New Prospects of Integrating Low Substrate Temperatures with Scaling-Sustained Device Architectural Innovation

Synthesis Lectures on Emerging Engineering Technologies

Editor
Kris Iniewski, ***Redlen Technologies, Inc.***

New Prospects of Integrating Low Substrate Temperatures with Scaling-Sustained Device Architectural Innovation
Nabil Shovon Ashraf, Shawon Alam, and Mohaiminul Alam
2016

Compound Semiconductor Material and Devices
Zhaoiun Liu, Tondgde Huang, Qiang Li, Xing Lu, and Xinbo Zou
2016

Advances in Reflectometric Sensing for Industrial Applications
Andrea Cataldo, Egidio De Benedetto, and Giuseppe Cannazza
2016

Sustaining Moore's Law: Uncertainty Leading to a Certainty of IoT Revolution
Apek Mulay
2015

New Prospects of Integrating Low Substrate Temperatures with Scaling-Sustained Device Architectural Innovation

Nabil Shovon Ashraf, Shawon Alam, and Mohaiminul Alam

www.morganclaypool.com

ISBN: 9781627058544 paperback
ISBN: 9781627058551 ebook

DOI 10.2200/S00696ED1V01Y201601EET004

A Publication in the Morgan & Claypool Publishers series
SYNTHESIS LECTURES ON EMERGING ENGINEERING TECHNOLOGIES

Lecture #4
Series Editor: Kris Iniewski, *Redlen Technologies, Inc.*
Series ISSN
Print 2381-1412 Electronic 2381-1439

New Prospects of Integrating Low Substrate Temperatures with Scaling-Sustained Device Architectural Innovation

Nabil Shovon Ashraf, Shawon Alam, and Mohaiminul Alam
North South University

SYNTHESIS LECTURES ON EMERGING ENGINEERING TECHNOLOGIES #4

MORGAN & CLAYPOOL PUBLISHERS

ABSTRACT

In order to sustain Moore's Law–based device scaling, principal attention has focused on toward device architectural innovations for improved device performance as per ITRS projections for technology nodes up to 10 nm. Efficient integration of lower substrate temperatures (<300K) to these innovatively configured device structures can enable the industry professionals to keep up with Moore's Law-based scaling curve conforming with ITRS projection of device performance outcome values. In this prospective review E-book, the authors have systematically reviewed the research results based on scaled device architectures, identified key bottlenecks to sustained scaling-based performance, and through original device simulation outcomes of conventional long channel MOSFET extracted the variation profile of threshold voltage as a function of substrate temperature which will be instrumental in reducing subthreshold leakage current in the temperature range 100K–300K. An exploitation methodology to regulate the die temperature to enable the efficient performance of a high-density VLSI circuit is also documented in order to make the lower substrate temperature operation of VLSI circuits and systems on chip process compatible.

KEYWORDS

threshold voltage, substrate temperature, Fermi potential, intrinsic carrier concentration, bulk potential, depletion charge, metal-to-semiconductor work function difference, flat-band voltage, subthreshold leakage current, thin-film microcoolers

Contents

CHAPTER 1

Review of Research on Scaled Device Architectures and Importance of Lower Substrate Temperature Operation of *n*-MOSFETs

1.1 INTRODUCTION AND SCOPE OFTHIS E-BOOK

With regard to the current research impetus that has been prevalent in the device engineering and modeling arena, most technological denouements are accompanied by room temperature or 300K device architectural-based new alternative and emerging device structures to the mainstream CMOS-based devices. The key goals in pursuit of these alternative device structures are to improve the off-state leakage current (I_{off}), superior I_{on} (on-state drive current)/I_{off} ratio, enhanced inversion layer mobility and steep subthreshold slope (less than 60 mV/decade even at room temperature 300K). Various novel device structures such as Ge-on—Silicon-on-Insulator (GESOI), staggered heterojunction vertical Tunnel FET, III-V MOSFET and Tunnel FET, Gate-all-around (GAA) nanowire tunnel FET, etc., are the competitive device architectures that are being vigorously employed to sustain Moore's Law-based scaling with simultaneous accordance with the key goals or benchmark figures of merit mentioned above. The process and fabrication complexities associated with these device architectures are proving to be very challenging for manufacturing professionals to keep the device or substrate temperature at 300K yet achieve the subthreshold slope lower than 60 mV/decade, the room temperature limit of thermal velocity related carrier surmounting of source side barrier of a natural MOSFET. Tunnel FET exhibits a room temperature subthreshold slope which is less than 60 mV/decade by employing different transport features to inject carriers into the channel and does not need thermal velocity related excitation to emit the carriers over the source side barrier. After a close examination and forethought into these exciting developments that have come about only in the last four or five years, the authors of this E-book felt the urge to pursue reduced (less than 300K) substrate or lattice temperature operation of present day MOSFET structures and other alternative device structure examples mentioned above to keep pace with Moore's Law-based scaling curve and achieve

the key goals mentioned above in a more efficient way, with fewer device-centered and process-centered hindrances or bottlenecks encountered during batch manufacturing and on-chip systems integration and circuit performance. Even though the concept of low temperature electronics or cryo-cooling was developed much earlier and researchers from IBM have already analyzed device performance based on low temperature cooling, one reason the industry has been hesitant to incorporate the cooling integrated device structures with silicon dies is because of the cost associated with the cooling need which will negatively impact the consumer markets. With this context, the authors of this E-book opine that even the room temperature 300K alternative and emerging MOSFET structure such as the GAA III-V tunnel FET requires many process sequences in ultimate fabrication, which has already proved to be highly expensive compared to the required price per chip when ULSI systems-on-chip are considered. Instead of operating the device at 77K or at liquid helium or nitrogen temperature that most earlier research on low temperature electronics concentrated upon, the device substrate temperature can be tailored to operate at 250K or 200K which is not a substantial reduction from 300K, and the overhead for cooling related solid state ICs to be integrated with silicon die will not be substantial when the substrate temperature has to be within 100K of room temperature. With this implementation of reduced substrate temperature, the E-book systematically discusses the different modeling features of threshold voltage to eventually enable the device engineers to attain a host of performance benchmarks that they are investing in emerging device structural solutions, albeit at 300K operation. The authors thus feel that this E-book will be very useful to the readers or engineers concentrating on device physics-based analyses of lower substrate temperature operational benefits that are first extracted from long channel MOSFET device features and extended to today's short channel-based device architectures following a scaling ratio similar to constant field-based scaling theory employed for MOSFET scaling. At the end of the composition of this E-book, diverse set of richly blended list of references is enumerated for the devoted and inspired perusers to meditate and garner relevant facts and information which have been principally guiding impetus behind most of the contents of this E-book particularly the write-ups in the relevant sections on device architectural issues sustained scaled operation of MOSFETs and previous researches conducted on low temperature electronics, i.e., the breakthrough mainstream theme of this E-book.

1.2 BASIC OVERVIEW AND OPERATIONAL SALIENT FEATURES OF *n*-CHANNEL MOSFET DEVICE TRANSPORT

Although the CMOS device is the core element of every integrated circuit building blocks of microprocessor logic and memory units, most textbooks and research articles focus of *n*-channel MOSFET because electrons are the channel carrier with higher intrinsic and surface mobility and also since for *n*-MOSFET, if we consider the enhancement mode type operation, positive gate voltage bias is needed for channel conduction. It is easier to design power supply blocks for generating positive supply voltages rather than negative supply voltages required for channel con-

duction in p-channel MOSFETs where the transport is governed by holes, including the fact that holes have intrinsic mobility 2.5 order lower than electrons as carriers. Since this E-book details the scaling-based operational characteristics' improvements of n-channel MOSFETs, we would like to briefly document some of the salient device physical analyses on n-channel MOSFET turn-on and turn-off characteristics with effects of substrate doping, substrate bias, and oxide thickness and how the device depletion width profile evolves from long channel regime to short channel counterpart. As shown in Figure 1.1, the metal-oxide-semiconductor (MOS) field-effect transistor (FET) structure can be intuitively configured from the concept of the metal-oxide semiconductor (MOS) capacitor. In n-MOS capacitor structure, the substrate is p-type and grounded.

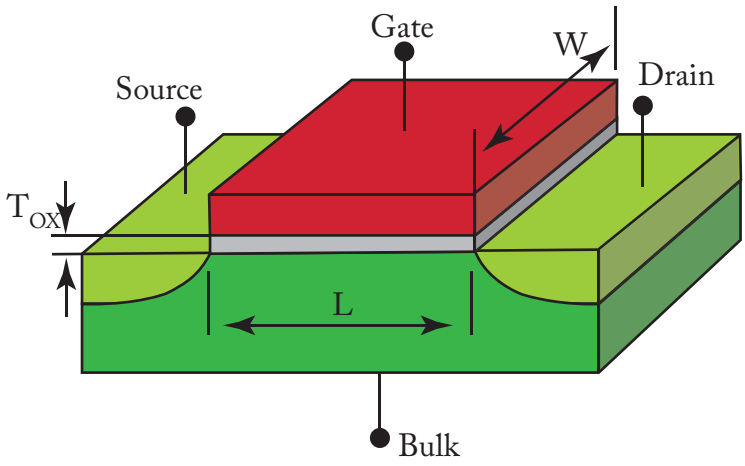

Figure 1.1: MOSFET structure.

A thin oxide is grown on p-type substrate and on top this oxide layer metal coating is overlaid to configure the gate terminal. The top metal terminal is called the gate terminal as the bias voltage applied to it, the thin inversion channel (n-type), can form underneath the oxide layer at Si:SiO_2 interface or the interface can be depleted of n-type carriers or accumulated by p-type hole carriers. When negative voltage is applied to top metal contact of this MOS capacitor from the point of flat band voltage, holes get accumulated to the surface attracted by the negative potential on the gate and the negative bias thus constitutes the accumulation region. But these mobile carriers are not minority carriers, such as electrons in p-type substrate that can be isolated from the bulk by the depletion region and utilized for conduction at the surface. Rather the accumulation regime signifies a redistribution of majority holes from the top surface to the device bulk. From the accumulation regime as the gate voltage is increased beyond flat band voltage toward more positive polarity, the positive charge on the gate repels the mobile holes at the surface and they are pushed deeper toward the bulk, leaving the immobilized negatively charged depletion

ions near the Si-SiO_2 surface forming the depletion region as the region is depleted of mobile p-type holes. If we continue applying increasing positive voltage on the gate, the depletion width expands and minority carriers such as electron-hole pairs get generated in the depletion region, and the electron carriers thus generated by thermal excitation get attracted toward the surface and start forming thin inversion charge density. At a certain positive gate voltage for n-MOS capacitor device, the depletion width reaches its maximum and the inversion charge concentration at the Si:SiO_2 interface becomes equal to the majority acceptor dopant concentration at the bulk, clearly establishing the condition that the surface has been "inverted" or the n-type minority carrier at the bulk has just become the n-type majority carrier at the surface. The gate voltage which establishes this condition is known as the threshold voltage. As the gate voltage is increased incrementally over threshold voltage, there is almost no addition to the depletion region by the additional surface band bending, but almost all of it goes into increasing the inversion charge density at the surface. At this point the concept of field effect transistor can be visualized. If for this gate voltage on the metal contact maintaining inversion charge density at the Si:SiO_2 interface we can add a source junction with significant concentration of n-type carriers, we can inject electrons and maintain this inversion charge density at the surface by modulating the metal-contact on the oxide layer, i.e., gate potential to source junction contact voltage difference. On the other side of the semiconductor surface if we can create another highly doped n-type junction and put enough high positive voltage upon it, the inversion carriers at the surface will move out or be "collected" out of this junction terminal known as drain. Thus, the thin inversion charge density is constantly supplied by source contact and flows out of drain contact—essentially forming the concept of MOSFET (metal-oxide-semiconductor-field-effect-transistor). The switching time of MOS capacitor and MOSFET is orders of magnitude difference since in the MOS capacitor the initial formation of inversion charge density takes time because of a thermally activated inversion charge or electrons in the depletion region, but in the case of MOSFET, these inversion carriers are injected and maintained by the source contact being ohmic and controlled by gate-to-source potential difference and hence take consideraby less time in generation of minority carriers. Now we focus on how gate-to-source voltage controls the injection of minority electrons into the channel, i.e., the thin inversion layer adjoining source and drain n^+ contact regions. First of all, there is a reverse bias region between n^+ source and p-type body, and based on body doping and rather degenerate n^+ doping a built-in voltage is formed across the source-to-body reverse biased diode. This reverse biased diode ideally inhibits carrier injection into the channel from the source at zero voltage on the gate and small positive or negative voltage. Therefore, on the energy band diagram, the gate-to-source barrier is very high so that with the acquired thermal or drift velocity it will be almost impossible for carriers to surmount this barrier. As the gate voltage is increased with the source at ground potential, this voltage gets coupled to the reverse biased depletion region of the source-to-body junction close to the source and reduces the built-in voltage of the source-to-body reverse biased diode. As a result of this, the gate-to-source barrier is progressively reduced and at threshold, the barrier is of such height that the carriers easily surmount the barrier through their

thermal velocity and populate the channel from the source side. Another concept that is misinterpreted is the reference potential by which gate voltage difference or drain voltage difference is computed. Ideally the *p*-type substrate contact potential is taken as reference and mid-channel potential is considered as the reference point for measuring threshold voltage or maximum depletion width. Proper knowledge out of device physical principles tells us since the source controls the injection of carriers into the channel, all references must be made with regard to source potential, including computations of threshold voltage, and maximum depletion width with regard to surface band bending potential must be referred from the source junction. Also at long channel regime, when there is ideal spacing between source and drain junctions, a thin oxide ensures proper termination of gate field lines, and the depletion region profile is rather flat for most of the gate bias ranges for small drain biases. Under this condition, since the depletion region profile is flat from source to the drain in the channel, maximum depletion width will be spatially nonvariant from source to the drain, and mid-channel reference will not alter the correct computation of threshold voltage. But as we enter the short channel regime, 2D field effects both from gate and drain terminal make the depletion region profile highly non-uniform from source to drain, having a peak near the middle of the channel within a small range. At this condition, calculation of maximum depletion width in the mid-channel region will introduce error as the source side maximum depletion width extension will be critical here to correctly compute the threshold voltage. A four-terminal MOSFET is equipped with additional terminal at the body to modulate the carrier transport in the channel of MOSFET. A negative bias on the *p*-type substrate contact will increase the reverse bias width of the source-body reverse bias diode and transistor threshold voltage, for perfectly "on" characteristics will shift toward a higher value. When 2D effects or potential contours become important in short channel MOSFETs, application of this negative bias on *p*-type body is crucial in controlling the non-uniform depletion width profile from source side into the channel toward the drain side. A positive bias voltage on the body on the other hand forward biases the source-to-body junction diode and hence threshold voltage for conducting channel will be set at a lower gate voltage than optimum. A positive voltage on the bulk will also reduce the vertical field from the gate and the reverse bias depletion field from the drain, making the 2D effect less severe for short channel MOSFET operation, essentially smoothing the asperities on the depletion width edge profile from source to drain. Interested readers are hereby directed to some excellent text books available for institutional instruction and research purposes covering MOSFET device physics and carrier transport [1–14].

1.3 REVIEW OF CHALLENGES AND BOTTLENECKS EXPERIENCED OVER SUSTAINED MOSFET DEVICE SCALING

Inquisitive readers, who may comprise technology professionals, researchers, educators, and pupils from around the world, may be interested in some classic references cited in order comprising of original research articles particularly on device scaling based architectural aspects and their

robust design schemes [19–38]. We begin the discussion in this section by citing the insightful excerpts from the classic textbook *Fundamentals of Modern VLSI Devices*, edited by Y. Tuar et al. [4]. A CMOS circuit typically consists of an n-channel MOSFET and a p-channel MOSFET connected in series between the power supply terminals so that when the circuits are driven inactive (stand by condition), there is negligible standby power dissipation. Significant power is dissipated only during dynamic activity (simultaneous switching of n- and p-MOSFET device core and also for dynamic logic or pseudo-logic device core) or switching of the circuit. By ingeniously designing the "switching activities" of the circuits on a chip to minimize active power dissipation, engineers have been able to integrate hundreds of millions of CMOS transistors on a single chip and still have the chip readily air-coolable. Until recently, the integration level of CMOS was not limited by chip level power dissipation but by chip fabrication technology. Another advantage of CMOS circuits comes from the ratio-less full rail-to-rail logic swing, which improves the noise margin even practically constrained by the requirement of low voltage supply operation, and this unique advantage makes a modular cell based CMOS chip easier to design. As linear dimensions reached the 0.5 μm level in the early 1990s, the performance advantage of bipolar transistors was outweighed by the significantly greater circuit density of CMOS devices fostered by concomitant reduction of manufacturing cost per chip. The system performance benefit of integrated functionality superseded that of raw transistor performance, and practically all the VLSI chips in production today are based on CMOS process technology. From the processing and manufacturing aspects, advances in lithography and etching technologies to enable resolution of critical dimension (CD) of gate length of CMOS have guided the industry to scale down transistors in physical dimensions and to pack more transistors in the same chip area. Such progress, combined with a steady growth in chip size, resulted in an exponential growth in the number of transistors and memory bits per chip. Dynamic random access memories (DRAMs) have characteristically contained the highest component count of any IC chips. This has been so because of the small size of the one transistor memory cell based memory architecture and because of the large and often insatiable demand for more memory in computing systems. One remarkable feature of silicon devices that fuels the rapid growth of the information technology industry is that their speed of propagation increases and their manufacturing cost decreases as their device feature size is reduced. This is one of the principal precursors of scaling induced device architectural evolution—a mainstream topic where much discussion has been laid out by noted researchers, industry collaborators, and institutional educators from the past to date and covered again in this introductory section of the E-book. The transistors manufactured today are 20 times faster and occupy less than 1% of the complete die area of those built 20 years ago. For instance, the increase in the clock frequency of microprocessors is the result of technology, logic, and device innovational schemes synergized into microprocessor core architectures through years of accumulative research built upon a single MOSFET device architecture engineering, enabling today's terahertz transmission of binary bits in its simplest form.

In lieu of attempting to critically analyze some of the classic and epochal research articles that have fueled the interests of device and manufacturing research professionals over the years, the authors of this E-book are urged by the need of well-rounded documenting of curving out novel and educational features from a list of references, authored by highly illustrious research professionals whose articles have served the purposes of CMOS device-based instruction in research institutions and also provided the pathway for the global semiconductor industry to benchmark CMOS digital and analog VLSI performance through device, process (front-end-of-line), and packaging (back-end-of-line) innovations. In this connection, the authors recommend the excellent review articles co-authored by H.S.-P Wong et al. on (i) Nanoscale CMOS [22], (ii) Device Scaling Limits of Si MOSFETs and their Application Dependencies [20], and (iii) Beyond the Conventional Transistor [23]. For the ease of understanding of the readers, the authors of this E-book are encouraged to chronicle the remaining subsections of this introductory section on the basis of a recent review article by Subhra Dhar et al. [33]. We also refer the reader to references [20, 22] and [23] to better supplement the discussion enumerated in [33]. The article by Subhra Dhar et al. on Advancement in Nanoscale CMOS Design En-Route to Ultra-Low Power Applications uses technologically sound and comparatively relaxed demonstration of central concepts behind scaling based device architectural evolution with relevant focus on performance parameters optimization to enable chip scaling. While discussing the review contents of this reference [33], the authors of this E-book would also investigate and justify the use and modeling of substrate and lattice temperature to corroborate such a scaling scenario, and where reduced substrate temperature operation of an entire chip will deservedly serve as phenomenal panacea to inconsistencies and discrepancies that might originate at a relevant technology node, due to performance bottlenecks [see 29, 30, 40–42, 54, 61–63]. The goal of operating the entire chip at reduced substrate and lattice temperature is to preserve the architectural innovation based device performance improvements but also improve intrinsic device reliability and life-time failure. When intrinsic parametric benefits aided by lowered temperature operation duly get integrated into the transport characteristics of devices as coming out of evolutionary architectural design, significantly higher performance metrics can be accrued and mastered that would not have been otherwise possible by keeping the doorway of architectural innovation open but excluding thermal engineering features from the chip. We now critically document the illustrations of article [33] in terms of device performance optimization, sustaining scaling and device attributes that need constant manipulation and control.

The process parameters in low power design are channel length, oxide thickness, threshold voltage, and doping concentration in the channel that fundamentally contribute to device performance. The low-power design technique should be such that it is cost sensitive to the process parameter variations. As technology scales down, the variations of these device-performance centric parameters are expected to be significantly disparate in future generations. The variations of leakage power and propagation delay in the transistors on a given die embedding a particular core architecture are different for different low power design methodologies applied to different

multi-core ASIC and digital circuits. The role of threshold voltage (V_{th}) and subthreshold swing (S) has become increasingly demanding with VLSI applications emphasizing ultra-low voltage subthreshold logic combined with low power high speed design. The gate oxide thickness in recent process technologies has approached the limit where the direct tunneling carrier electrons to the gate terminal starts to play a significant role in both off-state and on-state MOSFET transistor operation modes. This phenomenon, in addition to subthreshold leakage, results in a dramatic total static leakage power dissipation; for instance, the reference [44] addresses this issue. Thus, better design strategies to control the total leakage power are necessary. It is well know that gate tunneling currents are highly sensitive to the voltage variation across the gate oxide. Supply voltage attenuation commensurate with scaled die operation can give significant reduction in gate leakage power consumption. What has been apparent to the device and circuit designers is that traditional and conceptual circuit design techniques to mitigate the impact of gate leakage, which when provisioned will add to the chip area, would also be much less efficient than the use of alternative high-k materials in the standalone device fabrication itself. The fluctuations of oxide thickness variations in atomic scale result in exponentially diversified gate leakage current values whose cumulative effects from various circuit lay-outs of the die will far more outweigh the relative advantages deduced from reduced supply voltage (V_{DD}) and multi-threshold voltage CMOS operations based on scheduling the sleep mode operation of the circuit. This is the principal reason behind the use of high-k or higher-K dielectrics alternatives to SiO_2 in the device fabrication itself that will cut the gate leakage power at a faster rate than the circuit designers can demonstrate by the various circuit leakage power reduction methodologies [35]. With every technology scaling, as the device length between source and drain junction is also scaled proportionately, substrate or body doping needs to be increased proportionately to control the source side and drain side depletion extensions or encroachments into the substrate underneath the channel. Higher substrate doping results in degraded subthreshold swing, impeding the acceptable supply voltage scaling essential for acceptable standby power dissipation. Furthermore, supply voltage scaling (V_{DD}) necessitates reduction of threshold voltage (V_{th}) which exponentially increases the off-state leakage current I_{off}. I_{off} reduction is critical where the chips are often in standby (sleep) mode of operation and even during fractional clock cycle when the circuit is active, an acceptable I_{on}/I_{off} ratio still needs to be maintained since leakage power consumption is rapidly increasing at a much faster rate compared to dynamic power. Some of the front-end-of-line (FEOL) challenges faced by device designers in the deep submicron and ultra deep submicron regime of MOSFET operations are (i) diminishing $V_{gs} - V_{th}$ overdrive to sustain voltage scaling, (ii) larger V_{th}/V_{dd} where V_{th} scaling is much slower compared to V_{dd} from one technology node to the next, (iii) ultra shallow junctions necessitating dopant levels reaching the maximum solid solubility limit, and (iv) dopant loss due to inefficient activation anneal of source and drain side junctions owing to an extremely stringent thermal budget where rapid thermal annealing is employed, (v) increase of threshold voltage variability amongst intra-die and inter-die across a wafer due to statistical dopant fluctuations on very small geometry device area MOSFETs, (vi) controlling drain-induced-barrier-

lowering (DIBL) through increasing substrate doping adversely affecting the channel electron mobility through increased Coulomb and surface roughness scattering due to enhancement of vertical gate field, (vii) higher substrate doping exposing too much sensitivity to substrate biases that are routinely used to upshift or downshift the nominal threshold voltage (V_{th}) of a circuit, and (viii) the random fluctuations of gate oxide thickness, making oxide scaling to less than 1.5 nm, almost unfeasible compared to other parametric scaling and beyond 1 nm limit; self-assembly of monolayers and atomic layer deposition (ALD) technique nevertheless highly resourceful still make the targeted conformal thickness of gate oxide deviate from optimal precision.

1.4 DEVICE PARAMETERS CRITICAL FOR PERFORMANCE ENHANCEMENT FOR GENERALIZED SCALING AND AT THE END OF MOORE'S LAW

We again critically document the reference paper [33] in relation to this subsection. Device scaling is based upon simple principles: by reducing the sizes of devices and interconnects, the packing density, power density, clock speed, and on-current performance of transistors can be improved. The principal effect with technology scaling is centered upon MOSFET channel length, where the generalized scaling factor is deduced from the reduction trend of channel length as adopted by ITRS and global semiconductor researchers, professionals, and manufacturers who also support Moore's Law-based scaling as company giant Intel Inc. emerges to be the largest proponent of Moore's Law-based scaling in its every technology node production of consumer products. This scaling scenario, stemmed originally from Dennard et al. papers on scaling theory [25–28], is based on the principle that when the dimensions of a MOSFET are scaled down, both the voltage level and the gate oxide thickness are also scaled. The supply voltage V_{dd} has been scaled down in direct proportion to channel length reduction in order to keep standby leakage current under control which has a direct bearing on power consumption. The transistor threshold voltage V_{th}, although does not scale in direct proportion to channel length reduction, does follow a reduction of 70.7% of scale factor and must be reduced to maintain a high drive current by making the gate overdrive factor nearly constant through scaling nodes, and for high enough drift and mobility of channel carriers, V_{th} reduction also maintains high enough inversion charge in the channel even when gate voltage is reduced. The presence of high enough inversion charge density enhances Coulomb field related mobility at a lower vertical gate field through dopant screening. In a given technology generation, since the source-body and drain-body depletion widths are predefined based on the doping, the rate at which barrier increases as a function of distance from the source into the channel is constant. As the channel length is further reduced, the uniformity of mid-channel potential is disturbed by the higher penetration fields of the drain potential entering laterally into the channel and becoming denser toward the surface channel region between source and drain. This drain field when terminating on neutral source junction reduces the gate controlled barrier at the junction, essentially injecting more electron carriers from source contact edge into the channel. This feature, in addition to short channel MOSFET scaling scenario, re-

duces the threshold voltage at a much steeper rate than conventional scaling theory will predict. Therefore, extra precaution and maximum variation of threshold voltage window must be stipulated by the device designers to ameliorate the excessive subthreshold leakage that results from steeper threshold voltage lowering, violating the design of minimum allowable threshold voltage roll-off from a circuit operational perspective.

The gate material has long been polysilicon with SiO_2 as the insulator between the gate and the channel. Aggressive scaling of CMOS technology in recent years has reduced the silicon dioxide (SiO_2) gate dielectric thickness below than 2 nm. Gate oxide thickness has direct bearing on gate integrity of channel, and thinner gate oxide empowers the gate potential to primarily control the channel charge distribution and channel potential even when low to moderate drain bias is present. As a result, reduction of gate oxide enables the gate to turn on the inversion channel faster, maintain high I_{on}/I_{off} ratio, and also turn off the inversion channel faster—a necessary ingredient for higher transistor on—current and higher switching speed. The device engineers found it much more practical to sustain the scaling scenario down to 25 nm gate length with reduction of gate dielectric in line with scaling factor, and transition to higher-k materials has only become practical in the last five years due to excessive gate tunneling leakage currents at the lower gate oxide thickness in the vicinity of 1.5 nm [59, 64–69]. It is also necessary to enhance the body doping to an almost degenerate condition in order to ensure MOSFET operation at a substantially reduced channel length to prevent a source-to-drain direct punchthrough current. At such excessive substrate doping, a subthreshold slope degrades as the depletion region capacitance is enhanced due to much thinner depletion width. Also band-to-band tunneling (BTBT) current becomes significant at source to body and drain to body junctions when substrate doping becomes of the order of source and drain junction doping values. BTBT results as the band bends so sharply at the depletion regions of source-to-body and drain-to-body junctions that electrons from the valence band can tunnel into empty conduction band via the narrow width created by a steep band bending well, and this can happen when the gate potential is low or intentionally kept negative for an n-MOSFET device to facilitate efficient off-current bias conditions. BTBT enhances gate induced drain leakage (GIDL) primarily at negative to zero gate voltage, and low to moderate drain voltage as drain voltage enhances the GIDL current by electron tunneling, by increasing reverse bias induced more band bending at the drain-to-body junction. Finally, excessive substrate doping reduces the amount of inversion charge density at a nominal reduced gate voltage and therefore Coulomb scattering can be significant. Even when the gate induced vertical field through the oxide is low, the excessive body doping can impart substantially surface potential bending which promotes interface roughness scattering through the oxide field, and this reduces mobility at a much faster rate than Coulomb scattering-related mobility reduction.

When V_{dd} is reduced toward shorter channel lengths, it becomes increasingly difficult to satisfy both the performance and off-current requirements, i.e., maintaining the required I_{on}/I_{off} ratio. Trade off between leakage current and circuit speed emanates due to subthreshold slope factor nonscalability which is intrinsically concatenated to device or substrate temperature reduc-

tion. Therefore, only the gate oxide thickness was reduced in due proportion to scaling factor but as we have entered the nanometer regime in the last 7–8 years, supply voltage V_{dd} has not been scaled down in proportion to channel length L and V_{th} has not been scaled in proportion to V_{dd}. This resulted in nearly constant or increased V_{th}/V_{dd} ratio degrading propagation delay and exacerbating switching speed. Also due to scaling, the contact area of source and drain junction along with shallow junction depth all are reduced and as a result the source, drain contact resistance, source to channel overlapping resistance, the shallow junction spreading resistance all are enhanced and impede speed of the device and the drive current level by imparting parasitic effects through RC delay where gate capacitance, gate side-wall capacitance, gate to junction spacer capacitance all get coupled to parasitic source and drain resistance contributing agents and degrade the device performance and AC gain whose impacts become mostly visible at high frequency, switching analog and mixed signal applications. For more insightful and resourceful feedback on the content of this subsection, see [43–58, 60].

1.5 ROLE OF SUBSTRATE TEMPERATURE MODELING AND CONTROL

This particular section sets the platform for the motivation and impetus behind the contents of this E-book and the readers are directed to the reference [24] of Yuan Taur on CMOS Design near the Limit of Scaling—particularly his expert narration of a projected outlook on extending CMOS scaling to 10 nm. From the scaling scenario that has enabled the improved device architecture at the 10 nm era, the principal drawback results from decreased I_{on}/I_{off} ratio, decreased effective mobility, and subthreshold slope nonscalability. Through the introduction of multi-gate device architecture and the Tunnel-FET structure, a subthreshold slope closer to a 60 mV/decade limit at $T = 300$K and even lesser of the order of 30–40 mV/decade for Tunnel FET structures have been achieved. But these devices, particularly TFET, still suffer from much lower I_{on} value of their conventional MOSFET counterpart and hence a much reduced I_{on}/I_{off} ratio where I_{off} is decades of magnitude lower than conventional silicon MOSFET, but at the same time I_{on} is also 3–4 orders of magnitude smaller than conventional silicon MOSFET. Even with a combination of strain and stress enhancing integration of materials in the channel and capping material on gate, the mobility slightly improves and still tends to be dominated by interface roughness scattering due to a vertical field that exists across the 1 nm scaled gate oxide even with thicker high-K gate stacks; as for successful integration of high-k materials with silicon substrate, interfacial sub 1 nm gate oxide cannot be fully removed or scavenged. Oxide tunneling nearly becomes unavoidable even with multigate nanowire architecture; if not we try to implement the intrinsic attributes of the device and integrate the benefits that we accrue from intrinsic parameter engineering such as substrate temperature with the device architectural based parametric innovations. The impetus for this book hinged on ingenious optimization and reduction of substrate or lattice temperature often associated with chip cooling that will enable sustained device performance even when the ultimate deciding factor, End of Moore's Law seems attractive. As the article [24] clearly states,

the benefits are derived primarily from two aspects of MOSFET characteristics—higher carrier mobility and steeper subthreshold slope. According to the paper by Shin-ichi Takagi et al. [70], electron mobility enhances by a factor of more than 3 times from 300K to 77K substrate temperature depending on the magnitude of the vertical field and also when the body doping is less or near intrinsic. Similarly, the hole mobility improves by a factor of more than 2.5 times for the same temperature range. In addition, the MOSFET subthreshold slope steepens at a factor inversely proportional to the lattice temperature, making it much easier to turn off a device at lower than room temperature operation. Also at near intrinsic body doping where the impact of substrate temperature is maximum, the depletion region silicon capacitance (C_D) decreases due to wide enough depletion width, and depletion capacitance (C_D)/oxide capacitance (C_{ox}) decreases and this gets coupled to standard thermal energy based reduction (contribution from kT term), making subthreshold slope more steep through this additional beneficial contribution apart from simple proportional reduction of T (temperature). At lower than room temperature operation, the intrinsic carrier concentration of Si MOSFET reduces by orders of magnitude, and as a result the bulk potential and surface band bending at low enough gate voltage can still be substantially large, making the threshold voltage at a lower temperature larger than room temperature value. This has tremendous implications on the short channel behavior of the MOSFET. First for room temperature operation of MOSFET optimized for below 25 nm regime, absolute or nominal value of threshold voltage is already scaled and lowered but scaling induced 2D effects from drain, quantum mechanical effects from subbands, fringe fields from the sidewalls, higher-K gate capacitance all reduce the V_{th} even further, essentially making the V_{th} variability window difficult to track across die to die in a fabricated wafer. On the contrary, if the threshold voltage (V_{th}) at a reduced substrate temperature is shifted to a higher value than room temperature V_{th}, even with accommodation of considerable V_{th} variability window ΔV_{th} including the factors responsible for short channel effects, the final V_{th} that is measured across die-to-die will still be low enough as permitted by scaling but not too low to fail the V_{th} (min) window edge of V_{th} variability. One of the principal drawbacks at low temperature operation of MOSFET is the slower or retarding scaling of maximum depletion width W_{dm} with channel length L. As the channel length is scaled, for room temperature MOSFET, substrate doping has already been determined to be high enough to sustain scaling, and intrinsic carrier concentration is $10^{10}/\text{cm}^3$. This makes the bulk potential considerably smaller compared to the gate voltage, and as the depletion width is directly proportional to the square root of the bulk potential effect at the surface band bending and inversely proportional to the square root of substrate doping, it is the substrate doping compared to bulk potential which reduces depletion width further as technology scales. For substantially low temperature operation as may be needed to enhance V_{th} to reduce off state leakage current and produce steeper subthreshold slope, this exposes the maximum depletion width W_{dm} nonscalability with channel length at a T around 100K. This is due to the reason that surface potential becomes non-varying and large at a lower T, sometimes larger than applied low gate voltage, owing to saturating $\ln\left(\frac{N_A}{n_i}\right)$ effect where N_A is of the order of 10^{16}–$10^{18}/\text{cm}^3$ as device scales from

1 micron to a few 10 nm but n_i (intrinsic carrier concentration of silicon) can be so low that careful calculation at a T around 100K, n_i value is less than 1. Hence, ln $\left(\frac{N_A}{n_i}\right)$ reaches its maximum saturating limit incompatible with supply voltage reduction. The inversely proportional contribution to W_{dm}, square root of substrate doping alone cannot reduce W_{dm} in the adequate proportion such that the short channel efficiency factor $L_{eff} > 2W_{dm}$ is satisfied. Additionally, one device physical attribute at low temperature operation of extrinsic semiconductor is that incomplete ionization of acceptors of n-channel MOSFET becomes more prominent at lower T, and as depletion width scales inversely with doping value, it gets further boosted by the effect of incomplete ionization at every technology node when all the device designers will require is substantial reduction of maximum depletion width W_{dm} to enable gate length scaling. Therefore, this nonscalability feature of maximum depletion width at a sufficiently low substrate temperature of operation must be taken into account. By selecting a substrate temperature sufficiently close to room temperature where the advantages of improved carrier mobility and subthreshold slope factor can still be realized, use of low substrate doping can still reduce ln $\left(\frac{N_A}{n_i}\right)$ considerably so that a moderate and in-line scaling of W_{dm} can still be possible. An intrinsic or near intrinsically doped n-channel MOSFET is being designed today to control V_{th} adjustment by metal gate work function engineering with the additional benefit of higher effective channel mobility. A near intrinsic channel in conjunction with not too low substrate temperature below $T = 300$K operation of MOSFET can keep the W_{dm} scaling possible to avert the short channel effects. At the lower end of substrate temperature reduction, depletion width below the gate enhances at a faster rate which is obviously a concern for scaling enabled device architectural innovation. But since the depletion region near the drain also widens to include the increased surface band bending at lower T due to additional drain bias, drain can ideally lower the turn on V_{th} and cancel out the effect of increased W_{dm} by providing on current performance even at the expense of increased depletion width considering scaled operation and reduced substrate temperature. Non-uniform lateral doping where the doping is abruptly high near the source region and gradually decreases toward the channel up to the drain can ideally potentially reduce the overall maximum depletion width under the gate, and this reduction will be more effective as the source to drain distance narrows or constrict to a few atomic distance sustaining scaled operations. If the source to drain distance is of the order of a few nanometers, the variational effect on depletion layer width due to lateral non-uniform channel doping will be minimal since at the scaled operation of MOSFET, source side depletion width thickness is mostly relevant compared to mid-channel thickness, which is generally used in modeling equations for extracting various parameters owing to long channel and sub quarter micron channel length technology. Also inverse retrograde doping in the vertical direction (High-low) from channel surface to a few nm distance into the body can also alleviate the problem associated with depletion width enhancement at lowered substrate temperature operation. Inverse retrograde doping is engineered vertical non-uniform channel doping where the doping is steeply high at the channel surface to a few nm channel thickness that mostly embody the pathway for carrier flow from source to drain and then the doping is gradually reduced to average or effective body doping

along the vertical direction toward the bulk where surface potential bending gradually ceases to exist. Since the surface doping is high, the source side depletion region in two dimensions will include the effect of both steeply high vertical substrate doping and steeply high lateral substrate doping. Therefore, the extent to which a depletion region widens will be significantly curbed to represent the new modified W_{dm} that will still be sufficient to control the short channel and 2D effects. Substrate doping engineering in this way will therefore neutralize the impact of W_{dm} for scaled technology node operations of MOSFET at the lower substrate temperature condition.

Now that we have enumerated the efficient operation of MOSFET at the scaling limit, taking advantage of reduced substrate temperature, we will document some of the other benefits in terms of thermal management of the die under long term operation, reliability effects that arise from rise of chip junction temperature, and lastly a discussion of the relevant cooling technologies that have been introduced by technologists to implement the localized and globalized intra-die and inter-die cooling solutions. The following enumerations have been adapted from the classic reference article of Professor Kaustav Banerjee et al. from their published article, "Cool Chips: Opportunities and Implications for Power and Thermal Management" [39]. Cooling of MOSFET to reduced substrate temperature operation below room temperature benefits the back-end-of-line (BEOL) performance and reliability. Lower operating temperatures lead to smaller wire resistance per unit length, noting that metals exhibit a positive temperature coefficient and semiconductors a negative temperature coefficient. The reduction in wire resistance further enables delay in signal lines and static IR drop—a cause of major concern in extended power/ground rail-to-rail voltage application. Reliability of interconnect lines have been mostly associated with electromigration (EM) tolerance and inter-layer dielectric breakdown, also known as Time Dependent Dielectric Breakdown (TDDB), and these two metrics are found to be adequately devised by the engineering of chip cooling. This is due to the fact that at reduced thermal energy of electrons, the electrons have reduced thermionic drift velocity and also due to the increase of mean free path between collisions at lowered carrier or substrate temperature, the electrons therefore experience a lower statistical number of collisions stemming the otherwise observed growth of localized concentrations induced hot spots and steep thermal gradients, both of which are the known precursors to electromigration (EM) induced failure. TDDB is caused by defect induced conductive bridging of dielectric that exists between neighboring metal lines and also as a gate dielectric that separates the gate to channel conduction when the transistor is in the OFF state or the gate current is meant to be negligibly small. During the time the transistor is biased in the ON state, electrons and holes as generated by impact ionization near the drain can surmount the oxide tunneling barrier near the drain and create oxide traps and defects in the oxide, essentially serving the catalysts for TDDB failure at a later stage, by essentially forming the conductive bridging filament by oxide traps accumulation. At reduced temperature, the thermal energy of the electrons and holes are considerably low and at a given drain and gate bias, statistically a reduced number of electrons and holes are injected into the oxide, creating a substantially reduced number and distribution of traps along the oxide energy band, and hence TDDB lifetime of gate dielec-

tric or inter-layer dielectric becomes more extended as reliability measures. The benefits of lower temperature operation in the BEOL wire scaling is such that for semiglobal and global wires, more aggressive interconnect scaling (narrower width conductors with fixed or narrower spacing between them) can be allowed under cooled operation without degrading RC delay and EM reliability. On the other hand, this provision of scaling wires can offset any inductive effects that may become prominent due to reduced resistance per unit length at lower substrate temperature. Also at lower temperatures, intra-wire capacitance per unit length can be reduced significantly for smaller aspect ratio wires (reduced metal thickness with constant width) while maintaining the same resistance per unit length. This will lower the delay per unit length thereby enhancing the rate at which bits can be transmitted per unit chip frequency, i.e., bandwidth. Cooling techniques can be broadly classified into two types based on cooling power consumption. Passive cooling denotes cooling by conduction (heat sink) and or natural convection by air, while active cooling represents different types of cooling schemes with associated external cooling power. Typically, pure passive cooling is only applicable for systems with low power consumption due to its low heat removal capacity which is limited by the structural profile of the heat sink (size, number of fins, fins aspect ratio and spacing, thermal conductivity etc.) and surrounding temperature. Cooling techniques for sub-ambient temperature operation, including refrigeration and cryogenics, are only applicable for specialized use to achieve required performance when cost and cooling power consumption are not the primary concerns. The efficiency of various refrigeration techniques can be compared by the coefficient of performance (COP) which is determined by the ration of cooling capacity to power consumption by the refrigerator ($COP = Q_{cooling}/W_{power}$), where $Q_{cooling}$ is the cooling capacity and W_{power} is the power consumption by the refrigerator. Moreover, the limit of heat removal capacity (Q) can be further improved by using liquid refrigerants as coolants. The concept of microchannel cooling with liquid has been introduced and investigated for applications with higher heat removal requirements (large size or array of high performance ICs). The latest technological development in properly designing microfabrication based cooling devices ensures the design of innovative microcoolers such as solid state thin film thermoelectric coolers (TEC) that will deal with the reliability concerns that emanate from the presence of hot spots (larger thermal gradient) on the substrate of high performance ICs. Recently, thin film TEC is becoming an attractive option mainly due to its compact structure and larger cooling capability [40, 71].

In this introductory review and chronology, the authors have summarized the significant potential benefit of low substrate temperature on scaled device performance both from leakage power concern and drive current enhancement. Additionally the scaling induced bottlenecks that vie with realizing the reduced temperature operation of MOSFET have been systematically addressed, and possible remedies based on device parametric engineering have been identified. The authors also briefly discussed the various cooling device solutions that have been developed by the microelectronic professionals to implement at the device manufacturing level. The next chapters onwards will concentrate on modeling solutions from the perspective of lower substrate temper-

ature operation on threshold voltage control and its variability assessment and comparing the threshold voltage distribution profile in typical 100K–500K substrate temperatures distribution for various technology nodes considering mostly the long channel *n*-MOSFET case and projecting the tolerable threshold voltage window down to 10 nm technology node by judicious analysis of data from simulations performed on long channel *n*-MOSFET counterparts.