2 DIFFICULT CHALLENGES

The goal of the semiconductor industry is to be able to continue to scale the technology in overall performance. The performance of the components and the final chip can be measured in many different ways; higher speed, higher density, lower power, more functionality, etc. Traditionally, dimensional scaling had been adequate to bring about these aforementioned performance merits but it is no longer so. Processing modules, tools, material properties, etc., are presenting difficult challenges to continue scaling. We have identified these difficult challenges and summarized in Table MM1below. These challenges are divided into near-term 2015-2022 and long-term 2023-2030.

Table MM1: Process integration difficult challenges.

Near-Term 2015-2022	*Summary of Issues*
1. Scaling Si CMOS	• Scaling of fully depleted SOI and multi-gate (MG) structures • Implementation of gate-all-around (nanowire) structures • Controlling source/drain series resistance within tolerable limits • Further scaling of EOT with higher K materials (K > 30) • Threshold voltage tuning and control with metal gate and high-K stack • Inducing adequate strain in advanced structures
2. Implementation of high-mobility CMOS channel materials	• Basic issues same as Si devices listed above • High-K gate dielectrics and interface state (D_{it}) control • CMOS (*n*- and *p*-channel) solution with monolithic material integration • Epitaxy of lattice-mismatched materials on Si substrate • Process complexity and compatibility with significant thermal budget limitations
3. Scaling of DRAM and SRAM	• DRAM— • Adequate storage capacitance with reduced feature size; implementing high-κ dielectrics • Low leakage in access transistor and storage capacitor; implementing buried gate type/saddle fin type FET • Low resistance for bit- and word-lines to ensure desired speed • Improve bit density and lower production cost in driving toward $4F^2$ cell size • SRAM— • Maintain adequate noise margin and control key instabilities and soft-error rate • Difficult lithography and etch issues
4. Scaling high-density non-volatile memory	• Endurance, noise margin, and reliability requirements • Multi-level at < 20 nm nodes and 4-bit/cell MLC • Non-scalability of tunnel dielectric and interpoly dielectric in flash memory – difficulty of maintaining high gate coupling ratio for floating-gate flash • Few electron storage and word line breakdown voltage limitations • Cost of multi-patterning lithography • Implement 3-D NAND flash cost effectively • Solve memory latency gap in systems
5. Reliability due to material, process, and structural changes, and novel applications.	• TDDB, NBTI, PBTI, HCI, RTN in scaled and non-planar devices • Gate to contact breakdown • Increasing statistical variation of intrinsic failure mechanisms in scaled and non-planar devices • 3D interconnect reliability challenges • Reduced reliability margins drive need for improved understanding of reliability at circuit level • Reliability of embedded electronics in extreme or critical environments (medical, automotive, grid...)
Long-Term 2023-2030	• *Summary of Issues*
1. Implementation of advanced multi-gate	• Fabrication of advanced non-planar multi-gate and nanowire MOSFETs to below 10 nm gate length

structures	• Control of short-channel effects • Source/drain engineering to control parasitic resistance • Strain enhanced thermal velocity and quasi-ballistic transport
2. Identification and implementation of new memory structures	• Scaling storage capacitor for DRAM • DRAM and SRAM replacement solutions • Cost effective installation of high density 3-D NAND (512 Gb – 4 Tb) with high layer numbers or tight cell pitch • Implementing non-charge-storage type of NVM cost effectively • Low-cost, high-density, low-power, fast-latency memory for large systems
3. Reliability of novel devices, structures, and materials.	• Understand and control the failure mechanisms associated with new materials and structures for both transistor and interconnect • Shift to system level reliability perspective with unreliable devices • Muon-induced soft error rate
4. Power scaling	• V_{dd} scaling while supplying sufficient current drive • Controlling subthreshold current or/and subthreshold slope • Margin issues for low V_{dd}
5. Integration for functional diversification	• Integration of multiple functions onto Si CMOS platform • 3-D integration

2.1 NEAR-TERM 2015-2022

[1] Scaling of Si CMOS—

Implementation of fully depleted SOI and multi-gate (field-effect limited devices) will be challenging. Since such devices will typically have lightly doped channels, the threshold voltage will not be controlled by the channel doping. Among the most critical will be controlling the thickness and its variability for these ultra-thin bodies, and establishing a cost-effective method for reliably setting the threshold voltage. Threshold-voltage tuning and control with metal gate/high-K gate stacks has proven to be challenging, especially for low-threshold-voltages as V_{dd} continues to go down. This issue will be critical in fully depleted channels such as multi-gate and FDSOI, where the effective work-function needs to be in the bandgap (although at different values for *p*-MOSFETs and *n*-MOSFETs), and where the work-function is especially critical in setting the threshold voltage because of the lack of channel doping as a variable. Furthermore, since multiple threshold voltages are sometimes required, an ability to cost effectively tune the work-function over the bandgap would be very useful.

Additionally for multi-gate structures, the channel surface roughness may present problems in carrier transport and reliability. These issues will be more severe in nanowire structures.

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

Metal gate/high-K gate stacks have been implemented in the most recent technology generation in order to allow scaling of the EOT, consistent with the overall transistor scaling while keeping gate leakage currents within tolerable limits. Further scaling of EOT with higher-K materials (K>30) becomes increasingly difficult and has diminishing returns. The reduction or elimination of the SiO_2 interfacial layer has been shown to cause interface states and degradation of mobility and reliability. Another challenge is growing gate dielectrics on vertical surfaces in multi-gate structures. A fundamental burden placed on the overall gate capacitance is the non-scalable quantum capacitance in series with the gate dielectric capacitance.

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

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

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

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

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

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

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

[3] Scaling of DRAM and SRAM—

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

For SRAM scaling, difficulties include maintaining both acceptable noise margins in the presence of increasing random V_T fluctuations and random telegraph noise, and controlling instability, especially hot-electron instability and negative bias temperature instability (NBTI). There are difficult issues with keeping the leakage current within tolerable targets, as well as difficult lithography and etch process issues with scaling. Solving these SRAM challenges is critical to system performance, since SRAM is typically used for fast, on-chip memory.

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

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

Charge-trapping devices help alleviate the floating-gate interference and GCR issues, and the planar structure relieves lithography and etching challenges slightly. Recently, high-K IPD and metal gate for planar floating gate Flash memory have been successfully developed and products with 1/2 pitch as small as 16nm have been introduced. Scaling far below 16 nm is still a difficult challenge, however, because fringing-field effects and few-electron V_t noise

margin are still not proven and more important, electric breakdown between adjacent word lines may ultimately restrict word line 1/2 pitch to > 10nm.

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

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

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

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

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

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

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

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

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

2.2 LONG-TERM 2023-2030

[1] Implementation of advanced multi-gate structures—

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

[2] Identification and implementation of new memory structures—

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

Dense, fast, and low-power non-volatile memory will become highly desirable. Ultimate density scaling may require 3-D architecture, such as vertically stackable cell arrays in monolithic integration, with acceptable yield and performance. 3-D NAND flash will require > 100 layers of stacked devices and processing technology to achieve such structures and cost effective implementation are challenging. Cost effective implementation of non-charge-storage type of NVM is a difficult challenge, and its success may hinge on finding an effective isolation (selection) device. Non-charge-storage NVM may also need to be stacked into 3-D structures to reach Tb density. Without a built-in isolation device as flash memory, the stacking of these two-terminal devices is both costly and difficult. Much innovation is needed to continue increasing storage density to 1 Tb and beyond.

See Emerging Research Devices section for more detail.

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

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

therefore lead to disruption in reliability capabilities and it will take considerable resources to develop those capabilities.

[4] Power Scaling—

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

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

[5] Integration for functional diversification—

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

To improve density on the chip, the trend of the industry is 3-D integration. This induces stress, higher temperature of operation, parasitic capacitances, interference, isolation requirement, process requirements and their compatibility with one another, and device reliability.