RC24006 (W0607-057) July 14, 2006
Physics

IBM Research Report

A Perspective on Today's Scaling Challenges and Possible Future Directions

Robert H. Dennard, Jin Cai, Arvind Kumar
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

A Perspective on Today's Scaling Challenges and Possible Future Directions

Robert H. Dennard, Jin Cai, and Arvind Kumar
IBM T.J. Watson Research Center, Yorktown Heights, NY 10598 USA
dennard@us.ibm.com

Abstract

Progress in scaling of MOS transistors and integrated circuits over the years is reviewed and today's status and challenges are described. Generalized scaling is updated for the present leakage-constrained environment to project results of continued scaling at a constant power-supply voltage. Alternatives to achieve energy-efficient operation at lower voltages are discussed. Particular attention is given to threshold variability issues and to the design challenges in reducing and controlling variability using back-gate devices. The importance of the depth of the inversion layer below the silicon surface as a limit to the effectiveness of gate-insulator scaling is illustrated by a design study. Low-temperature operation is considered as a possible future direction for continuing scaling.

1. Introduction

Scaling of microelectronic devices and circuits to smaller and smaller dimensions has been amazingly successful since the first scaling principles were introduced in the early 1970's [1]. Since then the key device dimensions including the effective gate insulator thickness have been reduced more or less by a factor 100. Many challenges have been met to achieve this, but today even more challenges have to be faced if progress is to continue. It is well known that transistor off current now limits further scaling of the threshold voltage, V_T, which in turn limits scaling of the power supply voltage for highest performance applications. Also, because of the growth in gate oxide tunnelling current, gate insulator scaling has come to an end unless a high-κ solution succeeds. Variability problems are increasing due to line edge control and roughness,

doping fluctuations, and soft errors. For the near term, strain engineering and hybrid surface orientation are being pursued to keep performance moving forward. Several alternative structures are promising for the future, but appear challenging to build and only offer incremental benefit in performance.

This paper reviews the generalized scaling principles and updates them to show the effects of constant voltage scaling on power density. It also illustrates the energy versus performance tradeoff for optimum results over a range of supply voltage. It shows that threshold variability exacts a large penalty in energy per computation, and argues that a method to adjust thresholds to the optimum value can have a large impact on future system-level performance. Then it reviews the potential of a back-gated fully-depleted thin silicon device to provide this adjustment. It also shows design results aimed at reducing doping fluctuations in such devices, which illustrate the basic design constraints. Particular focus is given to a study illustrating how the confinement of the quantized weak inversion layer in the turned-off device varies with the choice of gate workfunction, how this affects the short-channel characteristics, and how this poses a limit on the effectiveness of scaled gate insulators. Finally it discusses the possible role of low-temperature operation in ultimate integrated silicon devices.

2. Review and Update of Generalized Scaling

Our concept of scaling in the deep submicron CMOS era from a decade ago until recently is called generalized scaling, which is illustrated in Table 1. It has been broadened from the original where the electric field was kept constant and the devices and wires were scaled together. Most device physical dimensions are divided by a factor of a_D, while the electric field is allowed to be multiplied by a factor ε so that voltage can be reduced more gradually than the device dimensions [2]. The wiring dimensions and

the device width are divided by a factor α_W [3]. Even if the electric field factor ε increases, for some time it has been thought that a reasonable goal is to increase the circuit speed by a factor α_D. This assumes any tendency to increase the average carrier velocity because of the higher lateral field is offset by mobility reduction from the higher vertical field and increased limitation effects of parasitic resistance and capacitance. At that speed, the active power for a given circuit scales as $\varepsilon^2/\alpha_D\alpha_W$ while the power density scales as $\varepsilon^2\alpha_W/\alpha_D$, assuming the density is dominated by the interconnections and accordingly varies with α_W^2.

Table 1
Generalized scaling approach

Physical Parameter	Generalized Scaling Factor
Gate Length, L	$1/\alpha_D$
Gate Insulator, t_{ox}	$1/\alpha_D$
Voltage, V	ε/α_D
Wiring Width	$1/\alpha_W$
Channel Width, W	$1/\alpha_W$
Circuit Speed (goal)	α_D
Circuit Power	$\varepsilon^2/\alpha_D\alpha_W$

Thus it is seen that power and power density are vitally affected by the electric field factor, ε. A plot of ε as a function of channel length for high-performance MOS technology, given in Fig. 1, was prepared from personal knowledge and archives of the authors. This shows how ε has increased rapidly through the history of scaling down channel length. Part of the increased field is clearly associated with the transition to CMOS and the desire to maintain a 5V power supply as long as possible. The trend line over many generations shows that ε is proportional to $1/\sqrt{L}$, and thus V is proportional to $\sqrt{L}$. We believe this trend arose to maintain smooth performance growth with scaling

by reducing V and V_T gradually, while avoiding the rapid growth in leakage power if V_T were scaled more rapidly.

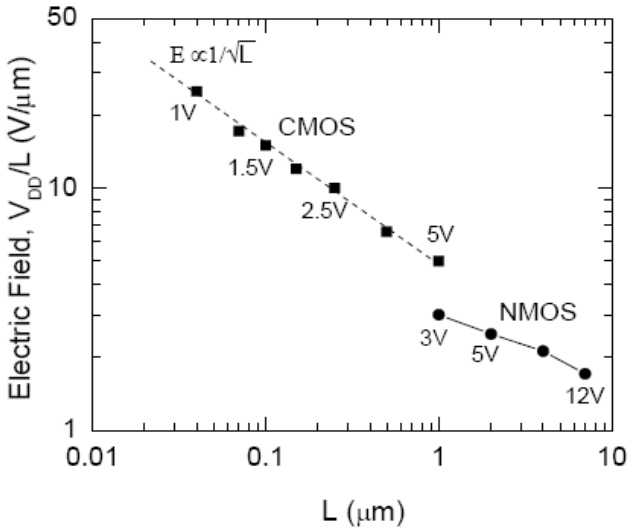

Fig. 1. Evolution of electric field strength for high-performance MOS technology as a function of channel length.

The generalized scaling relationships of Table 1 assumed in the past that the device leakage was not significant. However, at the 90nm generation with gate lengths in the order of 50-70nm, the point has been reached for high-performance CMOS with a supply voltage in the order of 1-1.2V where the leakage power at high operating temperature for worst-case (low) threshold voltages is a significant part of the total power. This represents a point where the V_T has reached an optimum value for this particular supply voltage. For the next generation, scaling the voltage lower and the V_T lower would result in higher total power for the given performance compared to keeping the supply voltage and V_T the same as in the previous generation. In fact, if the power supply voltage is reduced, the optimum V_T for operation at that voltage is actually higher [4] and the optimum performance for operation at that voltage level must decrease accordingly.

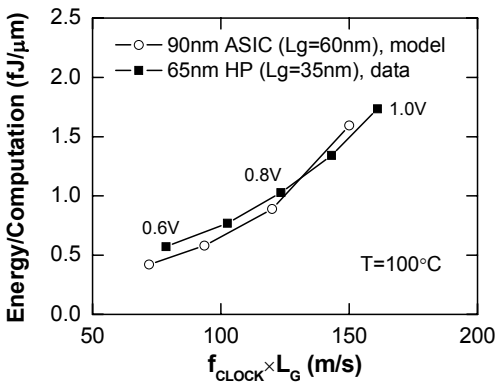

Fig. 2. Energy operation normalized by device width vs. performance normalized by gate length for 20 FO=4 inverters between latches.

Fig. 2 gives a modelled result of energy per operation versus performance for a 90nm ASIC technology, where the V_T has been optimised at each supply voltage for a projected logic switching activity (10% of the clock frequency) following the methodology of [5] based on 20 stages of inverters with fan-out of 4 between latches. A measured result for early 65nm high-performance logic technology is shown for comparison, where the axes are normalized as shown. In this measurement V_T increases as the supply voltage is reduced due to reduced DIBL and fortuitously maintains optimum balance between ac and dc energy consumption. Because the optimisation is fairly flat over a broad range of dc/ac energy, measurements like this are insensitive to the details. Curve fitting shows that energy per operation varies with $V^{2.5}$ in this experiment. This is because the switching energy, often expressed as CV^2, is affected by the nonlinearity of the capacitance. The intrinsic charge transferred in a switching event is related to V- V_T, and here V_T increases somewhat as V decreases.

If dimensional scaling continues in the future without voltage scaling, a set of constant voltage scaling rules can be derived (for any given voltage) by setting $\varepsilon=\alpha_D$ in Table 1, with the results shown in Table 2. Again an important assumption is that speed

increases directly proportional to the device scaling factor α_D. It is seen that power/circuit becomes constant if wires and devices are scaled at the same rate ($\alpha_D=\alpha_W$). Power density then increases by α_D^2 which presents a severe cooling challenge.

Table 2
Constant voltage scaling results

Density varies with α_W^2
Speed varies with α_D
Power/circuit varies with α_D/α_W
Power density varies with $\alpha_D\alpha_W$
Energy/operation varies with $1/\alpha_W$

(Note that leakage current per device goes up approximately as $C_{ox}W/L$ [6]. This means leakage power per circuit scales as α_D^2/α_W if V and V_T are constant. Thus V_T needs to increase slightly with α_D to maintain optimum balance with the ac power which varies as α_D/α_W.)

Energy per operation (power delay product) in this scenario only improves to the degree the wire size is scaled. This assumes that the average wire capacitance is reduced accordingly, as wire lengths are reduced. It should be noted that increasing use of repeaters to minimize wire delay subtracts from the energy saving due to smaller, shorter wires. It is now well known that scaling wiring to dimensions approaching the electron mean free path causes a significant resistivity increase due to scattering at the wire surfaces. Along with surface roughness and grain boundary effects this leads to the reported measured results in Fig. 3, compared to a theoretical prediction for ideal surface scattering [7]. Since larger wires are used in the wiring hierarchy for longer interconnections, this problem appears to have significant impact only after several generations. However, current density in the wires will rise the same as the power

density by $\alpha_D\alpha_W$ as scaling continues if the voltage is kept constant for highest circuit performance. Thus, electromigration can become a serious concern for this scenario.

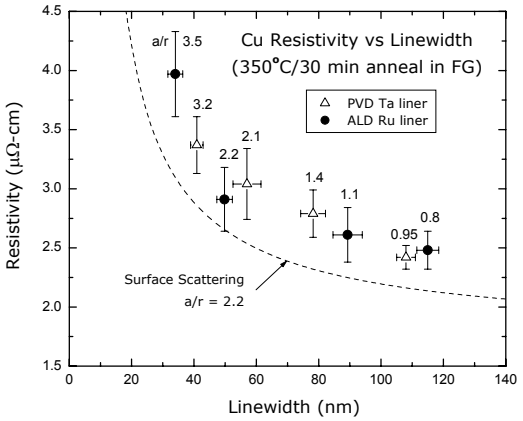

Fig. 3. Measured Cu line resistivity vs. linewidth for two liner processes, compared to ideal model with no liners.

3. Energy/Performance Considerations

As future scaling continues, as Fig. 2 illustrates, the clock frequency hopefully increases directly with the device scaling factor for a given voltage. Energy per computation is normalized in this plot to femtojoule (fJ) per micron of device width, W, to make the curves overlay, and it will thus be reduced by the wire scaling factor (as W required to drive the shorter wire is reduced) as shown in Table 2.

If increasing the individual processor speed is not attractive because of the increased power density and current density, it may be desirable to reduce the supply voltage and lower the energy/computation. A number of energy efficient processors could be placed on the same chip with much lower power density and current density, and possibly with greater net computation throughput depending on the system configuration and I/O bandwidth. Taking advantage of low voltage operation is not easy. The data in Fig. 2 show that the performance is very sensitive to the power supply voltage

at low voltages. Clearly it is also sensitive to threshold voltage variation, and this is shown in Fig. 4.

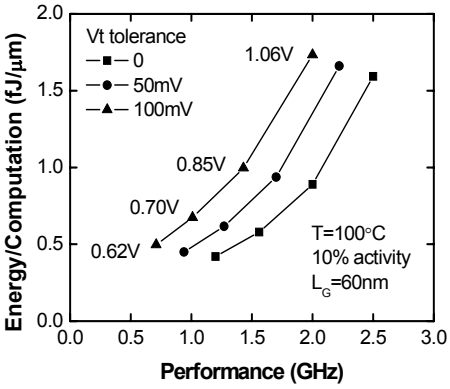

Fig. 4. Worst-case energy vs. performance taking threshold voltage tolerances into account.

The curve on the right in Fig. 4 illustrates energy vs. performance for an optimised situation with no considerations for V_T tolerance. Practically, today's manufacturing processes have a 10x spread in leakage current which represents about 100 mV of V_T variation at high operating temperature. Raising both p and n device thresholds by 100 mV to keep the worst-case leakage from exceeding the allowable value gives the performance result in the curve on the left. The energy/computation for this curve is the worst-case energy for a leaky low V_T chip running at this worst-case speed for a high V_T chip. It can be seen that a chip without tolerances could be operated at a lower voltage with 2x lower energy/operation at any given performance compared to the curve with 100 mV tolerance.

This suggests a strategy of adaptive bias control of body or back-gate potentials to tune out systematic threshold variations in suitable portions of a chip to attain a target V_T or a target performance at the lowest possible voltage. At low supply voltage, Fig. 4 shows nearly 2x performance difference at a given worst-case energy/operation. It can be seen that merely adjusting the power supply voltage adaptively can provide some of the

same benefits. This adjustment (with chip sorting) is being widely used today, but may be difficult to do for future complex systems with many processors per chip, and it cannot compensate for independent p and n variations. Also, statistical fluctuations in today's small devices make SRAM stability unacceptable at low supply voltages, a problem which is made worse by further scaling.

Assuming these variability issues can be solved by new device design approaches (as considered in the next section) and other technology challenges are met, the projected results of scaling are shown in Fig.5 for three generations of interest. These curves are derived by simply applying the scaling relationships of Table 1 to the rightmost curve of Fig.4, which is taken to represent the 90nm generation with a total n and p gate width of 3μm for a basic inverter, assuming that the device and wire dimensions both scale down by $\sqrt{2}$ each generation. It is to be noted that some of the speed improvement in practice is being gained by stress engineering and less by actual insulator and channel length scaling.

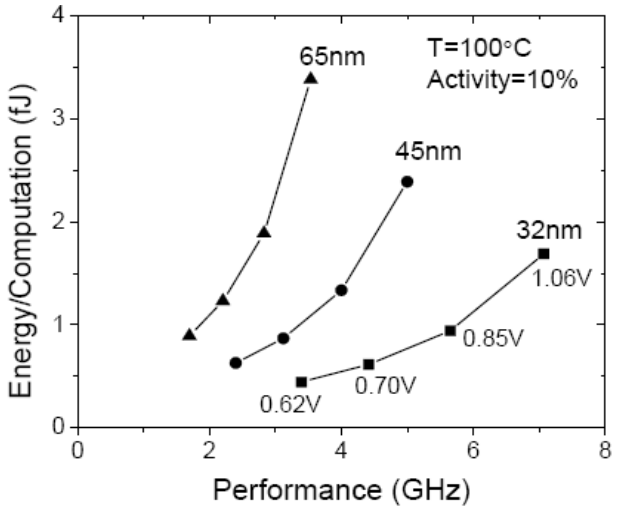

Fig. 5. Projected energy/operation for an F0=4 inverter vs. performance and voltage for indicated technology nodes assuming variability is controlled and scaling challenges are met.

Although the energy/operation at a given voltage decreases linearly with the wire scaling factor, the increased frequency and density lead to significantly increased power density as shown in Fig. 6, amounting to a factor of 4 increase in 2 generations of scaling. As noted previously the current density in all wires (assuming layouts are merely scaled) will increase by the same amount. This problem can be dealt with in a number of ways depending on the application. One way is to use innovative packaging approaches such as liquid cooling in microgrooves on the back surface of the chip [8]. Another is to change the system architecture to choose a design point which trades off some peak performance to lower the power density to a reasonable level. Figure 6 clearly suggests another alternative showing that a modest decrease of voltage can allow power density to remain constant moving across the plots from one generation to the next. Thus from 65nm to 32nm generations it is possible to place 4 times as many processors on a chip with no change of architecture and no increase in the total chip power, having each processor improve in speed by about 32 %.

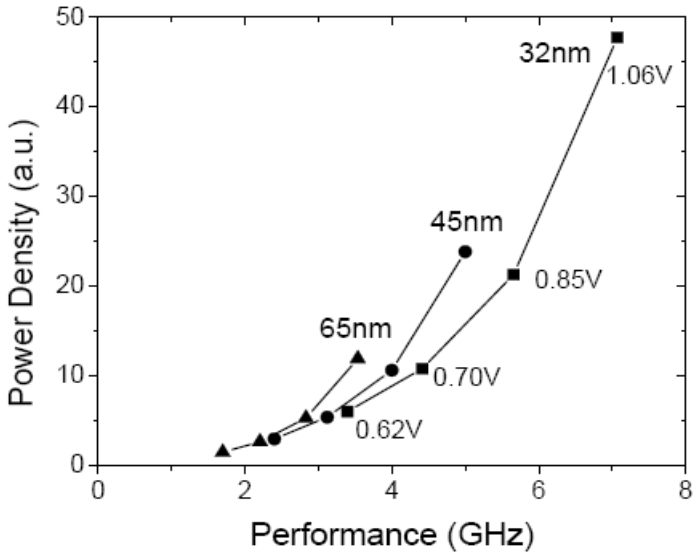

Fig. 6. Projected power density corresponding to Fig. 5.

The curves in Fig. 5 are repeated in Fig. 7 to further illustrate the various possible voltage scaling scenarios in terms of energy efficiency. Included here is a possible scenario to keep processor speed constant as scaling proceeds past 65nm. It is seen that two generations of scaling allow the potential to reduce energy/operation about 7x without loss of speed according to this analysis. Part of this benefit is due to the reduced effective device capacitance with voltage as described previously, which would not apply to circuits dominated by linear capacitance, e.g. long interconnection wires.

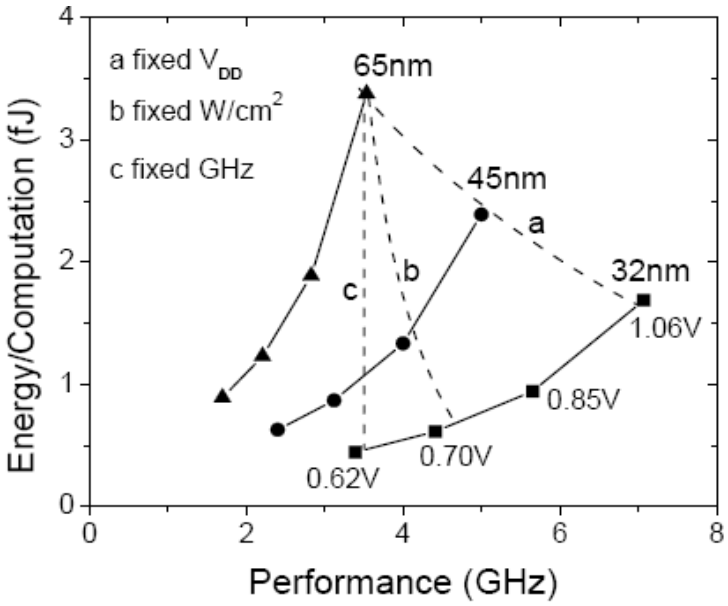

Fig. 7. Illustration of three different scaling scenarios

Clearly the simple picture presented in this section is very approximate and presents many challenges in design and technology. One challenge is that circuits with stacked devices will show faster speed degradation at reduced voltage than the simple inverters shown here, and some circuit redesign may be called for. Circuits optimised for lower activity with higher V_T also are challenging. SRAM has both these difficulties, but much work already under way to improve stability and yield can be applicable to lower

voltage operation [9]. The device work discussed in the next section to address variability issues will be key both to reducing voltage and to further scaling devices.

4. Design Issues with Back-Gated Thin SOI CMOS

A fully-depleted thin SOI structure with a back gate (Fig. 8) offers a fairly ideal device to optimise performance of CMOS processors at low supply voltage. The back gate can be used both to provide the adaptive control discussed in the previous section and to set the threshold voltage without body doping to avoid statistical V_T variations in small W devices. The present ITRS SOI thickness of 10nm for a gate length of 25nm is chosen here to illustrate design issues, using a 1.15nm gate oxide thickness. For a back gate without self alignment the buried oxide (BOX) needs to be thick to avoid parasitic capacitance to the drain but thin enough to give reasonable control voltage levels. A BOX thickness of 10nm is used in this study.

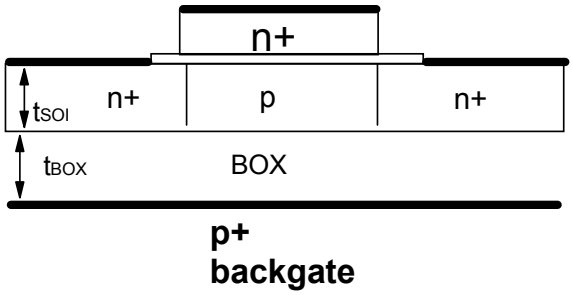

Fig. 8. Schematic of a fully-depleted thin SOI structure with a back gate.

Since halo implants are normally used to control V_T rolloff in short-L devices, eliminating body doping can be expected to give worse rolloff behaviour. Results of a study using a semi-classical drift-diffusion simulator including quantum-mechanical corrections are shown in Fig. 9, where halo doped devices with two different Gaussian implant profiles (σx=10nm and 20nm) are compared with an undoped device.

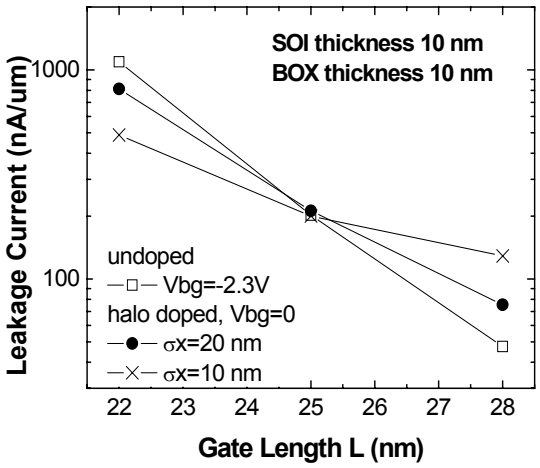

Fig. 9. Off-state leakage current as a function of gate length in back-gated structure, comparing undoped body to halo-doped body with two different Gaussian implant profiles. T=25°C and V_{DD}=1V.

All devices are designed to meet a leakage target at Lgate = 25nm, which is considered to be a 3-sigma short device for a 28nm nominal process, using zero voltage on the p+- doped back gate for the halo cases and a negative bias for the undoped device. The V_T -rolloff behaviour of the halo-doped devices, judged against the criterion that the leakage increase be less than 10x from the nominal gate length to the 6-sigma short gate length, is acceptable especially for the more abrupt implant. However, the undoped device with an n+-poly gate and a heavily doped p-type back gate requires a fairly large backgate voltage, V_{BG} = -2.3V, to achieve the right leakage and has more V_T rolloff. It was found that increasing the L by 4 nm gives acceptable rolloff and the required backgate voltage magnitude decreased somewhat.

A change to a more midgap workfunction gate material is another possible way to set V_T without body doping. This can be done in a thin SOI device without a back gate, or a work function change can be used with a back-gated device to reduce the magnitude of the back-gate voltage required to set the off current. It is well documented that an

undoped SOI device without a back gate can suffer severe short-channel behavior because there is no electric field from the depleted dopant atoms to confine the weak inversion layer in an off device toward the top surface [10]. Instead, fringing field lines from the drain can confine the weak inversion layer toward the back interface. To study this issue in back-gated devices, simulations were done comparing the previously discussed n+-poly gate design to metal-gate designs with workfunctions ¼ bandgap below band edge (QG) and at midgap, respectively. The results in Table 3 show the increased $L_{3\sigma}$ necessary to meet the rolloff criterion described above, the required backgate voltage to meet the leakage target at that length, and the degradation in subthreshold slope and DIBL (in spite of the elimination of poly depletion in the metal gate cases). This degradation is due to the loss of confinement and the resultant spreading of the weak quantized inversion layer in the turned-off device as the electric field from front gate to back gate is reduced (or reversed) for different gate work functions (see Fig. 10 in next section). It can be said that the greatly reduced capacitive coupling from the front gate to the weak inversion layer, compared to the capacitances from the drain and source, is responsible for the increased short-channel effects. The quarter gap (QG) metal gate case has only a modest degradation of short-channel behavior and greatly reduces the required back gate voltage.

Table 3
Results of design study for undoped devices with different gate workfunctions

	$L_{3\sigma}$ (nm)	Vbg (V)	SS(mV/dec)	DIBL (mV/V)
n+ poly	29	-1.65	83	72
QG	30	0.32	98	86
midgap	32	1.4	114	125

5. Carrier Confinement and Quantization Effects

Because the effectiveness of scaling the gate insulator is intimately linked to quantization of the inversion layer, further studies were done on a similar structure to the one studied above using a fully quantum-mechanical transport solver [11]. Both undoped and uniformly doped bodies were used and confinement was varied by changing the metal gate workfunction in $^1/_8$ bandgap increments. As the gate work function increases from band edge toward midgap, V_{BG} must be made more positive (undoped body) or N_A must be reduced (doped body) in order to achieve the I_{off} target, in this case 200 nA/um at temperature 100°C at low drain voltage. The top oxide and SOI thicknesses are 1 nm and 10 nm, respectively and the gate length is 25 nm. Gate leakage is turned off, for simplicity, but oxide penetration of the wave function is included.

The density of electrons in the weak inversion layer as a function of position below the top interface is shown in Fig. 10a for various workfunction values. This measurement is made along a vertical cut in the middle of the device and the drain voltage is kept low to avoid two dimensional effects. Fig. 10b shows the potential along the vertical cut, which approximates an "ideal" triangular potential well. It is clear that the reduction in confining field, as the work-function shift increases, causes the weak-inversion charge to spread out and move toward the center. As the electric field reverses, the charge largely moves through the center and is confined toward the back interface.

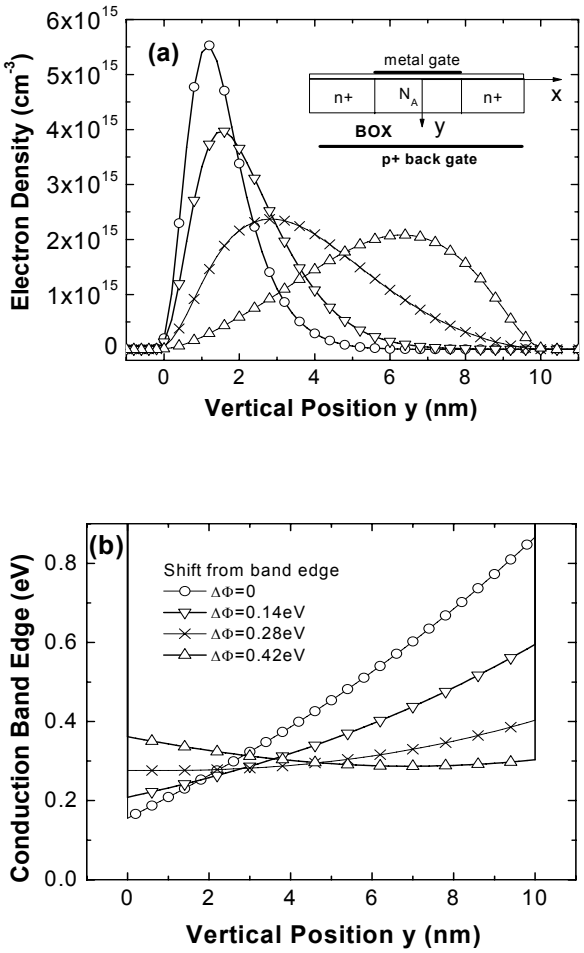

Fig. 10. (a) Electron density and (b) confining potential along a vertical cut in the SOI as function of position. Inset: Schematic of back gated UTSOI *n*FET with body doping N_A and *p*+ back gate at voltage V_{BG} used in this work.

In Fig. 11 we plot the effective vertical electric field, defined as the local field weighted by electron density, *n(y)*, $F_{eff} = \int F(y)n(y)dy / \int n(y)dy$, for the undoped and doped cases, evaluated at the channel center (x=0). Channel doping generates a somewhat weaker effective field than back-gate bias with an undoped body. As the work-function moves away from the band-edge, direct consequence of the reduction in the gate-channel coupling stemming from the loss of confinement is a degradation of the subthreshold swing, also shown in Fig. 11.

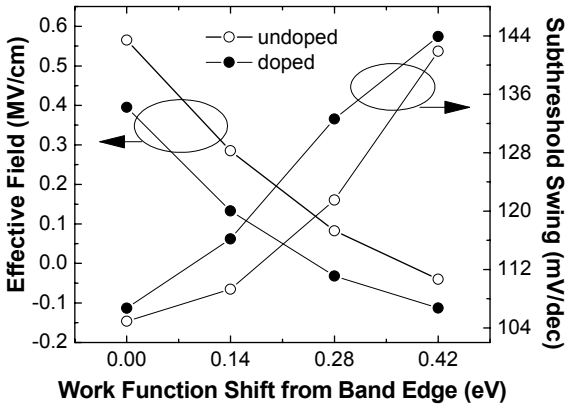

Fig. 11. Effective confining electric field and linear subthreshold swing at 100°C as a function of work function shift from the band edge.

Fig. 12 shows the position of the centroid of the electron distribution from the top interface at the channel center (x=0) as a function of areal electron density as the top gate voltage V_g is swept from the off-state (V_g=0 V) to the on-state (V_g=1 V). Even in strong inversion, shifting the work function away from band edge results in lower carrier density and a centroid farther from the interface. Also noteworthy is the significant difference in centroid position between the off-state and the strongly inverted state, which increases as $\Delta\Phi$ increases. The quantity $t_{inv,}$ measured in strong inversion, is often used to characterize the effectiveness of an insulator. However, the important short-channel characteristics are affected by the centroid position in the turned-off device, which can be converted to an effective oxide thickness (EOT) and added to the EOT of the gate insulator to obtain a total EOT that will be called t_{off} in this paper. We assert that t_{off} is a useful measure of how well a scaled insulator can control electrostatic behavior important to short-channel characteristics.

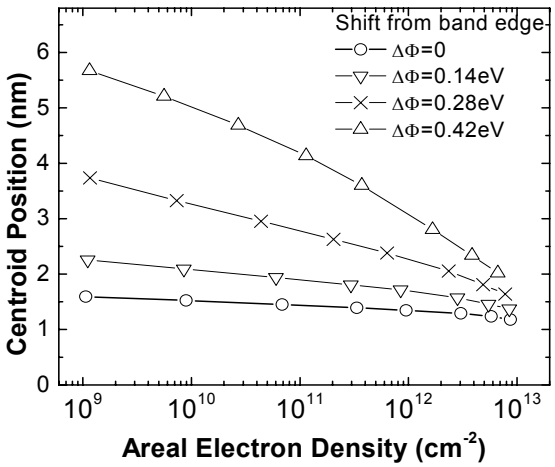

Fig. 12. Centroid position as a function of areal electron density in the channel, spanning from the off-state (leftmost) to the on-state (rightmost).

Fig. 13 plots the centroid position as a function of effective electric field for the undoped and doped cases and the four different work functions. A nearly universal correlation between effective confining field and centroid distance from the interface is observed. Both the undoped case with $\Delta\Phi$=0.42 eV and the quarter gap doped case ($\Delta\Phi$=0.28 eV) have approximately zero effective confining field, and their centroids are located nearly midway in the SOI body as a result of wave function repulsion from the two oxide barriers.

As a limiting case of high gate-channel coupling, we also consider the effect of increasing the top oxide dielectric constant from κ=3.9 to κ=7.8 in the undoped device with $\Delta\Phi$=0. To meet the off-current target, stronger confinement is required, as reflected by an increase in $| V_{BG} |$ by 1.35 V. Despite this stronger confinement, the centroid is still 1.35 nm from the top interface, corresponding to an effective oxide thickness of 0.45 nm that must be added to the 0.5 nm of this ultrathin effective gate dielectric to give t_{off} =0.95nm.Also, the stronger confinement in the “off” condition carries through to the “on”

condition and will affect the mobility, so a design with less confinement (ΔΦ positive by some amount) may be better. Thus the position of the centroid of the weak-inversion charge is seen as a major constraint on how far CMOS scaling can go with the device types in common use today even with high-k gate insulators. Structures with inherently better electrostatics, e.g. very thin SOI with double or wrap-around gates, are ultimately required if the practical difficulties with such devices can be solved.

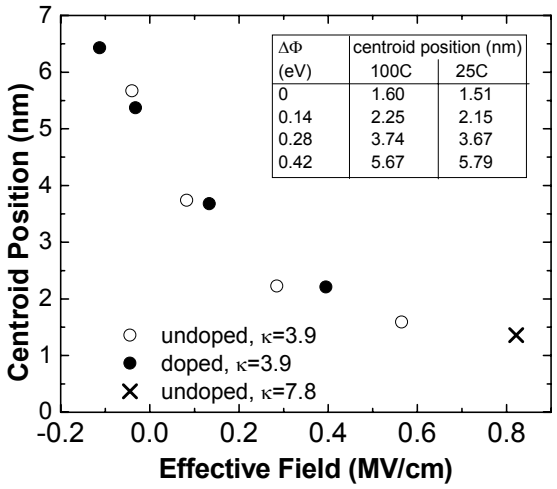

Fig. 13. Centroid position in off-state as a function of effective electric field at 100C. Inset shows shift in centroid position between 100°C and 25°C for undoped cases with k=3.9.

6. Potential of Low-Temperature Operation

Although cooling of CMOS to low temperature has many advantages and has been used to a limited extent in mainframes for performance and reliability improvement, it has never had broad application. Many of the present problems in scaling CMOS could be avoided and better performance achieved if absolute temperature, T, were scaled down along with dimensions in future CMOS generations. The benefits of cooling CMOS circuits are well known [12]. In the past this has been seen as a performance improvement, as much as 2x at 77K, due to greatly improved mobility, modestly greater saturation velocity, and improved conductivity in silicide and metal layers. In the present

environment, scaling the operating temperature would allow the threshold voltage to be scaled down along with dimensions and supply voltage (constant-electric field scaling) without increasing the device "off" current on a per square (W/L=constant) basis. This capability is illustrated in Fig. 14 which shows simulated characteristics of a 65nm-generation device (NFET 1) at 100°C and at -50°C versus a device (NFET 2) designed for and operated at -50°C. The much sharper turn-off behavior at low temperature, as characterized by the reduced subthreshold slope, is seen in NFET 1 but the threshold increases substantially. NFET 2 was optimized by reducing the halo dose so that it has the same leakage current at -50°C as the regular device at 100°C. The increased "on" current due to low-temperature operation, and further enhanced by the design optimization, is seen on the right-hand scale.

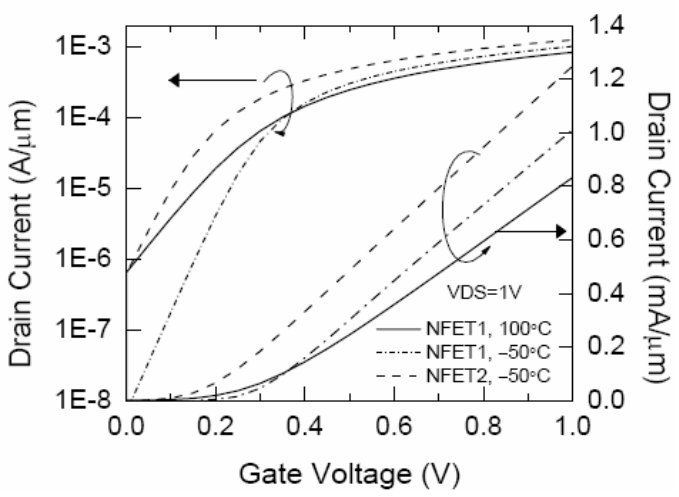

Fig. 14. Simulated Id-Vg characteristics for NFET1 at 100°C and -50°C, and NFET2 at -50°C. Channel doping is lowered in NFET2 to match the 100°C off current of NFET1.

Recent experimental work carried out to build and measure CMOS test circuits optimised for -50°C operation as discussed here gives the results shown in Fig. 15. The

improved subthreshold slope and higher mobility allow operation at much lower voltage without loss of performance and with much lower power. It is seen that the power-delay product improves by about 2.5x in this experiment. This could be very important in future ultimately-scaled CMOS in allowing very densely packed systems with shorter wires, and its advantages may offset the complexity and power consumption of the cooling system.

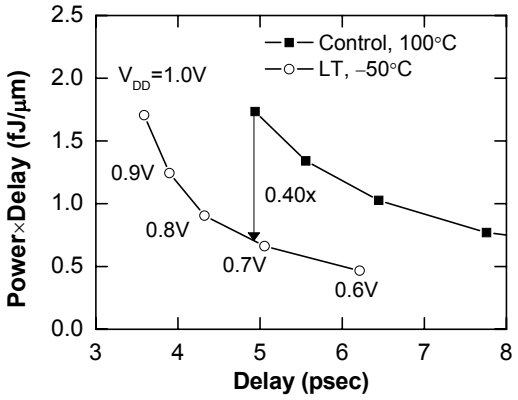

Fig. 15. Energy vs. delay data at various supply voltages for high performance 65nm node CMOS operated at 100°C and for re-optimised CMOS operated at -50°C.

It has been observed that band-to-band tunnelling is easier to avoid at low-temperature because of lower voltage and because the barrier height and the confining field needed to turn off the device both reduce with temperature. Thus band-to-band tunnelling should disappear in a fully-depleted device as the voltage is lowered below about 0.7V. On the other hand direct tunnelling through the lower barrier in the turned-off device limits how far the device can be scaled and still maintain the improved turnoff behaviour.

7. Conclusion

The present trend to scale technology for high-performance processors to smaller and smaller dimensions without reducing power supply voltage is difficult to sustain due to increasing power density and current density. Operating at lower voltage would offer

relief for these problems and much lower energy per computation, but the principal challenge to this is threshold variability. The ultimate silicon device may be one that minimizes V_T variability and/or allows adaptive control to adjust V_T to the optimum level. A fully-depleted thin SOI device with a back gate is promising for its ability to provide such an adjustment, and also can be designed without body doping to avoid random V_T fluctuation. The importance of quantization of the weak inversion layer of turned off devices on the short-channel behaviour is reconfirmed in this study and put in perspective as a limit of the effectiveness of scaling high-κ gate insulators. Low-temperature operation allows a path to low voltage without loss of performance, and it offers the possibility to remove band-to-band tunnelling as a constraint on future scaling.

Acknowledgement

The authors are grateful to S.E. Laux for his support of QDAME and to W. Haensch for helpful discussions.

References

[1] Dennard RH, Gaensslen FH, Yu HN, Rideout VL, Bassous E and LeBlanc AR. Design of ion-implanted MOSFETs with very small physical dimensions. IEEE J Solid-State Circ 1974;9(5):256-68.

[2] Baccarani G, Wordeman MR and Dennard RH. Generalized scaling theory and its application to a 1/4 Micron MOSFET design. IEEE Trans Electron Devices 1984;31(4):452-62.

[3] Davari B, Dennard RH and Shahidi GG. CMOS scaling for high performance and low power - the next ten years. Proc IEEE 1995;83(4):595-606.

[4] Frank DJ. Power constrained device and technology design for the end of scaling. IEDM Tech. Dig 2002: 643-46.

[5] Cai J, Taur Y, Huang SF, Frank DJ, Kosonocky S, Dennard RH. Supply voltage strategies for minimizing the power of CMOS processors. Symp VLSI Tech 2002:102-3.

[6] Swanson RM and Meindl JD. Ion-implanted complementary MOS transistors in low-voltage circuits. IEEE J Solid-State Circ 1972;SC-7(4):146-53.

[7] Rossnagel SM, Wisnieff R, Edelstein D and Kuan TS. Interconnect issues post 45nm. IEDM Tech. Dig 2005:95-7.

[8] Tuckerman DB and Pease RFW. High performance heat sink for VLSI. IEEE Electron Dev Lett 1981;EDL-2(5):126-9.

[9] Bhavnagarwala A, Kosonocky S, Radens C, Stawiasz K, Mann R, Qiuyi Ye, Chin K. Fluctuation limits & scaling opportunities for CMOS SRAM cells. IEDM Tech Dig 2005:659-62

[10] Trivedi VP and Fossum JG. Scaling fully depleted SOI CMOS. IEEE Trans Electron Devices 2003;50(10):2095-103.

[11] Laux SE, Kumar A and Fischetti MV. Analysis of quantum ballistic transport in ultra-small silicon devices including space-charge and geometric effects. J Appl Phys 2004;95(5):5545-82.

[12] Sun JY-C, Taur Y, Dennard RH and Klepner SP. Submicrometer-channel CMOS for low-temperature operation. IEEE Trans Electron Devices 1987;ED-34(1):19-27.