INTERNATIONAL TECHNOLOGY ROADMAP FOR SEMICONDUCTORS

2009 EDITION

PROCESS INTEGRATION, DEVICES, AND STRUCTURES

THE ITRS IS DEVISED AND INTENDED FOR TECHNOLOGY ASSESSMENT ONLY AND IS WITHOUT REGARD TO ANY COMMERCIAL CONSIDERATIONS PERTAINING TO INDIVIDUAL PRODUCTS OR EQUIPMENT.

Table of Contents

Process Integration, Devices, and Structures 1
- Scope 1
  - Logic 1
  - Memory 1
  - Reliability 1
- Difficult Challenges 2
  - Description of Process Integration, Devices, and Structures Difficult Challenges 2
- Logic Technology Requirements and Potential Solutions 6
  - Logic Technology Requirements 6
  - Logic Potential Solutions 9
- Memory Technology Requirements and Potential Solutions 11
  - DRAM 11
  - Non-volatile Memory 13
- Reliability Technology Requirements and Potential Solutions 19
  - Introduction 19
  - Reliability Requirements 20
  - Reliability Potential Solutions 20
- Cross TWG Issues 21
  - Modeling and Simulation 21
- Inter-focus ITWG Discussion 22
  - Front End Processes 22
  - Impact of Future Emerging Research Devices 22
  - Emerging Research Devices 22
- References 23

Link to MASTAR model

LIST OF FIGURES

Figure PIDS1 Scaling of Transistor Intrinsic Speed of High-Performance Logic 8
Figure PIDS2 Logic Potential Solutions 10
Figure PIDS3 DRAM Potential Solutions 13
Figure PIDS4 Non-Volatile Memory Potential Solutions 15

LIST OF TABLES

Table PIDS1 Process Integration Difficult Challenges 2
Table PIDS2 High-performance Logic Technology Requirements 7
Table PIDS3A Low Standby Power Technology Requirements 9
Table PIDS3B Low Operating Power Technology Requirements 9
Table PIDS4 DRAM Technology Requirements 11
Table PIDS5 Non-Volatile Memory Technology Requirements 14
Table PIDS5A Requirements for Spin-Torque Transfer (STT) MRAM 18
Table PIDS6 Reliability Technology Requirements 20

PROCESS INTEGRATION, DEVICES, AND STRUCTURES

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, encompassing parameters such as physical dimensions, key device electrical parameters, including device electrical performance and leakage, and reliability criteria. The focus is on nominal targets, although statistical tolerances are 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, memory (including both DRAM and non-volatile memory [NVM]), and reliability.

The main aims of the ITRS include identifying key technical requirements and challenges critical to sustaining 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-solution list here is not comprehensive, nor are the solutions in the list necessarily the most optimal ones. Given these limitations, the potential solutions in the ITRS are meant to stimulate but not limit research exploring new and different approaches.

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 performance, 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 long term, new structures such as ultra-thin body and multi-gate (MG) MOSFETs (such as FinFETs). 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.

MEMORY

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

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 of NVM is Flash, for both NAND and NOR architectures. There are serious issues with scaling 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. For DRAM, the key issue is dealing with increasing scaling difficulties, especially with ensuring very low levels of leakage.

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 emphasizes mostly front-end (transistor) reliability issues. The goal is to identify the challenges that are in need of significant research and development.

DIFFICULT CHALLENGES

Table PIDS1 Process Integration Difficult Challenges

Difficult Challenges for L_g ≥ 16 nm	*Summary of Issues*
1. Scaling of logic MOSFETs	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
2. Scaling of DRAM and SRAM	DRAM— Adequate storage capacitance with reduced feature size; implementing high-κ dielectric Low leakage in access transistor and storage capacitor Low resistance for bit and word lines to ensure desired speed Improve bit density and to 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
3. Scaling high-density non-volatile memory	Endurance, noise margin, and reliability requirements Non-scalability of tunnel dielectric and interpoly dielectric in flash Difficult lithography and etch issues with pitch scaling Maintain high gate coupling ratio in floating-gate flash
4. Reliability due to material, process, and structural changes	Threshold voltage shifts due to traps, carrier injection, and program/erase cycling in memory cells Mobility degradation due to mechanical stress relaxation or interface states New or changed failure mechanisms resulting from high-κ/metal gate and new doping/activation processes New failure mechanism resulting from fundamental length scales or new device structures Process variability
Difficult Challenges for L_g < 16 nm	*Summary of Issues*
1. Implementation of advanced non-classical CMOS structures	Advanced non-planar multi-gate MOSFETs below 10 nm gate length Control of short-channel effects Drain engineering to control parasitic resistance Strain enhanced thermal velocity and quasi-ballistic transport
2. Implementation of non-classical CMOS channel materials	Identification and demonstration of alternate channel materials New issues from materials, devices, and processing Integration of alternate channel materials on Si platform
3. Identification and implementation of new memory structures	Density and voltage scaling of NVM 3-D integration of NVM Implementing non-charge-storage type of NVM Scaling storage capacitor for DRAM DRAM and SRAM replacement solutions
4. Reliability of novel devices, structures, materials, and applications	Reliability characterization of new devices Dealing with fluctuations and statistical process variations Impact of microscopic physical effects Need for Design for Reliability tools
5. Power scaling	V_{dd} scaling Controlling subthreshold current
6. Beyond CMOS	Identification and implementation of non-CMOS devices and architectures Integration onto Si-CMOS platform See *ERD* and *ERM chapters* for more discussions and details

DESCRIPTION OF PROCESS INTEGRATION, DEVICES, AND STRUCTURES DIFFICULT CHALLENGES

NEAR-TERM (L_G ≥ 16 NM):

[1] Scaling of logic MOSFETs—

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 scaling.

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 challenges associated with high

channel doping and stochastic dopant variation in planar bulk MOSFETs will be avoided, but numerous new challenges are expected. Amongst 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.

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.

Metal gate/high-κ gate stacks have been implemented in the 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 increasing 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.

Threshold-voltage tuning and control with metal gate/high-κ gate stacks has proven to be challenging, especially for low-threshold-voltage devices. For planar bulk devices, this is mainly because of 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 at or near the valence band edge for *p*-MOSFETs. This issue will be even more critical in fully depleted channels such as multi-gate and FD 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. Furthermore, since multiple threshold voltages are required, an ability to cost-effectively tune the work-function over the bandgap would be very useful.

Enhanced channel-carrier mobility and high-field velocity due to 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] 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 $4F^2$ cell size, which will require high aspect ratio and non-planar structures. Novel 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.

[3] 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 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 steep challenges. Since NAND half-pitch has pulled ahead of DRAM and logic, lithography and 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 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 speed for both devices are still difficult challenges for MLC (multi-level cell) high-density applications.

[4] Reliability due to material, process, and structural changes—

In order to successfully scale ICs to meet performance, leakage current, and other requirements, it is expected that numerous major process and material innovations, such as high-κ gate dielectric, metal gate electrodes, elevated source/drain, advanced annealing and doping techniques, new 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. The most severe threshold voltage related failure is associated with the negative bias temperature instability (NBTI) observed in *p*-channel transistors in the inversion state. Burn-in may be impacted, as it may accelerate NBTI shifts. Introduction of high-κ gate dielectric may impact both the insulator failure modes (e.g., breakdown and instability) as well as the transistor failure modes such as hot-carrier effects, positive and negative bias temperature instability. The replacement of polysilicon with metal gates also impacts insulator reliability and raises new thermo-mechanical issues. The simultaneous introduction of high-κ and metal gate makes it even more difficult to determine reliability mechanisms.

At the heart of reliability engineering is the fact that there is a distribution of lifetimes for each failure mechanism. With low failure rate requirements we are interested in the early-time range of the failure time distributions. There has been an increase in process variability with scaling (e.g., distribution of dopant atoms, CMP variations, 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. It also translates into a need to increase the number of devices tested for reliability projection. We need to develop reliability testing tool to handle the vastly increased sample size (long-term reliability and large sample size are difficult combination), and the engineering software tools (e.g., screens, qualification, reliability-aware design) that can handle the increase in variability of the device physical properties.

LONG-TERM (L_G < 16 NM):

[1] Implementation of advanced non-classical CMOS structures—

For the long-term years, when the transistor gate length is projected to become 10 nm and below, 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. The other material and process solutions mentioned above, such as high-κ gate dielectric, metal gate electrodes, strained silicon channels, elevated source/drain, etc., are expected to be incorporated along with the non-classical CMOS structures. Body thicknesses well below 10 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.

[2] Implementation of non-classical CMOS channel materials —

Eventually, toward the end of the Roadmap or beyond, scaling of MOSFETs is likely to require alternate channel materials in order to continue to improve performance, power, density, etc. To attain adequate drive current for the highly scaled MOSFETs, materials with light effective masses are greatly beneficial in quasi-ballistic operation with enhanced thermal velocity and injection at the source end. The next materials of choice seems to be III-V materials or/and germanium (or SiGe). Eventually, other candidates such as semiconductor nanowires, carbon nanotubes, or graphene nanoribbons may be utilized.

Difficult challenges include developing a new knowledge base concerning new material and device properties, as well as processing issues, such as interface passivation to provide a low-defect interface between channel and gate dielectric. These could be quite different from the long-accumulated knowledge base on Si. Also, it is expected that such solutions will be integrated either functionally or physically with the high-performance, cost-effective, and very dense CMOS logic platform. Such integration requires epitaxial growth of foreign semiconductor on Si substrate which has shown to be challenging.

See Emerging Research Devices and Emerging Research Materials chapters for more discussions and details.

[3] Identification and implementation of new memory structures—

Dense, fast, and low-voltage 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. All of the existing forms of nonvolatile memory face limitations based on material properties. Success will hinge on finding and developing alternative materials and/or development of alternative emerging technologies. For example, the conflicting requirements of low-voltage operation and retention time make the tunnel-oxide scaling difficult, if not impossible. This fact makes the non-charge type of NVM, such as phase-change memory, attractive, but their ultimate scalability is also unproven. Ultimate density scaling may require 3-D architecture, such as vertically stackable cell arrays in monolithic integration, with acceptable yield and performance.

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 replacement solution which would be highly rewarding. See Emerging Research Devices chapter for more detail.

[4] Reliability of novel devices, structures, materials, and applications—

Many new materials and structures have been proposed, yet currently very little is known about the corresponding failure mechanisms. There is a need to have reliability characterization in place well in advance of application in order to develop appropriate reliability requirements and qualification procedures. For disruptive solutions it is likely that there will be 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 in design, screens, and tests. It also seems likely that there will be less-than-historic amounts of time and money to develop these new reliability capabilities. Disruptive material or devices lead to disruption in reliability capabilities and it will take considerable resources to develop those capabilities.

For such devices, the impact of statistical variations is not well understood. Compounding the issue is the ability to control the critical dimensions of the device is diminishing. Percent variation in gate length and width is increasing with each generation of technology.

Difficult challenge is also to deal with the impact of microscopic dimensions that may result in quantum effects, including line-edge roughness, ultra-thin body thickness and narrow width.

The fraction of electronic products that demand much higher level of reliability than is generally acceptable is rising. “Life at stake” applications are increasing (e.g., biotechnology products that are implanted into people’s bodies). The rise of these applications and the increasingly difficult challenge of assuring reliability are on a collision course. There is need for “Design for reliability” circuit tools such as “Reliability-aware design” and “Fault-tolerant design.” At some point, a paradigm change from ensuring all devices meeting specifications to accepting a certain probability of failure at the device level may become necessary.

[5] 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 causes the active power density ($\sim V_{dd}^2$) to climb with scaling, soon to an unacceptable level. Alternate channel materials and/or devices such as tunnel field-effect transistor can provide some relief in this area by potentially allowing more aggressive V_{dd} scaling or/and steeper subthreshold slope.

For high-performance logic, in the face of increasing chip complexity and increasing transistor leakage 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 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.

[6] Beyond CMOS—

Eventually, toward the end of the Roadmap or beyond, scaling of MOSFETs is likely to become ineffective and/or very costly, and advanced non-CMOS solutions will need to be implemented to continue to improve performance, power, density, etc. It is expected that such solutions will be integrated with a CMOS baseline technology.

LOGIC TECHNOLOGY REQUIREMENTS AND POTENTIAL SOLUTIONS

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 performance, and 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 leakage currents are limited by battery life. There are two major categories within low-power; low operating power (LOP) and low standby power (LSTP) logics. LOP chips are typically for relatively high-performance mobile applications, such as notebook computers, where the battery is likely to be high capacity and the focus is on reduced operating (i.e., dynamic) 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 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 rapidly scaled for high-performance logic. The transistors for LOP chips have somewhat lower performance and considerably lower leakage current, while the transistors for LSTP chips have both the lowest performance and the lowest leakage current of the three. For LOP logic, the gate length lags behind the high-performance transistor gate length by ~1 year in near-term years, reflecting historical trends and the need for low leakage current in mobile applications. For LSTP logic, the gate length lags that of high-performance logic by ~2 years in near-term years, reflecting the ultra-low leakage current required. However, both kinds of low-power transistors merge with the high-performance logic in gate length around year 2014 (*See the 2009 Executive Summary).*

It should be mentioned here that recent surveys and literature indicate that the gate-length scaling has been less aggressive than the past roadmap predictions. Realignment for this effect was the major change in the ITRS 2008 edition. Reiterating that 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. It is also observed that with the new L_g scaling model, the *CV/I* speed metric has a slope of ~13% increase per year instead of 17%. Similar changes were made to LOP technology whose physical gate length had a slow-down of 1-3 years, also with a change of slope. In this year's edition, a minor adjustment of 1-year slow-down is made compared to the 2008 edition for most logic devices.

For generating the entries in the logic technology requirements tables, the *MASTAR* MOSFET modeling software was used.[1,2,3] 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. An important calculated output parameter is the intrinsic MOSFET delay, $\tau = CV/I$, where C is the total gate capacitance (including parasitic gate overlap and fringing capacitance) per micron transistor width, V is the power supply voltage (V_{dd}), and I is the saturation drive current per micron transistor width ($I_{d,sat}$). τ is a reasonable metric for the intrinsic MOSFET delay, and hence $1/\tau$, the transistor intrinsic switching frequency, is used as a key transistor performance metric. (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,[4] 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.)

To reflect more accurately the transistor speed metric, added in this year 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 measured parameter, so we feel it is a more suitable parameter to monitor the real 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 also kept for continuity and comparison.

To determine the projected parameter values in a table, the main target is $1/\tau$ vs. years, for a given fixed off-current. Then the input parameters are tentatively chosen based on scaling rules, engineering judgment, and physical device principles. Using MASTAR, the input parameters are iteratively varied until the target is met, and the final set of values for the input parameters is entered into the table. *The MASTAR program and the specific MASTAR process and roadmap files used to generate the tables are on the ITRS website.*

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

In each of these logic devices, multiple parallel paths in structures are followed. Planar bulk CMOS is extended as long as possible, while advanced CMOS technologies—ultra-thin body fully depleted (UTB FD) silicon-on-insulator (SOI) MOSFETs and multi-gate (MG) MOSFETs (such as FinFETs), are implemented in 2013 or later, and run in parallel with the planar bulk CMOS for some period (for details see the logic tables). There is always a question that for the multi-gate structures, whether they will be on bulk wafers or SOI wafers. It is understood that their DC and AC performances are equivalent 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 UTB FD MOSFET scaling, and hence the ultimate MOSFET is projected to be the multi-gate device. 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 this roadmap are meant to reflect this.

For the high-performance logic tables (see Table PIDS2), the driver is the MOSFET intrinsic performance metric, $1/\tau$, although there is plan to switch to ring-oscillator speed eventually. Specifically, the target is an average 13% per year increase in $1/\tau$, which matches the historic rate of improvement in device performance in recent years, and has been slowed down from the previous 17% per year. Meeting this target is an important enabler for the desired rate of improvement in the chip clock speed. 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 tables. The NMOSFET saturation drive current, $I_{d,sat}$, pretty steadily increases over the course of the Roadmap in order to keep $1/\tau$ increasing at the desired 13% per year rate. 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 shows the scaling of $1/\tau$ for high-performance logic. Overall, the 13%/year target is met, with some precaution. For planar bulk, for 2009 and beyond, the $1/\tau$ 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 indicated in the MASTAR simulations, where the required channel doping increases sharply with year, to a very high value of 7.5×10^{18} cm^{-3} in 2015.) For UTB 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 4 nm in year 2019. This thin-body requirement is relaxed with the MG structure and scaling could continue until the end of this range 2024.

Table PIDS2 High-performance 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 somewhat slower than the transistor intrinsic frequency, as expected. For fan-out of one, the frequency ratio is about 5, where as for fan-out of 4, the ratio is about 10. The slopes for both cases are slightly less than that of *CV/I*, around 11%.

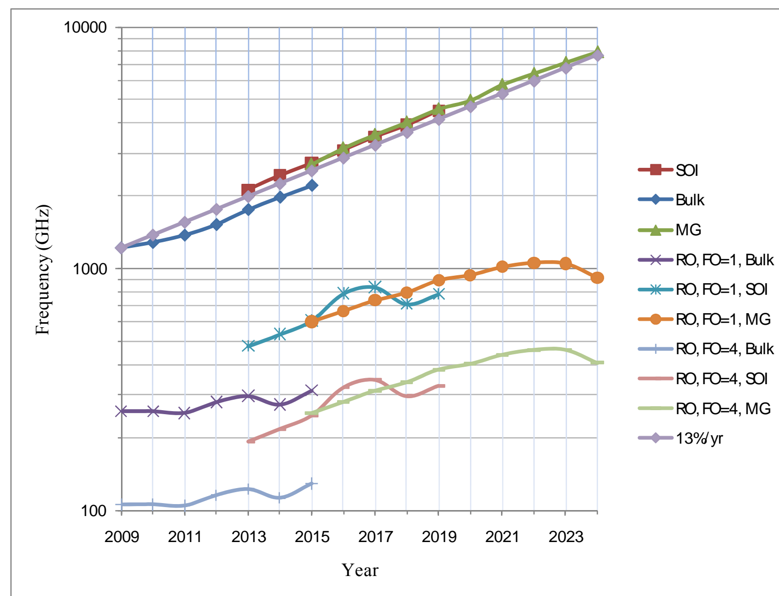


Figure PIDS1 Scaling of Transistor Intrinsic Speed of High-Performance Logic

In the legend section, the first set of 3 symbols represent the inverse of CV/I (1/τ). The second set represent inverse of the ring-oscillator delay per stage, for fan-out of 1. The third set represents the same but with fan-out of 4. The last curve is the reference of 13% increase per year.

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 more cores (multi-core) 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.

For high-performance chips, the high subthreshold leakage current must be dealt with to keep chip static power dissipation within tolerable limits. One common approach is to fabricate more than one type of transistor on the chip, including the high-performance, low-V_t device described above, as well as other MOSFET(s) with higher-V_t and sometimes larger EOT to reduce the leakage current. These alternate, lower leakage devices will have lower saturation drive current and hence poorer device performance (i.e., lower MOSFET intrinsic switching frequency, $1/\tau$) than the high-performance devices. The high-performance device is used just in critical paths, and the low leakage devices are used everywhere else. Extensive use of the low leakage devices can significantly reduce the chip static power dissipation without seriously degrading chip performance. Current circuit/architectural techniques to curtail static power dissipation include pass gates to cut off access to power/ground rails or other techniques to power down circuit blocks. Other potential techniques include well biasing, or using electrically or dynamically adjustable-V_t devices. Hence, a realistic picture of scaled high-performance ICs is that the static power dissipation is controlled by utilizing more than one type of transistor and by utilizing device/design/architectural techniques. In the technology requirements table, we have characterized only the high-performance transistor because this transistor is the technology driver.

For low-power chips, the important boundary is the source/drain subthreshold leakage current, $I_{sd,leak}$. For LSTP logic, $I_{sd,leak}$ is set at 50 pA/μm, while it is 5 nA/μm for LOP devices. All the other parameter values in the tables are chosen

iteratively to meet the $I_{sd,leak}$ targets, while optimizing $1/\tau$. Nevertheless, the resultant speed improvement in the device performance metric, $1/\tau$, is also around 13% improvement per year for both LOP and LSTP, the same as that of HP devices. Note that, to meet the leakage current requirements, the gate length scaling of low-power logic lags behind that of high-performance logic (see the logic tables for details). One key issue for LSTP logic is the slower scaling of V_{dd}. Refer to Table PIDS3a for LSTP data. This slow scaling 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 overdrive, ($V_{dd} - V_t$), must remain relatively large. Since dynamic power dissipation is proportional to V_{dd}^2, the dynamic power dissipation for the LSTP logic scales relatively slowly. But since the activity factor for this type of logic is expected to be relatively small, the lowered static power dissipation because of the very low leakage currents more than compensates for the dynamic power. In contrast to LSTP logic, V_{dd} scales relatively quickly for LOP logic (see technology requirements tables for LOP, Table PIDS3b), where, as mentioned above, the focus is on minimizing the operating power (i.e., the dynamic power dissipation, which is proportional to V_{dd}^2). However, since $I_{sd,leak}$ is larger than for LSTP logic, the saturation threshold voltage is low enough that the overdrive, ($V_{dd} - V_t$), is reasonable.

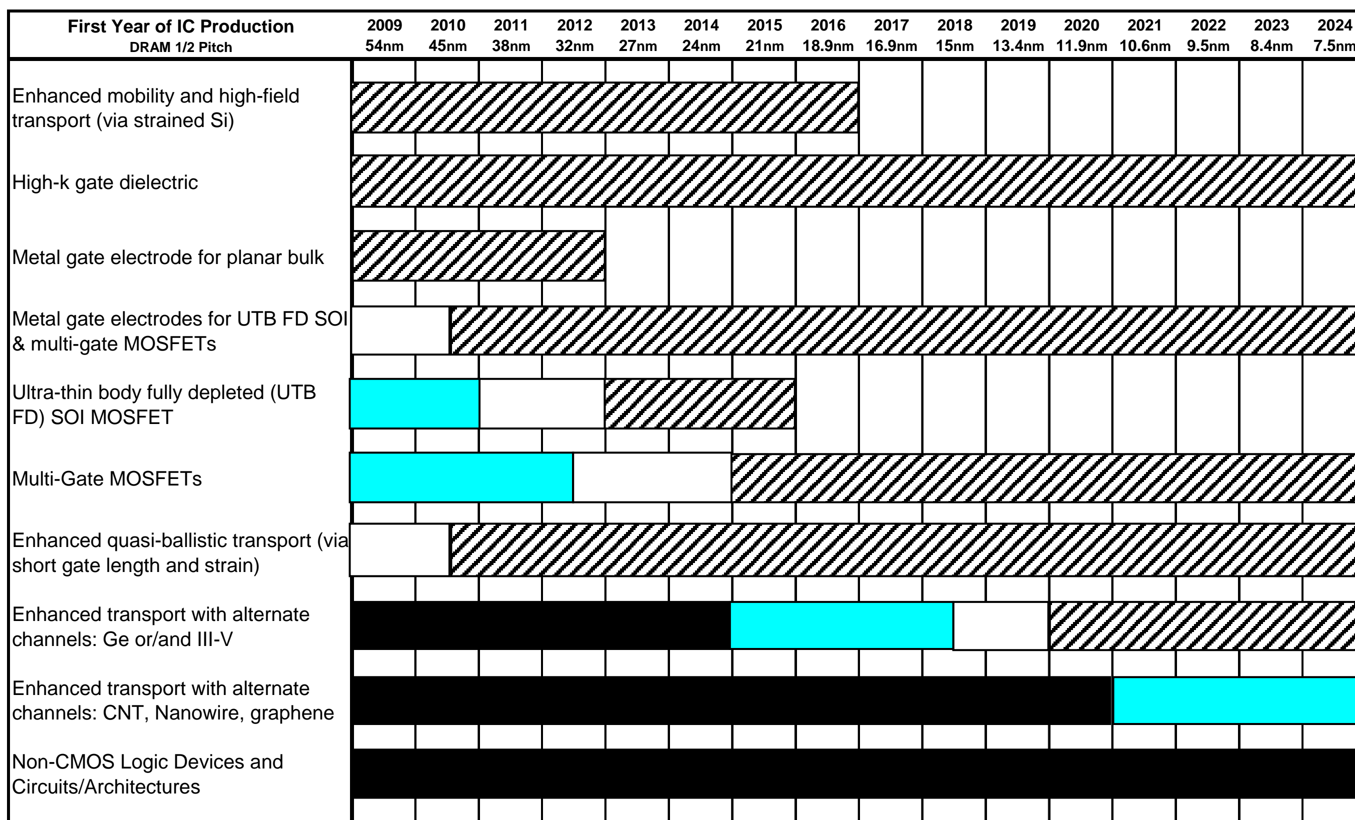
For low-power chips, the key goal is low power dissipation in order to enhance battery life, with a trade-off of low performance compared to high-performance chips. This overall goal is attained through the use of transistors with low $I_{sd,leak}$ as well as through the approaches utilized for high-performance logic: multiple transistors on chip and application of circuit and architectural techniques, including power management techniques to reduce chip leakage current in the standby mode. Eventually, effective dynamic threshold voltage adjust techniques may be feasible. The nominal targets for $I_{sd,leak}$ chosen in these LSTP logic tables are quite low, and reflect a transistor design emphasizing low leakage current in the active mode. In contrast, some companies will utilize transistors with significantly higher $I_{sd,leak}$ to get higher performance, and will thus rely more heavily on circuit and architectural techniques to lower overall chip power dissipation. Finally, for LOP logic, as discussed above, V_{dd} will be scaled relatively quickly to keep the dynamic power dissipation within tolerable limits.

Table PIDS3A Low Standby Power Technology Requirements

Table PIDS3B Low Operating Power Technology Requirements

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 PIDS2). 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 indicated by these colors. Another important general point is that each potential solution highlighted in the Potential Solutions figure 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 issues associated with these innovative solutions. Many of the potential solutions may be required first for high-performance logic. 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.



This legend indicates the time during which research, development, and qualification/pre-production should be taking place for the solution.

Figure PIDS2 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.) It was first implemented in 2004 for high-performance logic. 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., PMOSFET 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 NMOSFET and PMOSFET, to integrate mobility enhancement optimally with the overall process flow, and eventually to utilize mobility enhancement for advanced MOSFETs such as UTB 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 with scaling, 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.

In order to scale the basic MOSFET structure significantly beyond 2009 (corresponding to physical gate length of 29 nm for high-performance logic), key technology issues include 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 equivalent oxide thickness (EOT) is scaled correspondingly to control short-channel effects and to increase the inversion charge and saturation current drive.

But the effectiveness of continued EOT reduction becomes limited due to the non-scalability of poly-gate electrode depletion and inversion layer effects, which both increase the equivalent electrical oxide thickness in inversion. High-κ gate dielectric material is a solution to solve the problem of high gate leakage current, since the gate leakage current density corresponding to a given EOT is much smaller for high-κ than for oxy-nitride gate dielectric. Use of metal gate to replace poly-Si gate is effective in eliminating the poly-depletion phenomenon. For HP logic, high-κ gate dielectric and metal gate electrode have been introduced in 2009 in order to effectively prevent gate electrode depletion and hence allow acceptable scaling of the equivalent electrical oxide thickness in inversion. Low-power devices will follow in about 2 years. To set the threshold voltage correctly for planar bulk CMOS, the gate electrode work-function needs to be near the silicon valence band edge for PMOSFETs and near the silicon conduction band edge for NMOSFETs. Hence, different metals will probably be needed for the PMOSFET and NMOSFET.

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. These difficulties become worse with further scaling. A potential solution is to utilize ultra-thin body, fully depleted (UTB 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 doping 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, ultra-thin body, fully depleted MOSFETs are both more complex and even more scalable, and are projected to be implemented in 2015 for high-performance logic.

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 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.

Eventually, late in the Roadmap, more forward-looking solutions in utilization of alternate channel materials to further enhance the transport may be adopted. It is anticipated the first solutions would be III-V or/and Ge (or SiGe) channel materials, still based on MOSFET operation. Other possibilities beyond these are semiconductor nanowire, carbon nanotube, and graphene nanoribbon.

Finally, at the end of the Roadmap or beyond, MOSFET scaling will likely become ineffective and/or very costly. Completely new, non-CMOS type of logic devices and maybe even new circuits/architectures are a potential solution (see Emerging Research Devices section for detailed discussions). Such solutions may be integrated onto Si-based platform.

MEMORY TECHNOLOGY REQUIREMENTS AND POTENTIAL SOLUTIONS

DRAM

Technical requirements for DRAMs become more difficult with scaling (see Table PIDS4). The process associated with 193 nm argon fluoride (ArF) immersion high-NA lithography and double patterning technology are keys for 40 nm or smaller half-pitch DRAMs.

In recent years, DRAM cell structure was migrating to stack capacitor cell. Trench DRAM cell could not survive future scaling due to its difficulties of getting the adequate process and performance of memory cell. But, even in the stack capacitor cell, it also has many technology challenges for 40 nm or smaller size DRAM.

Table PIDS4 DRAM Technology Requirements

However, 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 that can stand up for a prolonged etch time. To overcome these challenges, the technology related to photoresists with a hard mask layer for pattern transfer is gaining importance. Furthermore, continuous improvements in lithography and etch will be needed.

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. In addition, the planar access device (cell FET) for the one transistor-one capacitor (1T-1C) cell is getting 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 compromise that, recessed channel cell FET is being adapted and optimization work have been done under half-pitch scaling. But below the 40's nm half-pitch, FinFET or 3-D type FET will be required to get the high drive current and low-voltage operation. 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 in the dielectric with scaling, process requirements for DRAMs such as front-end isolation, low-resistance materials for the word lines, self-aligned and high aspect ratio etches, planarization, and Cu interconnection are all needed for future high-density DRAMs.

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 and using high k ($HfSiO/Al_2O_3$, κ~10-25) as the capacitor of 70's nm half-pitch DRAM, ZrO/HfO as the capacitor of 50's nm half-pitch DRAM. And this material evolution will be continued and eventually ultra high-κ (perovskite $\kappa > 50$) material will be released in 2011. (See Figure PIDS4, DRAM Potential Solutions, for details.) 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 pedestal shape.

All in all, maintaining sufficient storage capacitance will pose an increasingly difficult requirement for continued scaling of DRAM devices. 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.

Keeping the chip size approximately constant as the DRAM capacity (number of bits per chip) is increased with scaling is very important from a chip cost point of view. In order to do so, the cell size factor (a) scaling (where a = [DRAM cell size]/[DRAM half pitch]^2) is critically important. Companies have started production of DRAMs with an a of 6 in 2006. And in order to accelerate the cost efficiency, $4F^2$ ($a = 4$) cell will be introduced in 2011. Migration of $4F^2$ cell required many technology challenges such as 3-D type cell FET, etc. [7]

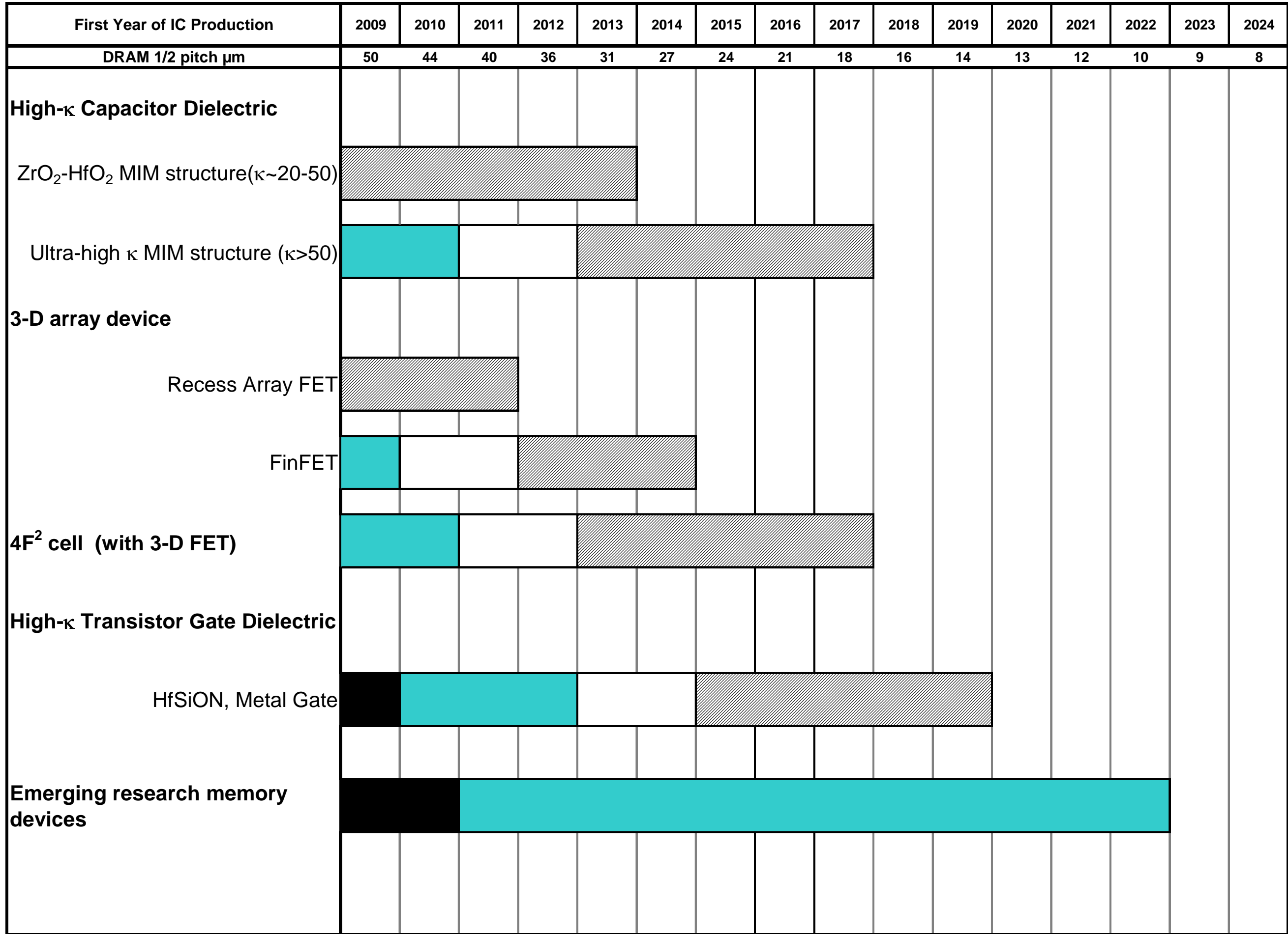


Figure PIDS3 DRAM Potential Solutions

NON-VOLATILE MEMORY

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 their applications, ranging from RFIDs that only require KByte of storage to high-density storage of tens of Gbit in a chip. The requirements tables are divided into three categories—NAND Flash, NOR Flash, and non-charge-storage memories. Each category may contain more than one approach. For example, NOR Flash memories are fabricated using both floating-gate device and nitride charge-trapping device, each with their own design rules and scaling trend. Overlapping and/or succession of different technologies for the same application are indicated as appropriate, since as technology evolves best density and performance may be achieved through multiple paths.

Information on each technology is organized into three categories. The requirements tabulation for each technology first treats the issue of 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 PIDS5 shows technology requirements for NAND Flash, NOR Flash and non-charge-storage memories for 2009 through 2024. 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). Until recently NVM half-pitches have lagged those for DRAM or CMOS logic devices in the same year. Rapid progress in NAND technology has not only reversed this trend but also has surpassed the half-pitches of DRAM and CMOS logic devices. This trend has not spread to other NVM applications yet, however.

Table PIDS5 Non-Volatile Memory Technology Requirements

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 stand-alone 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. (Figure PIDS4)

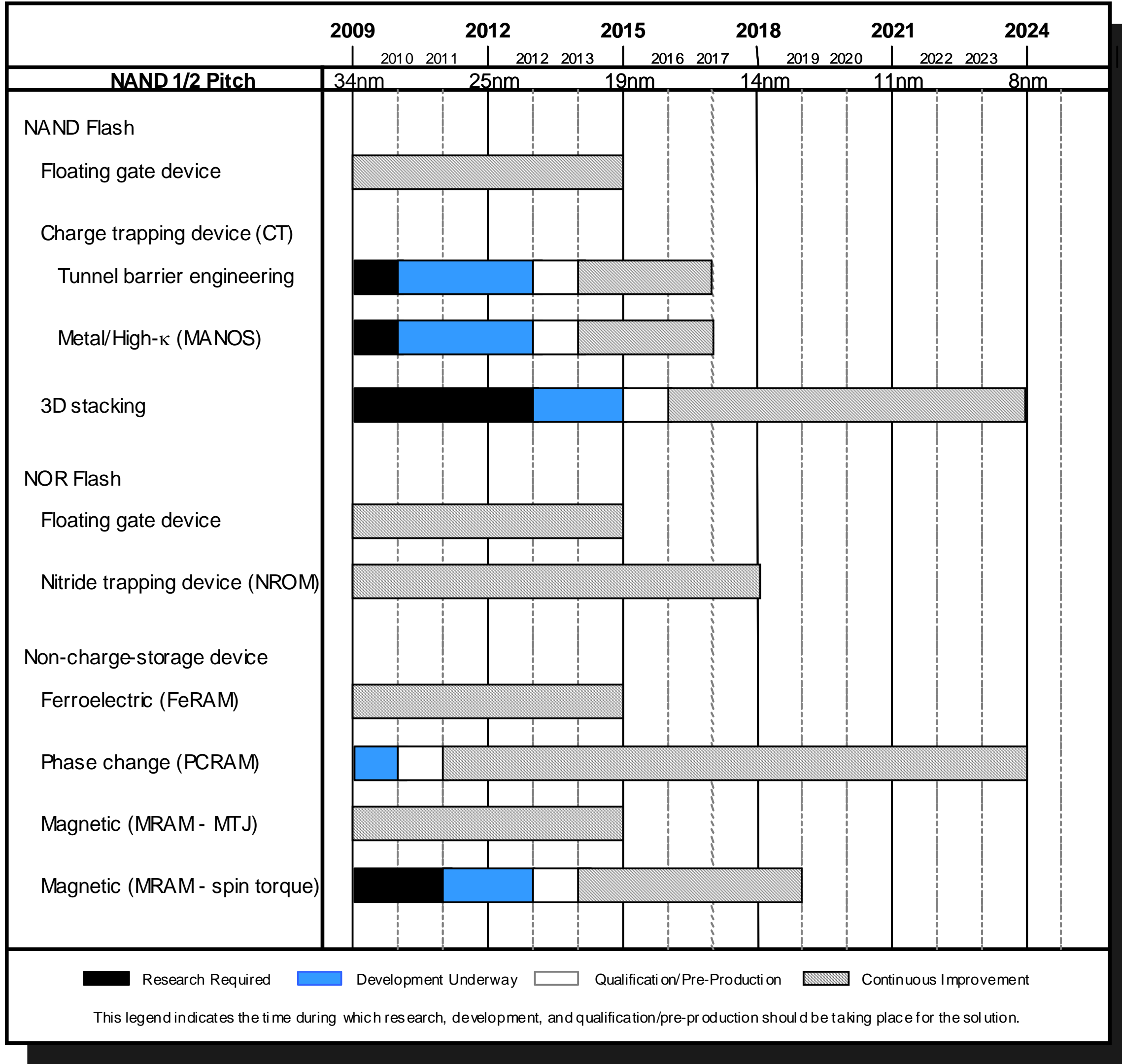


Figure PIDS4 Non-Volatile Memory Potential Solutions

FLOATING-GATE NAND FLASH

Floating-gate Flash devices achieve non-volatility by storing and sensing the charge on 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 scaling parameter, and must be ≥ 0.6. In most structures, to achieve a GCR ≥ 0.6, the word line (control gate) wraps 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 2009, the most advanced NAND technology (34 nm half-pitch) uses an IPD around 12 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. Both programming and erasing are through Fowler-Nordheim tunneling of electrons into and out of the floating gate. The low Fowler-Nordheim tunneling current allows the simultaneous programming of many bits, thus gives high programming speed. Since devices in the same bit-line string serve as pass transistors their leakage current does not affect programming or reading operation, and without the need for hot electrons junctions can be shallow. 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 NAND has weaker tunnel oxide limitation than NOR flash and is easier to scale.

How to maintain a GCR > 0.6 and to avoid floating gate to floating gate cross talk are two difficult challenges when scaling to 22 nm and below. Eventually, the few-electron limitation will cause unacceptable retention time distribution, for which there is currently no recognized solution.

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 bypassed by using charge-trapping devices. The single gate controls the MOS device channel directly and thus there is no GCR issue, and the cross talk between thin nitride storage layers is insignificant. 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.

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).[8] 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),[9] a high-κ blocking dielectric and 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.

NON-PLANAR AND MULTI-GATE DEVICES FOR NAND

Non-planar and multi-gate devices such as FinFET and 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 30 nm if innovative solutions are not found.

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. The memory density cannot be increased by continued scaling, but may be increased 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.[10] Another uses polycrystalline Si thin-film transistor (TFT) device.[11] 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 to design circuits that can handle devices that are slightly different in each layer. Although 3-D stacking can help increase the memory density beyond conventional scaling, its effectiveness diminishes after several layers are stacked. The complexity in interconnection increases and the array efficiency decreases

with the number of layers. In addition, the complex processing and the large number of masks cumulatively affect the yield, and the cost per bit benefit is only moderate.

Recently, a "punch and plug" approach is proposed to fabricate the bit-line string vertically to simplify the processing steps.[12, 13] This approach and other 3-D stacking that make devices in a few steps and not through repetitive processing,[14, 15, 16] hopefully, can provide a new low-cost scaling path to NAND flash. Note that in all 3-D approaches, the lateral (planar) half-pitch stops scaling and the density is increased by increasing the number of layers. In the requirement table, the half-pitch for 3-D NAND flash should be read as the equivalent density, no longer the lithographic half pitch. It should also be noted that charge-trapping devices are used for almost all 3-D stacking approaches, because (1) leakage of floating-gate tunnel oxide not made on Si substrate is intolerable, and (2) it is very hard, probably not possible, to build complicated floating-gate 3-D structures that can give high enough gate coupling ratio (GCR). Therefore, in the requirement table, only charge-trapping devices are considered with 3-D structures.

FLOATING-GATE NOR FLASH

A NOR Flash cell consists of a single MOS transistor serving both as the cell isolation device and the storage node. The threshold voltage of the transistor is modulated by charge stored in the floating gate and is used as an indication of the storage status. The storage cell may store single-level logic (SLC, actually means bi-level logic 1 and 0) or multiple logic levels (MLC, e.g., (11), (10), (00), and (01)). The memory array is an X-Y cross-wire structure, thus allowing random access of data. Programming is by channel hot electron or other variations of hot-electron generation, and erasing is by Fowler-Nordheim tunneling of electrons out of the floating gate. The generation of hot electrons requires high lateral electric field under the device and is provided by a steep junction profile. This in turn causes short-channel effect and leakage current that produces 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. The weaker market driving force in turn diminishes resources put on meeting technology challenges. In the requirement table, question marks are used in half-pitches < 32 nm to reflect the uncertainties of continued scaling below 32 nm.

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.[17] 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, one on the source side and one on the drain side, 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 NOR array. The cross talk between the two storage nodes in the same device cannot be completely eliminated. This so-called "second bit effect" restricts the threshold voltage window each storage node can carry, and the implementation of MLC in NROM poses a higher level of challenge than for floating-gate devices. However, NROM is intrinsically 2-bit/cell and a 4-level MLC implementation results in 4-bit/cell, compared to 16-level MLC required for floating-gate device for the same density. The virtual-ground array offers a factor of 1.5× to 2× density advantage over conventional NOR architecture using the same design rules, and the single poly process reduces the mask layers.

Charge-trapping devices do not have the gate coupling ratio issue floating-gate devices face; however, the scaling challenges are otherwise quite similar. The virtual-ground array and 2-bit/cell operation are sensitive to device leakage and the use of hot carriers for programming and erasing increases the vulnerability to reliability failures. The scaling

limitation is similar to floating-gate NOR–leakage from short-channel effect and junction breakdown, but without the severe limitation of tunnel oxide scaling the outlook may be somewhat better.

NON-CHARGE-BASED NON-VOLATILE MEMORIES

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

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[18] 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.

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 PIDS5A*). 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.[19] For details of STT MRAM please see the *ERD* and *ERM* chapters. Although STT MRAM is still under development and both device and materials study continues, but because of the closeness of product introduction, a special requirement Table PIDS5A is compiled this year, including a comparison with field switching MRAM. During the spin-transfer process, substantial current passes through the MTJ tunnel layer and this stress may reduce the writing endurance.

Table PIDS5A Requirements for Spin-Torque Transfer (STT) MRAM

(Field switching MRAM is shown for reference and is also included in the main NVM table PIDS5.)

PCRAM devices use the resistivity difference between the amorphous and the crystalline states of chalcogenide glass (the most commonly used compound is $Ge_2Sb_2Te_5$, or GST) to store the logic ONE and logic ZERO levels. The device consists of a top electrode, the chalcogenide phase-change layer, and a bottom electrode. The leakage path is cut off by an access transistor (or diode) in series with the phase-change element. The phase-change write operation consists of: (1) RESET, for which the chalcogenide glass is momentarily melted by a short electric pulse and then quickly quenched into amorphous solid with high resistivity, and (2) SET, for which a lower amplitude but longer pulse (> 100 ns) anneals the amorphous phase into low resistance crystalline state. The 1T1R (or 1D1R) 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 speed. 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 the 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 the maturity of PCRAM. 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. This property may help it secure a unique space in application.

RELIABILITY TECHNOLOGY REQUIREMENTS AND POTENTIAL SOLUTIONS

INTRODUCTION

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 scaling, the introduction of new materials and devices, increasing stresses (fields, current densities, temperatures), and increasing constraints of time and money. 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. 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.

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 leads 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.

There are even more profound reliability challenges associated with revolutionary changes associated with new materials and new devices. Recognized failure mechanisms can change. New materials, such as high-κ and low-κ dielectrics or metal gates, and new device architectures, such as multi-gate or FinFETs, can introduce new failure mechanisms or change the behavior of well-known failure mechanisms such as TDDB or BTI. For example, with the transition from oxynitride/poly-Si gates to high-κ/metal gates, a new failure mechanism, positive bias temperature instability (PBTI) in *n*-channel devices, has appeared. 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.

Moreover the speed of introduction of these new materials and devices is exceeding our capability to build up learning on new failure mechanisms and physics, whereas the failure rate requirement are become more and more demanding. The impact of an unrecognized failure mechanism that made 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.

There is consensus that one of the route to continue to the increase functionality of an IC is to integrate sensors and actuators on top of the CMOS platform. Such kind of “More than Moore” approach will greatly increase the complexity of reliability assurance. It is highly likely that such technology will come on line before the end of the road map and we

must prepare for it. The likelihood that each sensor/actuator brings along a unique set of reliability problem is high and will present a whole new challenge to the reliability effort.

RELIABILITY REQUIREMENTS

Reliability requirements are highly application dependent. For most customers, current overall chip reliability levels (including packaging reliability) need to be maintained over the next fifteen years in spite of the reliability risk inherent in massive technology changes. However, there are also niche markets that require reliability levels to improve. Applications that require higher reliability levels, harsher environments, and/or longer lifetimes are more difficult than the mainstream office and mobile applications. Note that 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, and defect screening 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.

The color-coding of the Reliability technology requirements is meant to represent the reliability risk associated with incomplete knowledge and tools for new materials and devices. The progression from white to yellow to striped indicates a growing reliability risk. The requirements first turn to yellow (Manufacturing Solutions are Known) in 2009 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 PIDS6 Reliability Technology Requirements

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 high reliability yields. 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. 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.

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. 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. 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. This need is driven by the decreasing process margin and increasing variability, which greatly degrades the accuracy of lifetime projection from a standard sample size. The ability to stress a large number of devices simultaneously is highly desirable, particularly for long term reliability characterization. Doing it at manageable cost is a challenge that is very difficult to meet and becoming more so as we migrate to more advanced technology nodes. A breakthrough in testing technology is badly needed to address this problem.

Cross TWG Issues

Modeling and Simulation

Currently, PIDS uses physical process parameters mostly provided by the Front End Processes group, and uses *MASTAR* to calculate the device *I-V* characteristics. Since MASTAR is based on analytical equations, even though it had been calibrated with real device data, prediction into the far future has some uncertainty. There might be advantages in the future to consider a full-blown physics-based TCAD device simulator to carry on such device simulations. Ideally, TCAD process simulation tool is also used to provide proper doping levels and geometries. But in order to do these, these

simulation tools need to be enhanced to deal with the key innovations requested by the *PIDS* section, including enhanced mobility, high-κ gate dielectrics, metal-gate electrodes, non-classical CMOS structures (ultra-thin body fully depleted SOI and multi-gate MOSFETs), and quasi-ballistic transport leading to enhanced saturation current, etc. These innovations will collectively drive major changes in process, materials, physics, design, etc. Similar features on III-V and Ge materials are also needed to justify as alternate channel materials. Other long-term issues requiring enhanced modeling and simulation include atomic-level fluctuations, statistical process variations, new interconnect schemes, and mixed-signal device technology. 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 and will need support from modeling and simulation to achieve the development speed required. Especially for devices that use SOI material, existing models (e.g., for dopant diffusion and activation, carrier transport or for stress) must be extended to cope with interface effects, which become increasingly important compared with bulk properties. Finally, non-classical CMOS devices require the development of appropriate compact models to support their introduction.

INTER-FOCUS ITWG DISCUSSION

FRONT END PROCESSES

There is strong linkage between the *Front End Processes (FEP)* and the PIDS chapters. Key areas of joint concern include predicting introduction years of SOI and multi-gate structures. For bulk devices, we face the difficult trade-offs of very high channel doping required to control short-channel effects. For ultra-thin body 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. Both groups feel that a device power metric needs to be added. 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 flash memories.

IMPACT OF FUTURE EMERGING RESEARCH DEVICES

EMERGING RESEARCH DEVICES

The *Emerging Research Devices (ERD) chapter* describes and evaluates potential technologies, including devices 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, 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, and cost per function, and higher functional density. Hence, the PIDS potential solutions tables include ERD solutions late in the Roadmap time period, and refer to the ERD chapter for details. Similarly, material-related topics come from the *Emerging Research Materials (ERM) chapter.* One issue to be addressed is to start a mechanism for PIDS to receive potential devices and materials that have been transited out of the ERD and ERM tables.

REFERENCES

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

[2] T. Skotnicki et al., "A new analog/digital CAD model for sub-half micron MOSFETs," Technical Digest of IEEE International Electron Devices Meeting, pp. 165–168, December 1994.

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

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

[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," Technical Digest of IEEE International Electron Devices Meeting, pp. 57–60, December 2003.

[7] See for example, Ki-Whan Song et al., "A 31ns Random Cycle VCAT-based 4F2 DRAM with Enhanced Cell Efficiency", *Symposium on VLSI Circuits Digest of Technical Papers,* p.132, 2009.

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

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

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

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

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

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

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

---

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

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

[17] 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).

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

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