INTERNATIONAL TECHNOLOGY ROADMAP FOR SEMICONDUCTORS

2011 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, and 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 high-κ gate dielectric, metal gate electrodes, strain enhancement, etc., and in the near future, new structures such as ultra-thin body fully depleted SOI, multi-gate (MG) MOSFETs (such as FinFETs), and alternate high-mobility channel materials. These innovations are expected to be introduced at a rapid pace, and hence understanding, modeling, and implementing them 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 excluded. The current mainstream NVM is flash. 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).

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 2011-2018 and long-term 2019-2026.

Table PIDS1 Process Integration Difficult Challenges

Near-Term 2011-2018	***Summary of Issues***
1. Scaling Si CMOS	Scaling planar bulk CMOS Implementation of fully depleted SOI and multi-gate (MG) structures Controlling source/drain series resistance within tolerable limits Further scaling of EOT with higher κ materials (κ > 30) Threshold voltage tuning and control with metal gate and high-κ stack Inducing adequate strain in new structures
2. Implementation of high-mobility CMOS channel materials	Basic issues same as Si devices listed above High-κ gate dielectrics and interface states (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

Table PIDS1	*Process Integration Difficult Challenges*
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
5. Reliability due to material, process, and structural changes, and novel applications.	TDDB, NBTI, PBTI, HCI, RTN in scaled and non-planar devices Electromigration and stress voiding in scaled interconnects Increasing statistical variation of intrinsic failure mechanisms in scaled and non-planar devices 3-D 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 2019-2026	***Summary of Issues***
1. Implementation of advanced multi-gate structures	Fabrication of advanced non-planar multi-gate 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) 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 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 2011-2018

[1] Scaling of Si CMOS—

Planar bulk CMOS devices compared to SOI and multi-gate structures have more difficulty in adequately controlling short-channel effects. Continued scaling will face significant challenges due to the high channel doping required to control short-channel effects and to set the threshold voltage properly, resulting in band-to-band tunneling across the junction, gate-induced drain leakage (GIDL), and degradation of carrier mobility. Furthermore, threshold voltage variation due to random (stochastic) dopant variation is projected to become more and more severe with short channel lengths.

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.

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. 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 to become worst with scaling.

Metal gate/high-κ 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-κ materials ($\kappa > 30$) 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-κ 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 stain 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 more than 10 years, and success has only started to appear only very recently. Nevertheless, there are still much work to be done in the areas of high-κ 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 non-volatile 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-κ 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 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. Scaling far below 20 nm is still a difficult challenge, however, because fringing-field effects and few-electron V_t noise margin are still not proven.

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.

[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-κ 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 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.

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, 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, 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 2019-2026

[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-κ gate dielectrics, metal gate electrodes, strained silicon channels, elevated source/drain, etc., are expected to

be incorporated. Body (fin) thicknesses below 5 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 may develop into more than 100 layers of stacked devices and cost effective implementation is 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. LOP technology is specifically designed to minimize the dynamic 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 LSTP 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, photophovoltaic, 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 reflect the MOSFET requirements of both high-performance (HP) and low-power 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 refers to chips for mobile systems, where the allowable power dissipation and hence the allowable off-currents are limited by battery life. There are two major categories within low-power; low operating (dynamic) power (LOP) and low standby (static) power (LSTP) logics. LOP chips are typically for relatively high-performance mobile applications, such as laptop computers, where the battery is likely to be of high capacity and the focus is on reduced operating power dissipation. LSTP chips are typically for lower-performance, lower-cost consumer type applications, such as consumer cellular telephones, with lower battery capacity and an emphasis on the lowest possible static power dissipation, i.e., the lowest possible leakage or off-current.

The transistors for high-performance ICs have both the highest performance and the highest leakage current of the three, and hence the physical gate length and all the other transistor dimensions are most aggressively scaled. This transistor typically constitutes a small minority of the transistors on a chip; it is used mainly in critical paths, while most of the transistors on the chip have higher threshold voltage and lower leakage current. This high-speed, high-leakage transistor tends to drive the technology. The transistors for LOP technology have the lowest V_{dd}, somewhat lower performance and off-current, while the transistors for LSTP chips have both the lowest speed performance and off-current (highest threshold voltage V_t) of the three.

The main indicator for low standby power is off-current, or source/drain leakage current $I_{sd,leak}$. It should be mentioned that other leakage currents coming 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 limitation. The main indicator for low dynamic power is CV^2. For this reason V_{dd} for LOP is the lowest. Starting from this year, CV^2 is added to the tables to monitor this quantity.

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 light effective masses are greatly beneficial in quasi-ballistic transport with enhanced thermal velocity and injection at the source end. In current view the materials of choice seem to be InGaAs for *n*-channel and germanium for *p*-channel. The higher performance will likely focus on delivering lower power for similar speed (*I/CV*) compared to the Si counterpart.

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 *CV/I* scaling, 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 17%/year. Recent surveys and literature indicate that the gate-length scaling has been less aggressive than the past. Similar trend of less rapid increase in circuit clock frequency had been observed at the same time. Realignment for this effect was the major change in the ITRS 2008 edition. Reiterating the change in 2008 in comparison to that in 2007, the physical gate length L_g scaling for HP logic is slowed down by 3-5 years, with a change of slope. The *I*/*CV* speed metric has a slope of ~13% increase per year instead of 17%.

The IC industry has begun to deploy architectural techniques such as multiple cores and multiple threads that exploit parallelism to improve the overall chip performance, and to enhance the chip functionality while maintaining chip power density and total chip power dissipation at a manageable level. With more than one central processing unit (CPU) core on chip, the cores can be clocked at a lower frequency while still getting better overall chip performance.

Thus, there is a trend for system designers to emphasize integration level, which enables multi-cores to be put on a chip, instead of raw transistor speed in optimizing system-level performance. In addition, system designers are sweeping ever more cache memory onto the processor chip in order to minimize the system performance penalty associated with finite-cache effects. As DRAM cells are significantly smaller than SRAM cells, another high-performance system technology trend is to integrate DRAM cells onto a processor chip for use in higher-level cache memory. With scaling, it is expected that these techniques will be more and more heavily exploited. In subsequent editions of the Roadmap, the Design and PIDS Working Groups will consider the impact of these architectural techniques, and in particular whether improved architectural parallelism may allow a slackening in the 13%/year transistor performance scaling target. Even though the same 13%/year slope is maintained for this 2011 edition, it is likely to be further reduced to 8% per year in the near future.

For generating the entries in the logic technology requirements tables, an MOSFET modeling software MASTAR was used.[2-4] The software contains detailed analytical MOSFET models that have been verified against literature data. It is well suited to efficiently analyzing technology tradeoffs for generating these tables, and has been used for the PIDS calculations for many years. For a given *CV/I* target, all input parameters are tentatively chosen based on scaling rules, engineering judgment, and physical device principles. These input parameters are iteratively varied until the target is met, and the final set of values for the input parameters is entered into the tables. *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 an analytical-based software, different than numerical-based TCAD programs. While it has the advantages of simplicity, the inputs are less fundamental compared to that of TCAD. Transport parameters are assigned as inputs to control the values of mobility and saturation velocity, and the degree of ballistic transport. In electrostatic control, for bulk devices the subtheshold slope is generated as output from MASTAR, but for fully depleted SOI and multi-gate structures, the subthreshold slope is assigned as another input parameter. To match the off-current as a pre-determined requirement, the gate work-function is varied until such off-current is met. The source/drain series resistance is another input parameter. In assigning its value, the ideal case without parasitic resistance is first calculated. The amount of resistance is then varied until the saturation current is reduced by 33-40%, depending on the year (linearly increasing over the range of 15 years).

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 that achieve the same target. For example, one technology would scale the EOT more aggressively by introducing high-κ 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 guide 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.

To reflect more accurately the transistor speed metric, added since the 2009 edition is the ring-oscillator speed, in delay per stage, for fan-outs of one and four. Ring-oscillator speed is slower than the intrinsic transistor speed, but is considered the fastest circuit speed that can be realized, and is a measurable parameter, so we feel it is a more suitable parameter to monitor a more realistic speed performance of a CMOS technology. For a CMOS inverter, the *p*-channel performance is also important but not captured in the past. In order to avoid having to double the table size from adding the *p*-channel MOSFET, only one parameter is entered—the ratio of $I_{d,sat}$ between the two types of channels. This is a reasonable compromise by assuming the capacitances associated with *p*-channel are similar, along with all other parameters such as threshold voltage and off-current. The inverter chain or ring-oscillator simulation is also conveniently performed by MASTAR. The *CV/I* metric is kept for continuity and comparison.

In each of these logic devices, multiple parallel paths in structures are sometimes followed. Planar bulk CMOS is extended as long as possible, while advanced CMOS technologies—ultra-thin body fully depleted (FD) silicon-on-insulator (SOI) MOSFETs and multi-gate (MG) MOSFETs (FinFETs) are implemented in later years, 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 difference substrates, so they do not affect the outcome of the performance prediction.[5] 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 can be scaled more effectively,

and hence are utilized later in the Roadmap. In fact, multi-gate MOSFET scaling is superior to FD SOI MOSFET scaling, and hence the ultimate MOSFET is projected to be the multi-gate device 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.

For the high-performance logic technology, as shown in Table PIDS2, the driver is the MOSFET intrinsic speed metric, $1/\tau$ or *I/CV*, although there is plan to switch to ring-oscillator speed eventually. Specifically, the target is an average 13% increase per year, which matches the rate of improvement in recent years. Meeting this target is an important enabler for the desired rate of improvement in the chip clock frequency. All the other parameter values in the table are chosen iteratively to meet this target, as explained above. Several important consequences of meeting this target are clear from the table. The *n*-MOSFET saturation drive current, $I_{d,sat}$, increases steadily over the course of the roadmap. The subthreshold source/drain leakage current, $I_{sd,leak}$, is fixed at a value of 100 nA/µm for all years, which has important consequences for the chip power dissipation (to be discussed below). Figure PIDS1 depicts the scaling of *I/CV* for high-performance logic. Overall, the 13%/year target is met. For planar bulk structure, the curve slopes increasingly downward from the 13%/year curve, mainly because of the scaling difficulties discussed in the Difficult Challenges section, Item 1. The scaling difficulties are also encountered in the MASTAR simulations, where the required channel doping increases sharply with year, to a very high value of 9×10^{18} cm^{-3} in 2017. For FD SOI, even though the pace is kept up with the 13% slope, the thin-body thickness requirement becomes extremely demanding, in the range of 3.5 nm in year 2019. This thin-body requirement is relaxed with the MG structure and scaling could continue until the end of this roadmap 2026.

Table PIDS2 High-performance (HP) Logic Technology Requirements

Figure PIDS1 also includes ring-oscillator speed which is defined as the reciprocal of the delay per stage, for both cases of fan-out of 1 and fan-out of 4. It is shown here that these frequencies are much slower than the transistor intrinsic frequency, as expected. For fan-out of one, the frequency ratio to intrinsic speed is about 5, whereas for fan-out of 4, the ratio is about 10. The slopes for both cases are also found to be much lower than that of *I/CV*, around 8%/year.

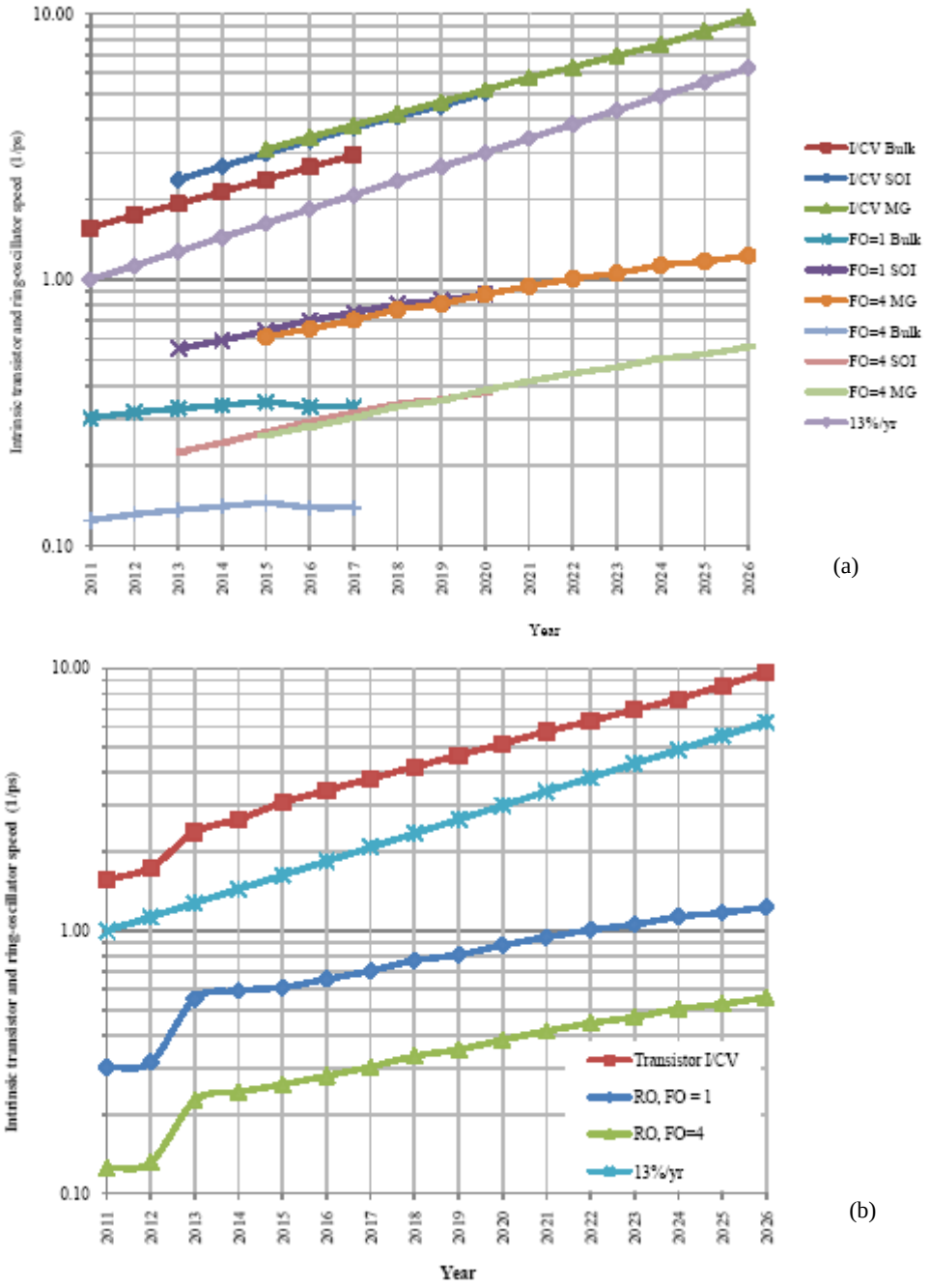


Figure PIDS1 (a) Scaling of Intrinsic Transistor Speed (I/CV) and Ring-Oscillator Speed for HP Technology. (b) Only the Highest Speed of Structures for Each Year is Shown.

For low-power chips, the important factor is the source/drain subthreshold leakage current, $I_{sd,leak}$ or off-current. For LOP logic (Table PIDS3), $I_{sd,leak}$ is set at 5 nA/µm, while it is 10 pA/µm for LSTP devices (Table PIDS4). All the other parameter values in the tables are chosen iteratively to meet the $I_{sd,leak}$ targets, while optimizing *I/CV*. A comparison of all logic technologies are presented in Figure PIDS2. The resultant speed improvement in the device performance metric, *I/CV*, is also around 13% improvement per year for both LOP and LSTP devices. Note that to meet the leakage current requirements, the gate length scaling of low-power logic lags behind that of high-performance logic. One key issue for LSTP logic is the slower scaling of V_{dd}. This is a result of the relatively slow scaling of the threshold voltage V_t required to meet the very low subthreshold leakage current targets. V_{dd} must follow V_t in scaling slowly because to obtain reasonable device performance, the gate overdrive, $(V_{dd} - V_t)$ must remain at a

reasonable level. Since dynamic power dissipation is proportional to V_{dd}^2, the dynamic power dissipation for the LSTP logic is not very different from that of HP. But since the activity factor for this type of logic is expected to be relatively small, the low static power dissipation more than compensates for the dynamic power. In contrast to LSTP logic, V_{dd} scales relatively quickly for LOP logic, where the focus is on minimizing the operating power (i.e., the dynamic power, proportional to V_{dd}^2).

Table PIDS3 *Low Operating Power (LOP) Technology Requirements*

Table PIDS4 *Low Standby Power (LSTP) Technology Requirements*

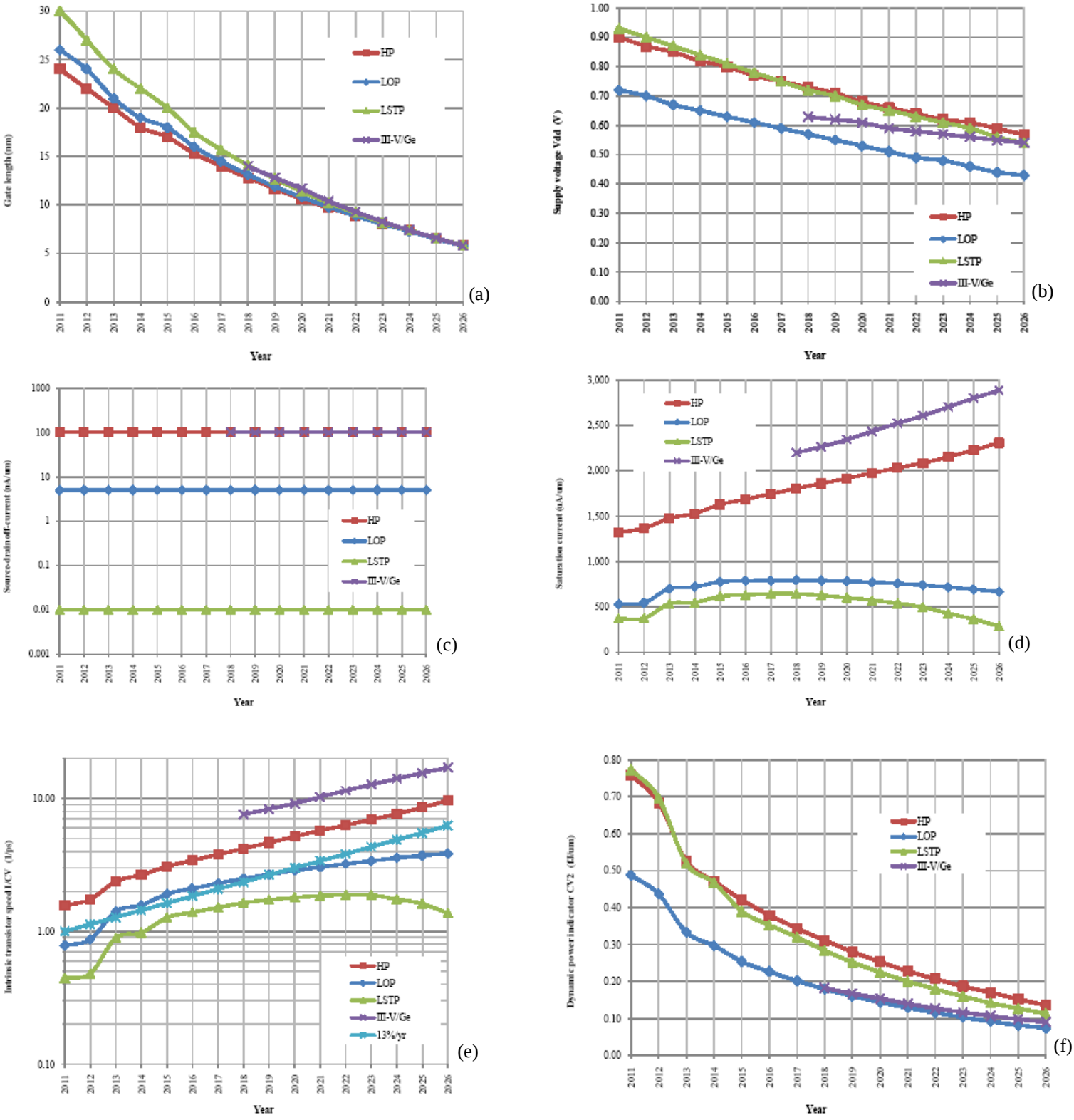


Figure PIDS2 Scaling Trend of Logic Technologies with Year; (A) Gate Length, (B) Supply Voltage, (C) Off-Current, (D) Saturation On-Current, (E) Intrinsic Speed (I/CV) Of Transistor, (F) Dynamic Power CV^2.

Starting from this year, we have included high-mobility channel III-V/Ge as a technology for both high speed and low dynamic power simultaneously (Table PIDS5). Such technology is anticipated to be in production in year 2018. The gate length is estimated to be lagging from that of HP by 1 year, since it involves a completely new set of materials. The main features are better seen also in Fig. PIDS2. Not only this technology offers higher speed than HP, but does so with the same low dynamic power as LOP.

Table PIDS5 *III-V/Ge High-performance Logic Technology Requirements*

Table PIDS6 *Comparison of HP, LOP, LSTP, and III-V/Ge Technologies*

	HP	LOP	LSTP	III-V/Ge
Speed (*I*/*CV*)	1	0.5	0.25	1.5
Dynamic power (CV^2)	1	0.6	1	0.6
Static power (I_{off})	1	$5x10^{-2}$	$1x10^{-4}$	1

Ultimately the trade-off between different logic technologies is speed vs. power, which is consisted of static power and dynamic power. A summary of all logic technologies for these metrics is captured in Table PIDS6. Here we only list the ratio in relation to the values in HP. It can be seen that between HP, LOP, and LSTP, there is trade-off of speed, static power, and dynamic because they are all Si-based technologies. Whereas for III-V/Ge, there is a net improvement since it is a completely different material system.

3.2 LOGIC POTENTIAL SOLUTIONS

There is a strong correlation between the challenges indicated by the colors in the technology requirements tables and the potential solutions (see Figure PIDS3). In many cases, red coloring (manufacturable solutions are not known) in the technology requirements 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 around two years in order to understand and deal with any new and different reliability, yield, and process integration issue associated with these innovative solutions. Many of the potential solutions may be required first for high-performance logic, followed by the low-power options. Finally, the industry faces a major overall challenge due to the sheer number of major technological innovations required over the next five years: enhanced mobility [6] and high-field transport, high-κ/metal gate stack (which are already implemented but requiring continuous improvement with scaling), ultra-thin body fully depleted SOI, and multi-gate MOSFETs, with quasi-ballistic transport.

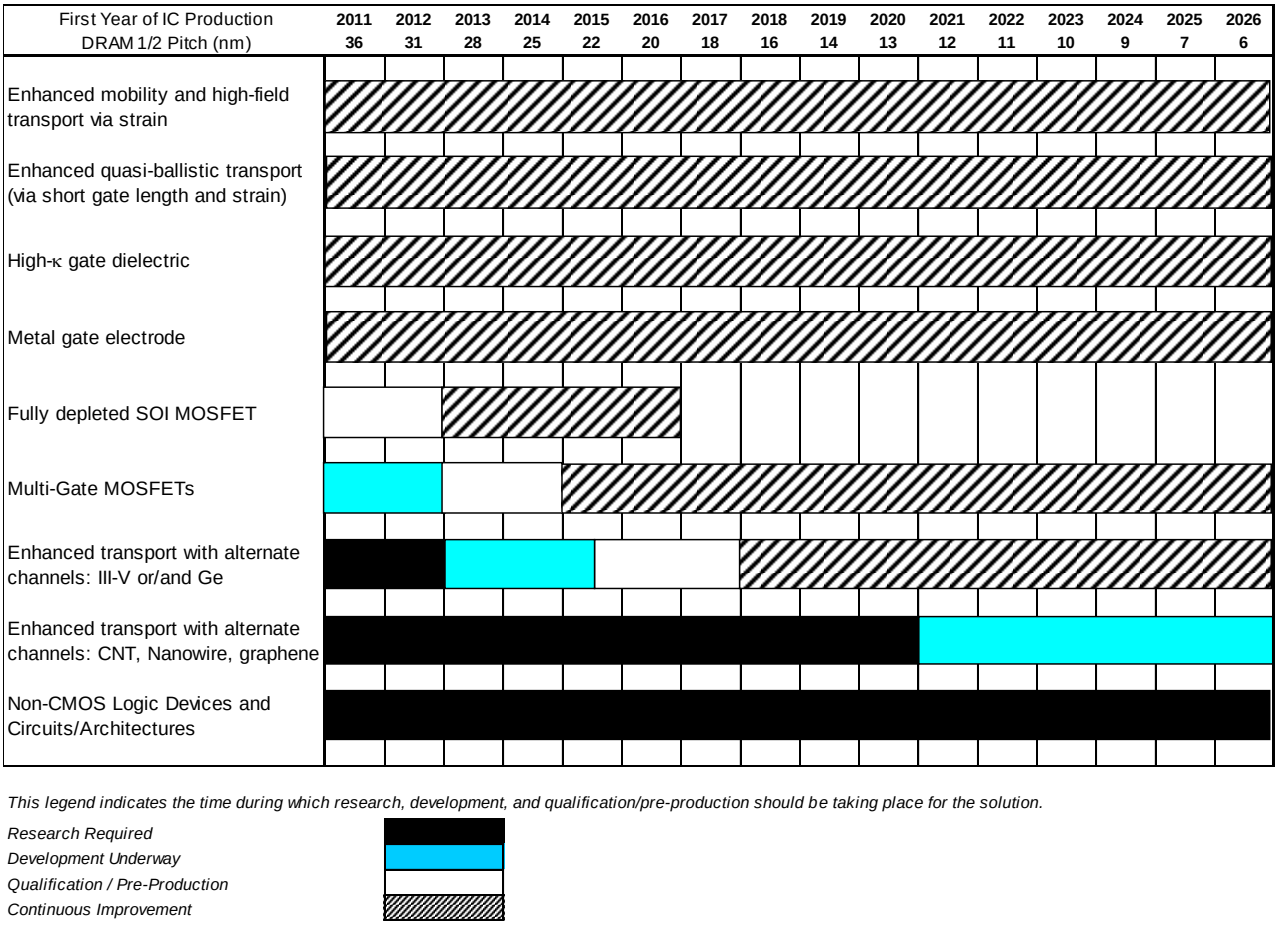


Figure PIDS3 Logic Potential Solutions

The first potential solution, enhanced mobility and high-field transport due to strain, is needed to enhance the saturation current drive to meet transistor performance targets. (Note that, in the Logic Technology Requirements tables, significantly enhanced mobility is assumed in the projections.) There are numerous techniques to implement enhanced mobility, including various types of process-induced local strain (such as heterojunction source/drain and strained liner layer) or by globally induced strain in a thin strained silicon layer, either on relaxed SiGe layers with controlled percentages of Ge or in SOI substrates. Other approaches include use of hybrid orientations (e.g., *p*-MOSFET mobility is highest for the (110) substrate orientation) or use of strained SiGe or (eventually) strained Ge channels. The potential solutions figure indicates that continuous improvement will be needed here, to increase the mobility enhancement to the maximum extent possible for both *n*-MOSFET and *p*-MOSFET, to integrate mobility enhancement optimally with the overall process flow, and eventually to utilize mobility enhancement for advanced MOSFETs such as FD SOI and multi-gate MOSFETs. In addition, continuous improvement will be needed to deal with the reduced effectiveness of process-induced strain techniques with scaling: as the spacing between transistors is reduced, techniques such as embedded SiGe or Si:C in the source/drain and the addition of stressed thin film silicon nitride liner layers over the top of the transistor tend to becomes less effective at inducing stress in the channel. Overall, continue to increase the strain is getting more difficult, and the improvement of mobility and high-field transport saturates at some high strain level.

As the gate length is scaled well below 20 nm, the fully depleted, lightly doped MOSFETs are likely to require enhanced quasi-ballistic transport to meet the performance requirements (see Effective Ballistic Enhancement Factor in the Logic Technology Requirements tables for detailed numbers). These enhancements will be obtained through reduced scattering in short channel length, through improved injection at the source, and through reduction of effective mass by strain.

In order to scale the basic MOSFET structure, one of the key technology issues is the device gate stack which consists of the gate dielectric and the gate electrode. As the physical gate length is scaled, ideally the gate dielectric thickness is

scaled correspondingly to control short-channel effects, to increase the inversion charge and saturation current drive. But the effectiveness of continued thickness reduction becomes limited due to the tunnel current leakage through the gate. High-κ gate dielectric material has been a solution to solve the problem of high gate leakage current, since the gate leakage current density corresponding to a given equivalent oxide thickness (EOT) is much smaller for high-κ materials than for oxy-nitride gate dielectric. New dielectric options such as TiO_2 can continue to increase the κ value.

Use of metal gate to replace poly-Si gate is effective in eliminating the poly-depletion phenomenon. Additionally, to set the right threshold voltage, metal gate electrode provides much more flexibility in varying the work-function due to more choices of metals. The gate work-function needs to be near the silicon valence band edge for *p*-MOSFETs and near the conduction band edge for *n*-MOSFETs. Hence, different metals used for the *p*-MOSFET and *n*-MOSFET provide flexibility that poly-Si gate does not have.

As scaling proceeds, it will become increasingly difficult to effectively scale planar bulk CMOS devices. In particular, adequately controlling short-channel effects is projected to become especially problematical for such short-channel devices. Furthermore, the channel doping will need to be increased to exceedingly high values, which will tend to reduce the mobility and to cause high leakage current due to band-to-band tunneling between the drain and the body. Finally, the total number of dopants in the channel for such small MOSFETs becomes relatively small, which results in large random fluctuations in the dopant placement and number, and hence unacceptably large statistical variation of the threshold voltage. A potential solution is to utilize ultra-thin body, fully depleted (FD) SOI MOSFETs. The channel doping is relatively light, and for such devices, the threshold voltage can be set by adjusting the work-function of the gate electrode, rather than by the doping level in the channel as in planar bulk MOSFETs. Metal gate electrodes with near-midgap work-functions will be needed to set the threshold voltage to the desired values. Because of the different work-functions in this case, the electrode material will presumably be different than those utilized for planar bulk MOSFETs. In fact, one electrode material with work-function tunable within several hundred meV on either side of midgap may be possible. Due to the lightly doped and fully depleted channel, the threshold voltage control by the work-function of the gate electrode, and the ultra-thin body, these SOI MOSFETs are considerably more scalable and can deliver higher saturation drive current than comparable planar bulk MOSFETs. Single-gate SOI MOSFETs are projected for 2013 for high-performance logic. Multi-gate MOSFETs, being also ultra-thin body and fully depleted, are both more complex and even more scalable, and are projected to be implemented in 2015 for high-performance logic.

Eventually later in the roadmap, more forward-looking solutions in 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 semiconductor nanowire, carbon nanotube, and graphene nanoribbon, also based on MOSFET operation.

Finally, beyond the roadmap range of this edition (2026), 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 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.

4 DRAM

4.1 DRAM TECHNOLOGY REQUIREMENTS

In general, technical requirements for DRAMs become more difficult with scaling (see Table PIDS7). In the past two years (ITRS 2009-2011), half-pitch (HP) scaling is relatively accelerated compare to the ITRS 2007-2009 terms. Because new technologies (e.g. 193 nm argon fluoride (ArF) immersion high-NA lithography with double patterning technology, improved cell FET technology include fin type transistor [7-9], buried word line/cell FET technology [10] and so on) are launched and these technology innovations lead to half-pitch acceleration. Result of that is the DRAM technology can be feasible to produce under 30 nm half-pitch.

Table PIDS7 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 ($ZrO2/Al_2O/ZrO2$) [11] as the capacitor of 40-30's nm half-pitch DRAM. And this material evolution will be continued and ultra high-κ (perovskite $\kappa > 50 \sim 100$) material will be released in 2013. 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 PIDS4, 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.

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 tens 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. Flash memories are based on 1T cells, where a transistor serves both as the isolation (or access) device and the storage node. Several non-conventional non-volatile memories that are not based on charge storage (ferroelectric or FeRAM, magnetic or MRAM, and phase-change or PCRAM) 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 isolation (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.

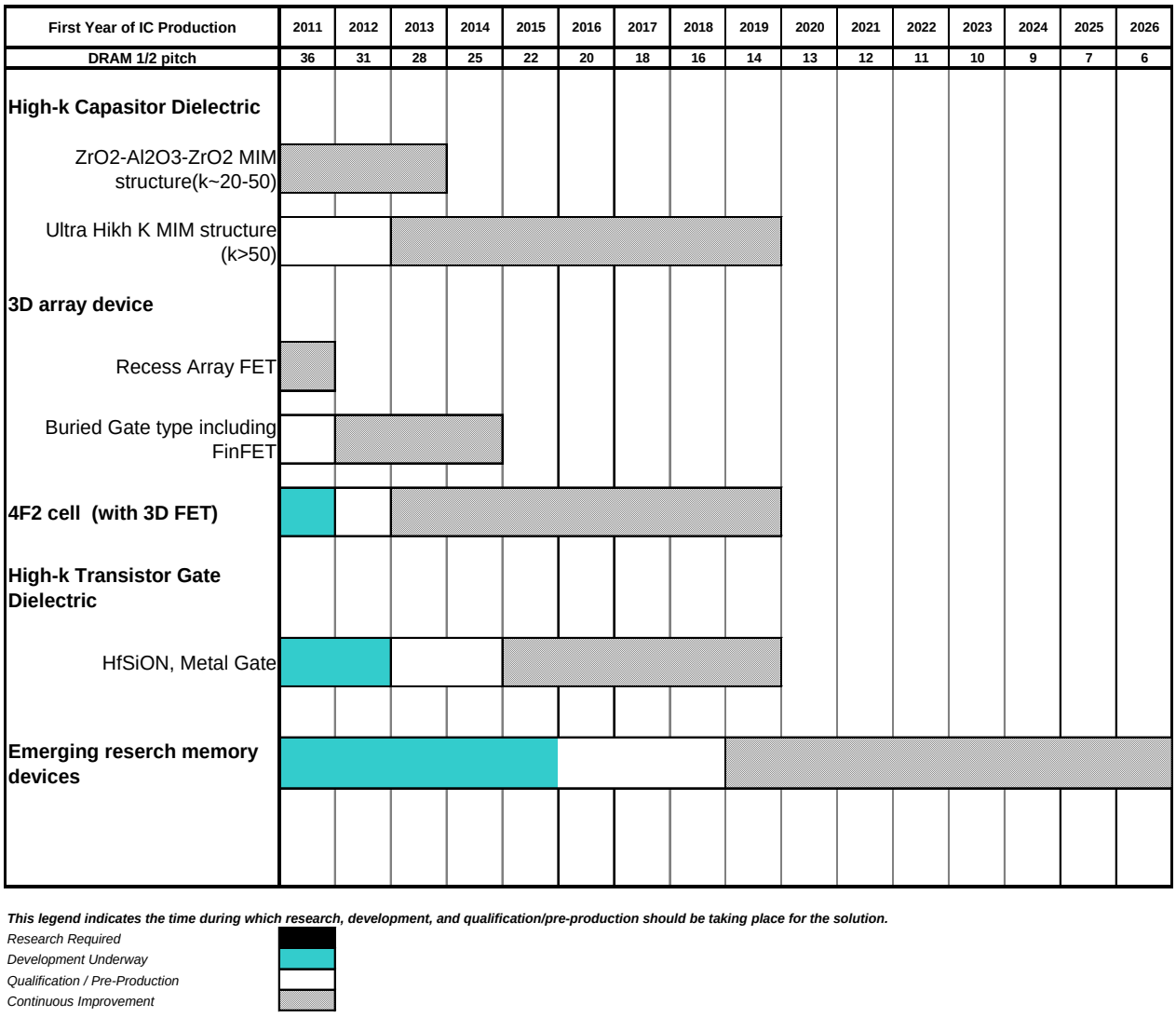


Figure PIDS4 DRAM Potential Solutions

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 they determine whether the device has adequate utility to be of interest to an end customer.

Table PIDS8a shows technology requirements for NAND flash and NOR flash, and PIDS8b non-charge-storage memories for 2011 through 2026. 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. This trend has not spread to other NVM applications.

Table PIDS8a Flash Memory Technology Requirements

Table PIDS8b Non-charge-based Non-Volatile Memory 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. The scaling issues and potential solutions for various non-volatile memory approaches are explained in the next several sections, and a summary is shown in the potential solution figure (Fig. PIDS10) at the end of this section (5.2).

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 polycide 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.

The interpoly dielectric 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. In 2010, the most advanced NAND technology (24 nm 1/2 pitch) uses an IPD around 11 nm. It is difficult to achieve the wrap around structure when the bit line spacing becomes 20 nm or less. Therefore, maintaining the GCR is a major challenge for floating gate flash device scaling.

A NAND flash cell consists of a single MOS transistor, serving mainly as the storage device. The NAND array consists of bit line strings of 32 devices or more with a selection device at each end. This architecture requires no direct bit line contact to the cell, thus allows the smallest cell size. 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 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. 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 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 importantly reducing the operation voltage, which is another scaling limiter.

To maintain a GCR > 0.6 and to avoid floating gate to floating gate cross talk are two difficult challenges when scaling below 20 nm. In addition, if new innovations (such as high-κ IPD) are not implemented and the voltages for write and erase are not reduced then word-line to word-line breakdown may limit the scaling into 1X-nm and below. Eventually, the number of storage electrons will be too low and will cause unacceptable retention time distribution and severe random (telegraph) noise, for which there is no solution currently recognized.

Despite the difficult challenges, FG NAND is expected to scale at least into 1X-nm nodes. After that, transition to charge-trapping and 3-D structures to further increase the packing density is forecasted.

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. Charge trapping devices has 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 must be 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 retention built-in field 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 and ruin the data retention.)

Several variations of SONOS have been proposed recently. 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) [12]. 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) [13], 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 20 nm 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 3-D NAND. However, most 3-D NAND proposals use charge trapping devices because of its relatively simple structure and tolerance to tunnel oxide imperfection.

5.2.1.3 NON-PLANAR AND MULTI-GATE DEVICES FOR NAND

Non-planar and multi-gate devices such as FinFET 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 requirement tables.

5.2.1.4 3-D 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 [14]. Another uses polycrystalline Si thin-film transistor (TFT) device [15]. 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 Fig. PIDS5, 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 3-D stacking.

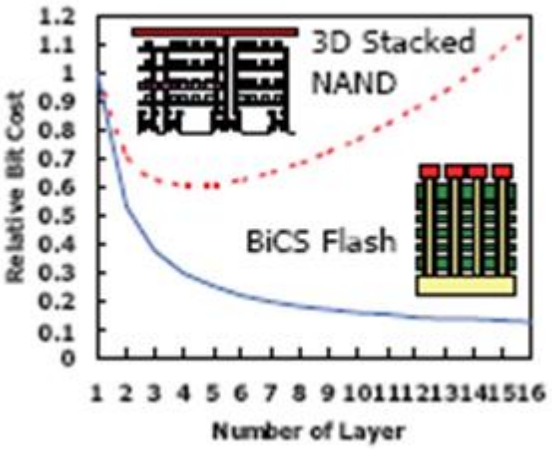


Figure PIDS5 Comparison of Bit Cost between Stacking of Layers of Completed NAND Devices and Making All Devices in Every Layer At Once (From Ref. [16])

Recently, a "punch and plug" approach is proposed to fabricate the bit line string vertically to simplify the processing steps dramatically [16,17]. This approach makes 3-D stacked devices in a few steps and not through repetitive processing, thus promises a new low cost scaling path to NAND flash. Figures PIDS6a and PIDS6b illustrate 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 Fig. PIDS5, this type of 3-D approach is much more economical than the stacking of complete devices, and the cost benefit does not saturate up to a quite high number of layers.

Various architectures for low-cost 3-D NAND have been proposed since BiCS, all employing the same principle of making all devices in a few simple operations [18-21]. These approaches may be put into three large categories: vertical channel, vertical gate, and floating gate, and are depicted in Figs. PIDS6, PIDS7 and PIDS8, respectively.

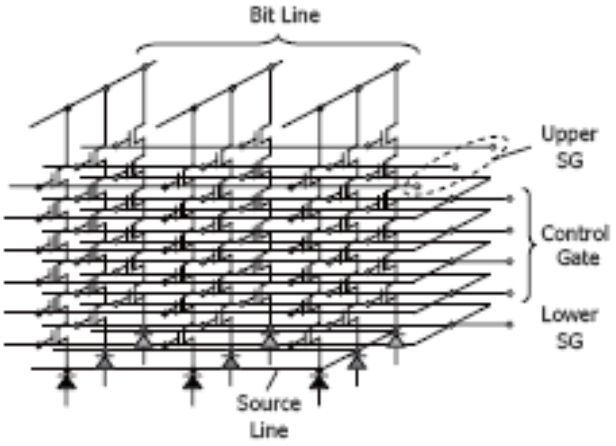


Figure PIDS6a A 3-D NAND Array Based on a Vertical Channel Architecture (From Ref. [16])

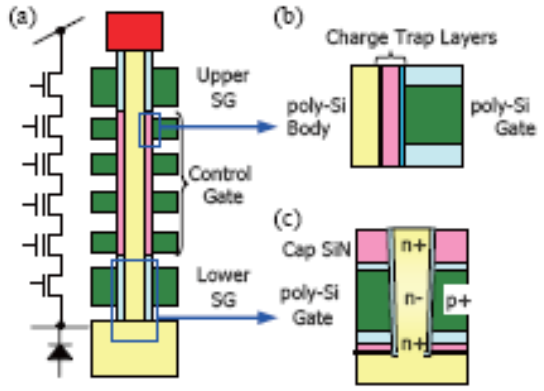


Figure PIDS6b *BiCS (Bit Cost Scalable)—A 3-D NAND Structure using a Punch and Plug Process (From Ref. [16])*

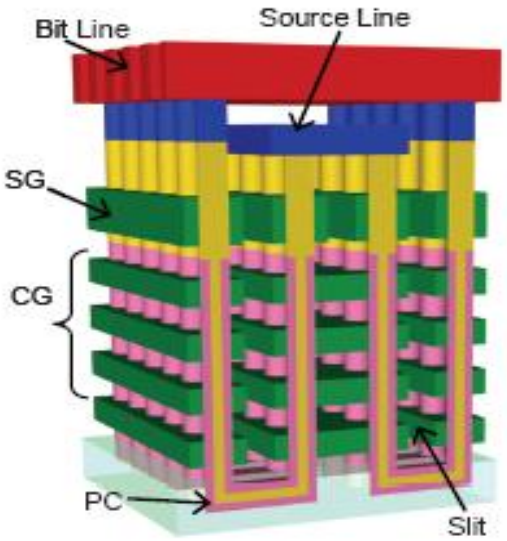


Figure PIDS6c *P-BiCS (Pipe-shaped BiCS)—An Advanced Form of BiCS 3-D NAND Array (From Ref. [17])*

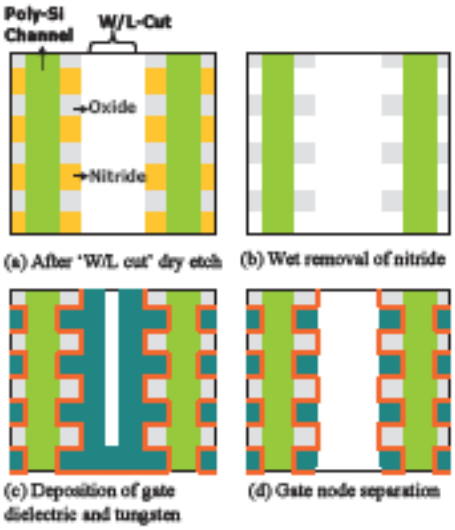


Figure PIDS6d *TCAT (Terabit Array Transistor)—A Gate Last 3-D NAND Array (From Ref. [18])*

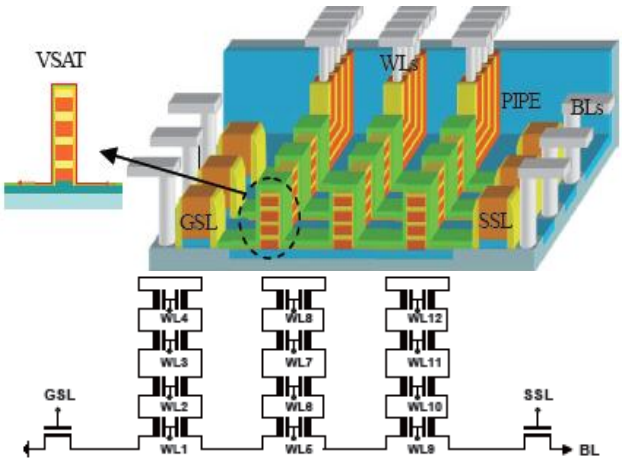


Figure PIDS6e *VSAT (Vertical Stacking of Array Transistors)—Equivalent to Folding up the Horizontal Bitline String Vertically (From Ref. [18])*

The basic architecture for vertical channel approaches is shown in Fig. PIDS6a, and may be achieved by a number of different structures – BiCS (Bit Cost Scalable, Fig. PIDS6b, Ref. [16]), P-BiCS (Pipe-shaped BiCS, Fig. PIDS6c, Ref. [17]), and TCAT (Terabit Cell Array Transistor, Fig. PIDS6d, Ref. [18]). 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-κ/metal-gate for faster program/erase. VSAT (Vertical Stacking Array Transistor, Fig. PIDS6e, Ref. [19]) has a different architecture. It resembles a folded-up 2-D NAND string, as shown in Fig. PIDS6e. All structures share a common feature that the transistor channels in the array are in the vertical direction. Their detailed working mechanisms can be found in the cited references.

The vertical gate architecture is shown in Fig. PIDS7a. The structure resembles the staking of 2-D NAND arrays side by side. The two VG approaches shown in Figs. PIDS7b and PIDS7c differ in decoding method, which is also a difficult challenge for vertical gate 3-D NAND.

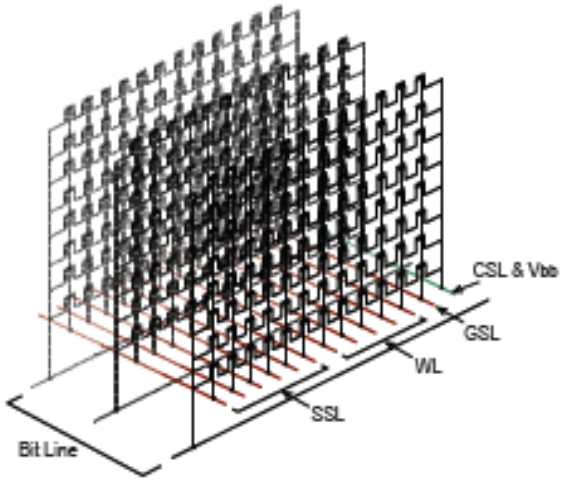
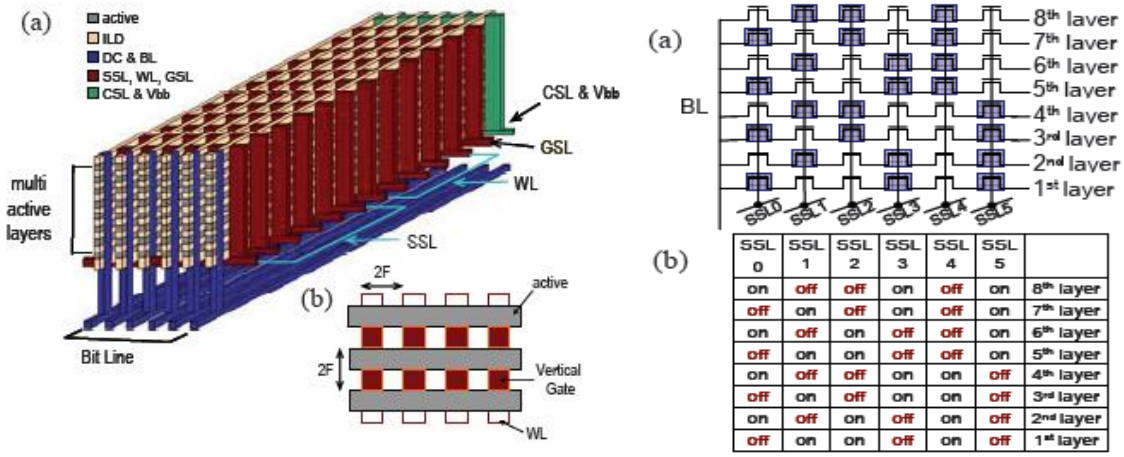


Figure PIDS7a Vertical Gate 3-D NAND Architecture

The bit-line strings are in the horizontal direction as in the conventional 2-D NAND. Each vertical "plane" of NAND devices is reminiscent to a 2-D array. (From Ref. [20])



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 PIDS7b A Vertical Gate 3-D NAND Array with Decoding Method (From Ref. [20])

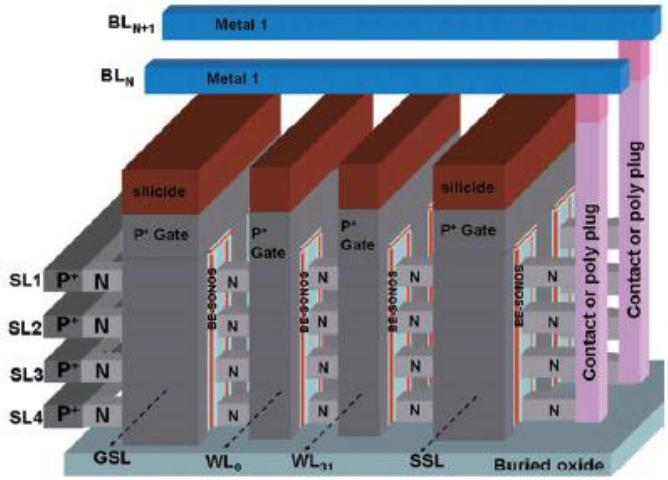


Figure PIDS7c Schematic Diagram of the PN Diode Decoded Vertical Gate (VG) 3-D NAND Architecture

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. (From Ref. [21])

Although most 3-D NAND structures use charge trapping (CT) devices, it is possible to construct a 3-D NAND array using a floating gate device. Figure PIDS8 shows such a structure using a surround gate floating gate device [22]. Please refer to the cited reference for the construction process and other details.

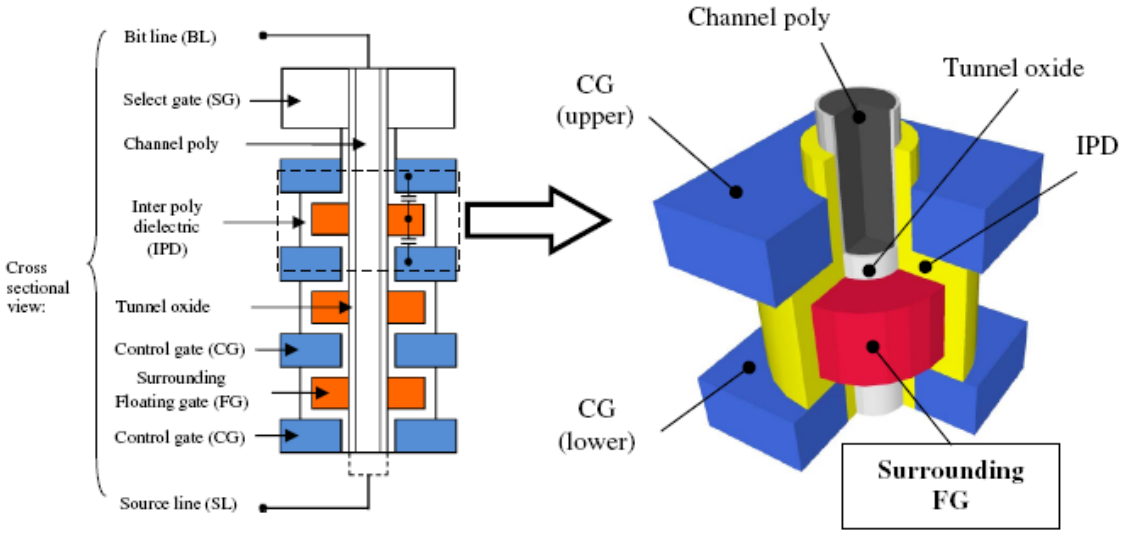


Figure PIDS8 *A Surround Gate Floating Gate 3-D NAND Structure (From Ref. [22])*

Even though charge trapping devices do not need the gate coupling ratio to operate and thus may be planar, but once put in a 3-D 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, 3-D NAND is subject to Z-direction interference [23]. The implication to cell size and scalability vary depending on the various 3-D 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 requirement table does not forecast a unique "node" or 1/2 pitch for all 3-D structures. Various 3-D approaches may be fabricated at different 1/2 pitches and with different layer numbers to achieve the same packing density. Thus the requirement table allows the choice of 1/2-pitch/layer-number that is suitable to a particular architecture.

3-D 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, 3-D 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 PIDS9 shows two schemes that may reduce the number of masks to make contacts [16,18]. Even in the best case, however, there is a considerable overhead cost for making the 3-D structure. If the 1/2 pitch for 3-D is substantially relaxed compared to 2-D 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 3-D architecture will differ.

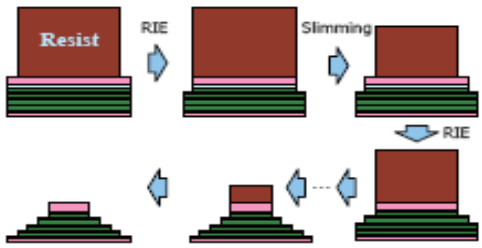


Figure PIDS9a *Scheme to Make Staircase Landing Pads for all Layers by Trimming One Single Layer of Photoresist (From Ref. [16])*

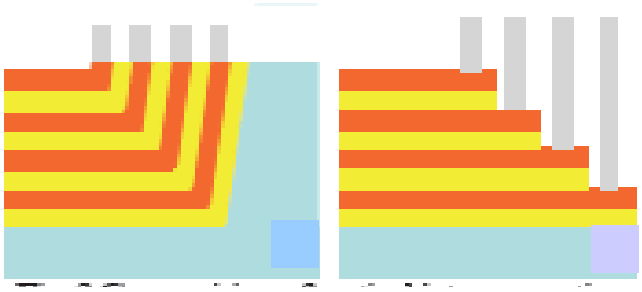


Figure PIDS9b A Scheme to Make Contacts using Tapered Deposition and Surface Contact (Left: surface contacts are made in one operation. Right: conventional staircase contacts.) (From Ref. [19])

Finally, it is important to clarify that 3-D NAND is different from the often mentioned 3-D integration of different chips using through silicon vias (TSV). 3-D 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 3-D NAND is a description of the final structure, not to be confused with an action of stacking actual devices.

5.2.2 NOR FLASH MEMORY

5.2.2.1 FLOATING GATE NOR FLASH

A NOR flash cell consists of a single MOS transistor serving both as the cell isolation (selection) 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 8 nm 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 32 nm half-pitch.

High density applications for NOR flash, especially in the 3G and beyond cell phone market, have also been steadily eroded by increasing popularity of other solutions such as DRAM/NAND SiP, that provide better performance. 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 stable, and may even expand. Therefore, the requirement table has projected a slow but steady scaling down to about 32 nm. Beyond that the technology challenges are steep, but more importantly 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 [24]. 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 1.5x to 2x 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 requirement table stays the same as floating gate NOR flash.

5.2.3 NON-CHARGE-BASED NON-VOLATILE MEMORY

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 superparamagnetism 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 two terminal devices still lacks a separate control (e.g. gate) that can turn the device off in normal state. Therefore, these devices use 1T-1C (FeRAM), 1T-1R (MRAM and PCRAM) or 1D-1R (PCRAM) 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 (selection) device, it is more difficult to design 3-D arrays that can be fabricated using just a few additional masks like those proposed for 3-D 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 data must be written back after reading. 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 [25] 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 2-D stacked capacitor through most of the near-term years. Beyond this the need for 3-D 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 to 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 65 nm node, and this is reflected in the requirement table (Table PIDS8b).

Recent advances in "spin-torque transfer (STT)" 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 [26]. For details of STT MRAM please see the ERD/ERM chapters. Although STT MRAM is still under development and both device and materials study continues but because of the closeness of product introduction, it is now included in the requirement table. 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 32 nm and below. New materials for perpendicular magnetization are still being researched, and are discussed in the ERM chapter.

5.2.3.3 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 (selection) 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 >100 ns) 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 < 5 nm has been recently demonstrated using carbon nanotubes as electrodes [27,28], and the reset current followed the extrapolation line from larger devices. In at least one case, cycling endurance of $1x10^{11}$ was demonstrated [29].

5.2.3.4 BEYOND FERAM, MRAM AND PCRAM

Beyond FeRAM, MRAM and PCRAM a large category of two-terminal resistive devices are being studied for memory applications. These resistive memories are still in research stage and are discussed in the ERD/ERM chapters.

The above potential solutions are summarized in Figure PIDS10 for easy comparison.

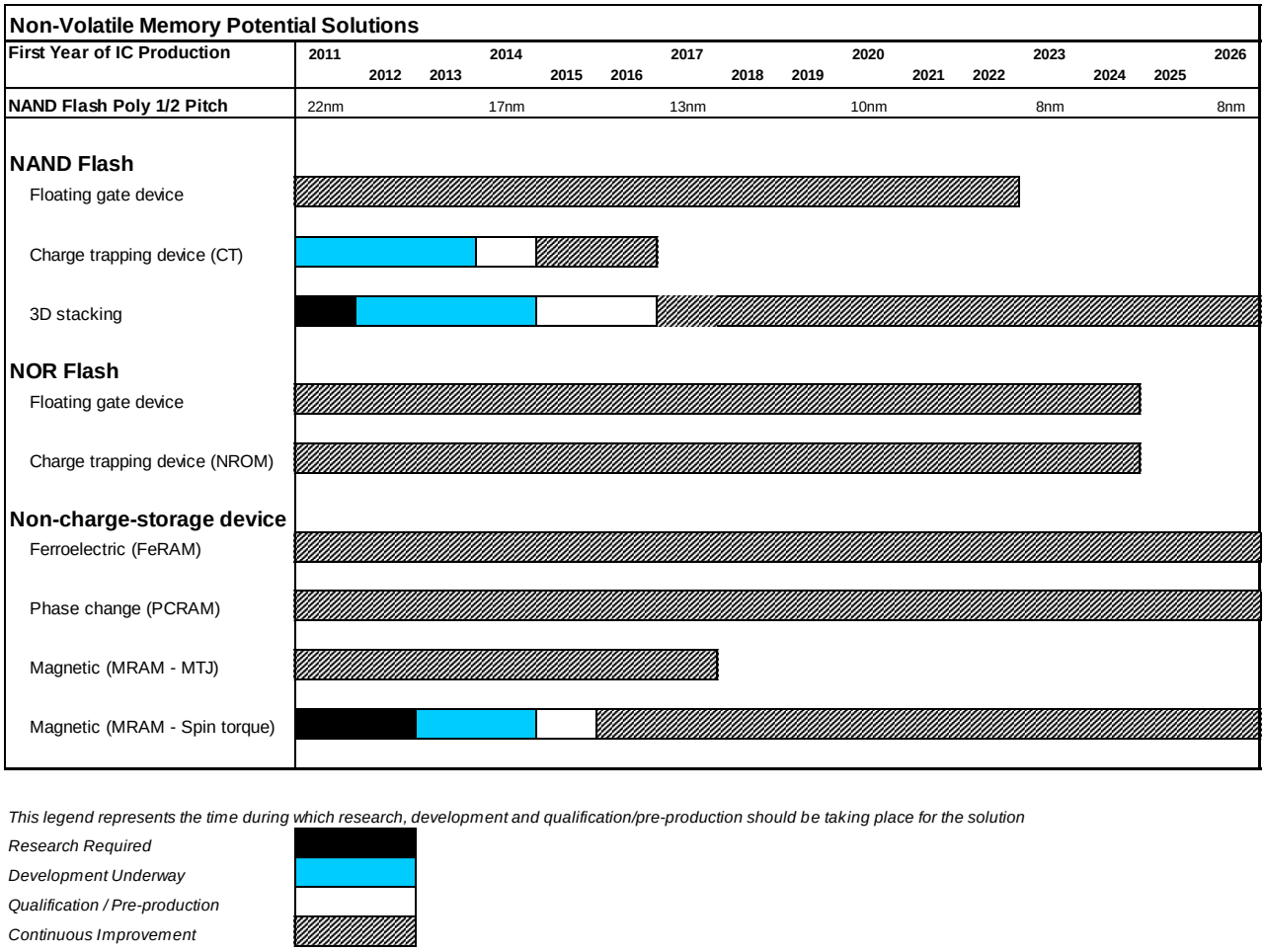


Figure PIDS10 Non-Volatile Memory Potential Solutions

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, (2) the introduction of 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, simulation suggests, a hot spot much smaller than the phonon mean-free-path exist around the drain junction of a MOSFET transistor. The temperature of such a hot spot may be hundreds of degree higher than predicted by heat diffusion, and can significantly affect the transistor reliability. Another new reliability issue due to the smallness of the transistor has already been reported to be a more severe issue than NBTI, namely random telegraph noise (RTN).

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.

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. For example, aluminum is stable after being deposited and the preferred path for electromigration is along grain boundaries. In contrast, there is grain boundary growth in copper after electroplating that can lead to stress voiding failures when a single via is connected to a wide metal line. In addition, in copper the preferred electromigration path is along the surface, making copper electromigration and stress voiding much more sensitive to the properties of the intermetal dielectric. This makes the reliability of copper lines much more sensitive to interfaces compared to aluminum. The electromigration in copper will also become worse as the cross section of the copper lines is reduced with scaling. New materials, such as high-κ and low-κ dielectrics or metal gates, and new device architectures, such as multiple 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 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-κ/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-κ 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. This 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 roadmap 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 community.

3. Mission profiles tend to be stretched further. For instance in sensor applications in automotive where temperatures exceeding 200°C 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 challenges 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 make 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 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 markets that require overall chip reliability levels to improve continuously. Applications that require higher reliability levels, harsher environments, and/or longer lifetimes are more difficult than the mainstream office and mobile applications. Note that even with constant overall chip reliability levels, there must be continuous improvement in the reliability per transistor and the reliability per meter of

interconnect 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, 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 need 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 statistic to support such guarantee.

The color-coding of the Reliability technology requirements in Table PIDS9 is meant to represent the reliability risk associated with incomplete knowledge and tools for new materials and devices. The progression from yellow to striped indicates a growing reliability risk. The requirements first turn to yellow (Manufacturing Solutions are Known) in 2011 indicating a relative smaller risk associated with scaling, increased power. It is expected more manufacturers will introduce high-κ/metal-gate transistor stacks during the time frame of now to 2012, which will present a considerable reliability risk. 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.

The requirements then turn to striped (Interim Solutions Known) in 2013. This date is approximate. It is meant to represent 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

6.2 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 atomistic 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 mean time, 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 including MEMS. 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 Cu, low-κ dielectric and alternate gate dielectrics 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. The decreasing process margin and increasing variability, which greatly degrades the accuracy of lifetime projection from a standard sample size, drive this need. 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 breakthrough 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%/yr to the current value of 13%/yr, and will likely be further reduced to 8%/yr next year, and this had been a consensus opinion of many TWGs including Design. For low-power technologies LOP and LSTP, the target metrics had largely come from Design. The over-all requirements or guide-lines of speed and power metrics among all logic technologies, summarized in Table PIDS6, is 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 [30] 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., "A new punchthrough current model based on the voltage-doping transformation," *IEEE Trans. Electron Devices*, vol. 35, no. 7, pp. 1076–1086, June 1988.

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

[4] 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.

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

[6] S. Takagi et al., "Channel Structure Design, Fabrication and Carrier Transport Properties of Strained-Si/SiGe-On-Insulator (Strained-SOI) MOSFETs," *IEDM Technical Digest*, pp. 57–60, December 2003.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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 *IEDM Technical Digest Tech.*, pp. 547-550, 2005.

[13] 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," *IEDM Technical Digest*, pp. 337-340, 2005.

[14] 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," *IEDM Technical Digest*, pp. 37-40, 2006.

[15] 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," *IEDM Technical Digest*, pp. 41-44, 2006.

[16] 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," *Symp. VLSI Technology*, pp. 14-15, 2007.

[17] 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," *Symp. VLSI Technology*, pp. 136-137, 2009.

[18] 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," *Symp. VLSI Technology*, pp. 192-193, 2009.

[19] 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)," *Symp. VLSI Technology*, pp. 186-187, 2009.

[20] 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," *Symp. VLSI Technology*, pp. 188-189, 2009.

[21] 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", *Symp. VLSI Technology*, 4B-1, 2011.

[22] 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", *IEDM Technical Digest*, pp. 668-671, 2010.

[23] 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.

[24] 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 Lett.*, **21**, pp. 543–545, Nov. (2000).

[25] 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," *Symp. VLSI Technology*, pp. 230-231, 2007.

[26] 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," *Symp. VLSI Technology*, pp. 234-235, 2007.

[27] 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.

[28] J. 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", *Symp. VLSI Technology*, 5B-4, 2011.

[29] 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 $4F^2$ Cell Size, Extendable to DRAM Applications", *Symp. VLSI Technology*, 19-3, 2010.

[30] See for example, *https://nanohub.org/resources/tools/?view=taxonomy*

MASTAR INSTRUCTIONS

DOWNLOADING AND INSTALLATION

The MASTAR application must be downloaded to your hard drive and installed there in order to run it.

Do NOT open the file from the internet.

Save the file to your hard disk by the following method:

1. Create a folder on your hard drive for the MASTAR installation file.

2. Select this interactive icon for MASTAR

then select "Save target as..." to save the file to your newly created folder on your hard drive.

►NOTE: The file name is "SetupMastar_5051_ITRS2011.exe".

3. Once the file has downloaded, go to your new folder and click on this downloaded file

►The installer will start, creating a new folder in your computer's Program Directory and placing the MASTAR application and supporting files in this folder labeled "Mastar_5051_ITRS2011."

5. When the installation is completed, open the Program Directory file folder labeled "Mastar_5051_ITRS2011."

6. Find the file named MASTAR.EXE.

7. Open the application by clicking on this file.

8. Be sure to register as a new user.

RUNNING THE MASTAR APPLICATION

For detailed instructions about MASTAR – download these *Modeling Instructions*.