability capabilities to catch up requires a substantial increase in reliability research-development-application and cleverness in acquiring the needed capabilities in much less than the historic time scales. Work is needed on rapid characterization techniques, validated models, and design tools for each failure mechanism. The impact of new materials like alternate channel material needs particular attention. Breakthroughs may be needed to develop design for reliability tools that can provide a high fidelity simulation of a large fraction of an IC in a reasonable time. As mentioned above, increased reliability resources also will be needed to handle the introduction of a large number of major technology changes in a brief period of time.

The needs are clearly many, but a specific one is the optimal reliability evaluation methodology, which would deliver relevant long-term degradation assessment while preventing excessive accelerated testing which may produce misleading results. This need is driven by the decreasing process margin and increasing variability, which greatly degrades the accuracy of lifetime projection from a standard sample size. The ability to stress a large number of devices simultaneously is highly desirable, particularly for long term reliability characterization. Doing it at manageable cost is a challenge that is very difficult to meet and becoming more so as we migrate to more advanced technology nodes. A break-through in testing technology is badly needed to address this problem.

7 CROSS-TWG ISSUES

7.1 FRONT END PROCESSES

There is strong linkage between the Front End Processes (FEP) and the PIDS chapters. Key areas of joint concerns include predicting introduction years of FD SOI and multi-gate structures. There are many parameters determined by process module capability that have significant influence on device characteristics. For example, for bulk devices, we face the difficult trade-offs of very high channel doping required to control short-channel effects. For fully depleted SOI and multi-gate MOSFETs, the key issue is controlling the required ultra-thin silicon body. All devices face the stringent requirement of source/drain series resistance, especially challenging with ultra-thin bodies. Another concern is V_{dd} scaling which affects almost all parameters, especially current drive, speed, EOT, and power density. For DRAMs, key areas of joint concern include implementation of metal-insulator-metal (MIM) storage capacitors with high-κ dielectric to scale the equivalent oxide thickness aggressively, as well as keeping the leakage of the access transistor ultra-low as the DRAM is scaled. For non-volatile memory, a key issue of joint concern involves the difficult trade-offs in scaling the interpoly and the tunneling dielectrics in FET flash memories.

Ideally, all parameters that are used common to both chapters should have the identical values. In reality we find it difficult to do a perfect job. The main reason is for PIDS, all parameters should be consistent with the over-all roadmap targets set by ORTC, such as device speed *I/CV*, gate length, V_{dd}, etc, as well as that from Design. All parameters also have to be self-consistent in MASTAR simulations. Secondly, in order to reconcile all parameters, there should be a few iterative cycles for each group to react and check for solutions in terms of both process capability and device performance. Understanding these challenges, there is plan for both groups to start the process earlier from now on to correct this short-coming.

7.2 DESIGN

The most immediate recipient of the outputs from PIDS is probably the Design TWG, so close interaction is a must. Most of the discussions surround the issues of speed and power requirements, and the trade-offs among them. The intrinsic transistor speed I/CV and its slope of increase per year is ultimately tied to the circuit clock frequency. This slope had been changed from 17%/year to the current value of 13%/year, and will likely be further reduced to 8%/year next year, and this had been a consensus opinion of many TWGs including Design. For low-power technologies (LP), the target metrics had largely come from Design. The overall requirements or guidelines of speed and power metrics among all logic technologies, summarized in Table PIDS5, are an example of the output of such interaction.

7.3 MODELING AND SIMULATION

Currently, PIDS uses physical parameters as inputs in MASTAR to calculate the device major characteristics, with certain assumptions in transport and electrostatics (subthreshold slope). Since MASTAR is based on analytical equations, even though it had been calibrated with device data, projection into the far future has some uncertainty. An approach to reduce the uncertainty in long range projection is to use TCAD tools, which rely on different assumptions and models to cross-check, and to determine some input parameters for MASTAR such as ballistic transport factor and subthreshold slope. Also in light of the newly introduced high-mobility channel materials of InGaAs and Ge, there are still a lot of uncertainties, such as the impact of low density of states in III-V. Close interaction and help from the Modeling and Simulation TWG has been most beneficial and needs to be continued. TCAD process simulation is also important to provide proper doping levels, defect transport and annihilation, contact interfacial properties, and geometries that can enhance accuracy of device simulation. Other long-term issues requiring enhanced modeling and simulation include atomic-level fluctuations, statistical process variations, and new interconnect schemes. With the shrinking of feature sizes, new process steps, architectures and materials reliability issues at the device, interconnect, and circuit levels will become even more important.

7.4 EMERGING RESEARCH DEVICES AND EMERGING RESEARCH MATERIALS

The Emerging Research Devices (ERD) chapter describes and evaluates potential technologies, including logic devices, memories, and architectures, beyond the current standard silicon CMOS technology. As such, it is concerned with the potential successor(s) to the CMOS described in the PIDS chapter. Toward or beyond the end of this roadmap period, when CMOS scaling will likely become ineffective and/or prohibitively costly, some version(s) of ERD technology will presumably be needed if the industry is to continue to enjoy rapid improvements in performance, lower power dissipation, lower cost per function, and higher functionality. Hence, the PIDS potential solutions tables for the late roadmap years include ERD solutions. Similarly, material-related topics come from the Emerging Research Materials (ERM) chapter.

8 REFERENCES

[1] M. Na et al., *IEDM Technical Digest*, p. 121, Dec. 2006.

[2] T. Skotnicki et al., "Innovative materials, devices and CMOS technologies for low-power mobile multimedia," *IEEE Trans. Electron Devices*, vol. 55, no. 1, pp. 96-130, January 2008.

[3] H. Mendez et al, "Comparing SOI and bulk FinFETs: Performance, manufacturing variability, and cost", *Solid State Technology*, Nov. 2009.

[4] T. Skotnicki, et al., "A new punchthrough current model based on the voltage-doping transformation," *IEEE Trans. Electron Devices*, vol. 35, no. 7, pp. 1076–1086, June 1988.

[5] T. Skotnicki et al., "A new analog/digital CAD model for sub-half micron MOSFETs," *IEDM Technical Digest*, pp. 165–168, December 1994.

[6] T. Skotnicki and F. Boeuf, "CMOS Technology Roadmap – Approaching Up-hill Specials," in *Proceedings of the 9th Int. Symp. On Silicon Materials Science and Technology*, Editors H.R. Huff, L. Fabry, S. Kishino, pp. 720–734, ECS. Vol. 2002-2.

[7] *http://nanohub.org/* last visit Jan 5th, 2014

[8] ITRS tool on nanohub.org (*https://nanohub.org/tools/itrs/*)

[9] R. Zhibin, R. Venugopal, S. Datta, M. Lundstrom, D. Jovanovic, J. Fossum, "The ballistic nanotransistor: a simulation study," *IEDM Tech. Dig.*, pp.715-718, 2000.

[10] M. Luisier, A. Schenk, W. Fichtner, G. Klimeck, "Atomistic simulation of nanowires in the sp3d5s* tight-binding formalism: From boundary conditions to strain calculations," *Phys. Rev. B.*, vol. 74, no. 20, pp. 205323 (2006).

[11] M. Luisier, G. Klimeck, "Atomistic full-band simulations of silicon nanowire transistors: Effects of electron-phonon scattering," *Phys. Rev. B.*, vol. 80, no. 15, pp. 155430 (2009).

[12] K. Banoo, M.S. Lundstrom, "Electron transport in a model Si transistor", *Solid-State Electronics*, Volume 44, Issue 9, 1 September 2000, Pages 1689-1695.

[13] R. Granzner, V.M. Polyakova, F. Schwierza, M. Kittlera, R.J. Luykenb, W. Rösnerb, M. Städele, "Simulation of nanoscale MOSFETs using modified drift-diffusion and hydrodynamic models and comparison with Monte Carlo results," *Microelectron. Eng.*, vol. 83, no. 2, pp. 241, 2006.

[14] M. R. Pinto, K. Smith and M. A. Alam, S. Clark, X. Wang, G. Klimeck, D. Vasileska, "Padre," *https://nanohub.org/resources/padre*. (DOI: 10.4231/D3RJ48T5G).

[15] X. Sun, X. Wang, Y. Sun, M. Lundstrom, "MIT Virtual-Source Tool," *https://nanohub.org/resources/vsmod*. (DOI: 10.4231/D3028PC40).

[16] M. Luisier, A. Schenk, W. Fichtner, and G. Klimeck, "Atomistic simulation of nanowires in the $sp^3d^5s^*$ tight-binding formalism: From boundary conditions to strain calculations," *Physical Review B*, vol. 74, no. 20, p. 205323, 2006.

[17] R. Kim and M. S. Lundstrom, "Physics of carrier backscattering in one-and two-dimensional nanotransistors,", *IEEE Transactions on Electron Devices*, vol. 56, no. 1, pp. 132-139, 2009.

[18] C. Jeong, D. A. Antoniadis, and M. S. Lundstrom, "On backscattering and mobility in nanoscale silicon mosfets," *IEEE Transactions on Electron Devices*, vol. 56, no. 11, pp. 2762-2769, 2009.

[19] M. Lundstrom, "Elementary scattering theory of the si mosfet," *IEEE Electron Device Letters*, vol. 18, no. 7, pp. 361-363, 1997.

[20] P. Palestri, D. Esseni, S. Eminente, C. Fiegna, E. Sangiorgi, and L. Selmi, "Understanding quasi-ballistic transport in nano-mosfets: part i-scattering in the channel and in the drain," *IEEE Transactions on Electron Devices*, vol. 52, no. 12, pp. 2727-2735, 2005.

[21] K. Natori, "Ballistic mosfet reproduces current-voltage characteristics of an experimental device," *IEEE Electron Device Letters*, vol. 23, no. 11, pp. 655-657, 2002.

[22] P. Palestri, R. Clerc, D. Esseni, L. Lucci, and L. Selmi, "Multi-subband-montecarlo investigation of the mean free path and of the kt layer in degenerated quasi ballistic nanomosfets," in *IEEE International Electron Devices Meeting, 2006*, pp. 1-4.

[23] A. Khakirooz, K. Cheng, A. Reznicek, T. Adam, N. Loubet, H. He, J. Kuss, J. Li, P. Kulkarni, S. Ponoth, et al., "Scalability of extremely thin soi (etsoi) mosfets to sub-20-nm gate length," *IEEE Electron Device Letters*, vol. 33, no. 2, pp. 149-151, 2012.

[24] T. Hiramoto, G. Tsutsui, K. Shimizu, and M. Kobayashi, "Transport in ultrathin-body soi and silicon nanowire mosfets," in *2007 IEEE International Semiconductor Device Research Symposium*, pp. 1-2.

[25] K. Uchida, J. Koga, and S.-i. Takagi, "Experimental study on electron mobility in ultrathin-body silicon-on-insulator metal-oxide-semiconductor field-effect transistors," *Journal of Applied Physics*, vol. 102, no. 7, p. 074510, 2007.

[26] K. Shimizu, G. Tsutsui, and T. Hiramoto, "Experimental study on mobility universality in (100) ultra thin body nmosfet with soi thickness of 5nm," in *IEEE International SOI Conference, 2006*, pp. 159-160.

[27] O. Faynot et al. FDSOI Workshop, October 15, 2009.

[28] T. Skotnicki, F. Arnaud and O. Faynot, "UTBB SOI – a wolf in sheep's clothing", pp. 72-79, *Future Fab International*, vol. 42, pp. 72-79, July 2012.

[29] M. Salmani-Jelodar, S. Kim, K. Ng and G. Klimeck, "Scaling Issues and Solutions for Double Gate MOSFETs at the end of ITRS," *ISDRS 2013***.**

[30] J. Lacord, G. Ghibaudo, and F. Boeuf, "Comprehensive and Accurate Parasitic Capacitance Models for Two- and Three-Dimensional CMOS Device Structures", *IEEE Trans. Electron Devices*, V. 59, No. 5, p. 1332, 2012.

[31] U. Avci and I. Young, "Heterojunction TFET Scaling and Resonant-TFET for Steep Subthreshold Slope at Sub-9nm Gate-Length", p. 96, *IEDM* 2013.

[32] A. Khan, C. Yeung, C. Hu, and S. Salahuddin, "Ferroelectric Negative Capacitance MOSFET: Capacitance Tuning and Antiferroelectric Operation", *IEDM Tech. Dig.*, p.255-258, 2011.

[33] 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", *Symp. VLSI Technology Digest of Technical Papers*, p.11, 2003.

[34] J. Y. Kim et al., "S-RCAT (sphere-shaped-recess-channel-array transistor) technology for 70nm DRAM feature size and beyond", *Symp. VLSI Technology Digest of Technical Papers*, p.34, 2005.

[35] Sung-Woong Chung et al., "Highly Scalable Saddle-Fin (S-Fin) Transistor for Sub-50 nm DRAM Technology", *Symp. VLSI Technology Digest of Technical Papers*, p.32, 2006.

[36] T. Schloesser et al., "$6F^2$ buried wordline DRAM cell for 40 nm and beyond", *IEDM Technical Digest*, p. 809, 2008.

[37] Deok-Sin Kil et al., "Development of New TiN/ZrO2/Al2O3/ZrO2/TiN Capacitors Extendable to 45nm Generation DRAMs Replacing HfO2 Based Dielectrics", *Symp. VLSI Technology Digest of Technical Papers*, p.38, 2006.

[38] H. T. Lue, S. Y. Wang, E. K. Lai, Y. H. Shih, S. C. Lai, L. W. Yang, K. C. Chen, J. Ku, K. Y. Hsieh, R. Liu, and C. Y. Lu, "BE-SONOS: A Bandgap Engineered SONOS with Excellent Performance and Reliability," in *Tech. Digest 2005 International Electron Devices Meeting*, pp. 547-550, 2005.

[39] Y. Shin, J. Choi, C. Kang, C. Lee, K.T. Park, J.S. Lee, J. Sel, V. Kim, B. Choi, J. Sim, D. Kim, H.J. Cho and K. Kim, "A Novel NAND-type MONOS Memory using 63nm Process Technology for Multi-Gigabit Flash EEPROMs," *Tech. Digest 2005 International Electron Devices Meeting*, pp. 337-340, 2005.

[40] S-M. Jung, J. Jang, W. Cho, H. Cho, J. Jeong, Y. Chang, J. Kim, Y. Rah, Y. Son, J. Park, M-S. Song, K-H. Kim, J-S. Lim and K. Kim, "Three Dimensionally Stacked NAND Flash Memory Technology Using Stacking Single Crystal Si Layers on ILD and TANOS Structure for Beyond 30nm Node," *Tech. Digest 2006 International Electron Devices Meeting*, pp. 37-40, 2006.

[41] E. K. Lai, H. T. Lue, Y. H. Hsiao, J. Y. Hsieh, C. P. Lu, S. Y. Wang, L. W. Yang, T. H. Yang, K. C. Chen, J. Gong, K. Y. Hsieh, R. Liu and C. Y. Lu, "A Multi-Layer Stackable Thin-Film Transistor (TFT) NAND-Type Flash Memory," *Tech. Digest 2006 International Electron Devices Meeting*, pp. 41-44, 2006.

[42] H. Tanaka, M. Kido, K. Yahashi, M. Oomura, R. Katsumata, M. Kito, Y. Fukuzumi, M. Sato, Y. Nagata, Y. Matsuoka, Y. Iwata, H. Aochi and A. Nitayama, "Bit Cost Scalable Technology with Punch and Plug Process for Ultra High Density Flash Memory," *Digest of Technical Papers, 2007 Symposium on VLSI Technology*, pp. 14-15, 2007.

[43] J. Jang, H.S. Kim, W. Cho, H. Cho, J. Kim, S.I. Shim, Y. Jang, J.H. Jeong, B.K. Son, D.W. Kim, K. Kim, J.J. Shim, J.S. Lim, K.H. Kim, S.Y. Yi, J.Y. Lim, D. Chung, H.C. Moon, S. Hwang, J.W. Lee, Y.H. Son, U.I. Chung, and W.S. Lee, "Vertical Cell Array using TCAT (Terabit Cell Array Transistor) Technology for Ultra High Density NAND Flash Memory," *Digest of Technical Papers, 2009 Symposium on VLSI Technology*, pp. 192-193, 2009.

[44] J. Kim, A.J. Hong, S. M. Kim, E.B. Song, J.H. Park, J. Han, S. Choi, D. Jang, J.T. Moon, and K.L. Wang, "Novel Vertical-Stacked-Array-Transistor (VSAT) for Ultra-high-density and Cost-effective NAND Flash Memory Devices and SSD (Solid State Drive)," *Digest of Technical Papers, 2009 Symposium on VLSI Technology*, pp. 186-187, 2009.

[45] W. Kim, S. Choi, J. Sung, T. Lee, C. Park, H. Ko, J. Jung, I. Yoo, and Y. Park, "Multi-layered Vertical Gate NAND Flash Overcoming Stacking Limit for Terabit Density Storage," *Digest of Technical Papers, 2009 Symposium on VLSI Technology*, pp. 188-189, 2009.

[46] C.H. Hung, H.T. Lue, K.P. Chang, C.P. Chen, Y.H. Hsiao, S.H. Chen, Y.H. Shih, K.Y. Hsieh, M. Yang, J. Lee, S.Y. Wang, T. Yang, K.C. Chen, and C.Y. Lu, "A Highly Scalable Vertical Gate (VG) 3D NAND with High Program Disturb Immunity using a Novel PN Diode Decoding Structure", *Digest of Technical Papers, 2011 Symposium on VLSI Technology*, 4B-1, 2011.

[47] S.H. Chen, H.T. Lue, Y.H. Shih, C.F. Chen, T.H. Hsu, Y.R. Chen, Y.H. Hsiao, S.C. Huang, K.. Chang, C.C. Hsieh, G.R. Lee, A. Chuang, C.W. Hu, C.J. Chiu, L.Y. Lin, H.J. Lee, F.N. Tsai, C.C. Yang, T.H. Yang, and C.Y. Lu, "A Highly Scalable 8-layer Vertical Gate 3D NAND with Split-page Bit Line Layout and Efficient Binary-sum MiLC (Minimal Incremental Layer Cost) Staircase Contacts", *Tech. Digest 2012 Electron Devices Meeting*, 2.3.1-2.3.4, 2012.

[48] R. Katsumata, M. Kito, Y. Fukuzumi, M. Kido, H. Tanaka, Y. Komori, M. Ishiduki, J. Matsunami, T. Fujiwara, Y. Nagata, L. Zhang, Y. Iwata, R. Kirisawa, H. Aochi and A. Nitayama, "Pipe-shaped BiCS Flash Memory with 16 Stacked Layers and Multi-Level-Cell Operation for Ultra High Density Storage Devices," *Digest of Technical Papers, 2009 Symposium on VLSI Technology*, pp. 136-137, 2009.

[49] S.J. Whang, K.H. Lee, D.C. Shin, B.Y. Kim, M.S. Kim, J.H. Bin, J.H. Han, S.J. Kim, B.M. Lee, Y.K. Jung, S.Y. Cho, C.H. Shin, H.S. Yoo, S.M. Choi, K. Hong, S. Aritome, S.K. Park, and S.J. Hong, "Novel 3-dimentional Dual Control-Gate with Surrounding Floating-Gate (DC-SF) NAND Flash Cell for 1Tb File Storage Application", *Tech. Digest 2010 International Electron Devices Meeting*, pp. 668-671, 2010.

[50] Y.H. Hsiao, H.T. Lue, T.H. Hsu, K.Y. Hsieh, and C.Y. Lu, "A Critical Examination of 3D Stackable NAND Flash Memory Architectures by Simulation Study of the Scaling Capability", *2010 International Memory Workshop*, pp. 142-145, 2010.

[51] B. Eitan, P. Pavan, I. Bloom, E. Aloni, A. Frommer, and D. Finzi, "NROM: A Novel Localized Trapping, 2 bit Nonvolatile Memory Cell," *IEEE Electron Device Letters*, **21**, pp. 543–545, Nov. (2000).

[52] Y. K. Hong, D. J. Jung, S. K. Kang, H. S. Kim, J. Y. Jung, H. K. Koh, J. H. Park, D. Y. Choi, S. E. Kim, W. S. Ann, Y. M. Kang, H. H. Kim, J.-H. Kim, W. U. Jung, E. S. Lee, S. Y. Lee, H. S. Jeong and K. Kim, "130 nm-technology, 0.25 μm^2, 1T1C FRAM Cell for SoC (System-on-a-Chip)-friendly Applications," *Digest of Technical Papers, 2007 Symposium on VLSI Technology*, pp. 230-231, 2007.

[53] K. Miura, T. Kawahara, R. Takemura, J. Hayakawa, S. Ikeda, H. Takahashi, H. Matsuoka and H. Ohno, "A novel SPRAM (SPin-transfer torque RAM) with a synthetic ferromagnetic free layer for higher immunity to read disturbance and reducing write-current dispersion," *Digest of Technical Papers, 2007 Symposium on VLSI Technology*, pp. 234-235, 2007.

[54] H. Noguchi, K. Kushida, K. Ikegami, K. Abe, E. Kitagawa, S. Kashiwada, C. Kamata, A. Kawasumi, H. Hara, S. Fujita, "A 250-MHz 256b-I/O 1-Mb STT-MRAM with advanced perpendicular MTJ based dual cell for nonvolatile magnetic caches to reduce active power of processors" *Digest of Tech. Papers, 2013 Symposium on VLSII Circuit*, pp. 108-109, 2013.

[55] F. Xiong, A. Liao, D. Estrada, and E. Pop, "Low-power Switching of Phase-Change Materials with Carbon Nano Tube Electrodes", published online in *Science Express,* March 10th, 2011.

[56] Liang, R.G.D. Jeyasingh, H-Y. Chen and H-S. P. Wong, "A 1.4uA Reset Current Phase Change Memory Cell with Integrated Carbon Nanotube Electrodes for Cross-Point Memory Application", *Digest of Technical papers, 2011 Symposium on VLSI Technology,* 5B-4, 2011.

[57] I.S. Kim, S.L. Cho, D.H. Im, E.H. Cho, D.H. Kim, G.H. Oh, D.H. Ahn, S.O. Park, S.W. Nam, J.T. Moon, and C.H. Chung, "High Performance PRAM Cell Scalable to Sub-20nm Technology with below 4F2 Cell Size, Extendable to DRAM Applications", *Digest of Technical papers, 2010 Symposium on VLSI Technology,* 19-3, 2010.

[58] K.H. Xue, C. A. Paz de Araujo, J. Celinska, C. McWilliams, "A non-filamentary model for unipolar switching transition metal oxide resistance random access memories", *J. Appl. Physics,* 109, issue 9, pp. 091602-091602-6, May 2011.

[59] R.S. Shenoy, K. Gopalakrishnan, B. Jackson, K. Virwani, G.W. Burr, C.T. Rettner, A. Padilla, D.S. Bethune, R.M. Shelby, A.J. Kellock, M. Breitwisch, E.A. Joseph, R. Dasaka, R.S. King, K. Nguyen, A.N. Bowers, M. Jurich, A.M. Friz, T. Topuria, P.M. Rice, B.N. Kurdi, "Endurance and scaling trends of novel access-devices for multi-layer crosspoint-memory based on mixed-ionic-electronic-conduction (MIEC) materials", *Digest of Tech. Papers, 2011 Symposium on VLSI Technology*, pp. 94-95, 2011.

[60] H.Y. Chen, S.M. Yu, B. Gao, P. Huang, J.F. Kang, H.-S.P. Wong, "HfOx Based Vertical Resistive Random Access Memory for Cost-Effective 3D Cross-Point Architecture without Cell Selector", *Tech. Digest 2012 International Electron Devices Meeting,* pp. 497-500, (20.7.1-20.7.4), 2012.

[61] Linder, B. P., E. Cartier, S. Krishnan, and E. Wu, "Improving and optimizing reliability in future technologies with high-κ dielectrics", Int. Symp. VLSI Technology, Systems, and Applications (VLSI-TSA), 2013.