Effectiveness of Low Power Dual-V_t Designs in Nano-Scale Technologies Under Process Parameter Variations

Amit Agarwal, Kunhyuk Kang, Swarup K. Bhunia, James D. Gallagher, and Kaushik Roy

School of Electrical and Computer Engineering
Purdue University, West Lafayette, IN 47906, USA

<amita, kang18, bhunias, jdgallag kaushik> @ ecn.purdue.edu

ABSTRACT

This paper explores the effectiveness of dual-V_t design under aggressive scaling of technology, which results in significant increase in all components of leakage (subthreshold, gate and junction tunneling) while having large variations in process parameters. The present way of realizing high-V_t devices results in high junction tunneling leakage compared to low-V_t devices, which in turn may result in negligible leakage savings for dual-V_t designs in scaled technologies. Moreover, increase in process variation severely affects the yield of such designs. This paper suggests important measures that need to be incorporated in conventional dual-V_t design to achieve total leakage power improvement while ensuring yield. It also shows that different process options, such as metal gate work function engineering, are required to realize high-performance and low-leakage dual-V_t designs in sub-50nm technologies.

Categories and Subject Descriptors

B.6.3 [**Logic Design**]: Design Aids – *optimization;* B.7.1 [**Integrated Circuits**]: Types and Design Styles – *advance technology, algorithms implemented in hardware.*

General Terms

Algorithms, Performance, Design and Reliability.

Keywords

Dual-V_t, leakage, yield, process variation, metal gate.

1. INTRODUCTION

Aggressive scaling of CMOS devices to achieve higher integration density and performance, results in exponential increase in subthreshold leakage and worse short channel effects (SCE), e.g. increased drain induced barrier lowering (DIBL), V_t roll-off, and reduced on-current to off-current ratio. To avoid such SCE, oxide thickness scaling and higher and non-uniform doping ("halo" and "retrograde well") needs to be incorporated as the devices are scaled in the nanometer regime. The low oxide thickness gives rise to high electric field, resulting in considerable direct tunneling current (gate leakage). Higher doping results in high electric field across the p-n junctions (source-substrate or drain-substrate), causing significant junction band to band tunneling (BTBT leakage) of electrons [1]. Another leakage mechanism called gate induced drain leakage (GIDL)) which is the product of small transistor geometries and is not a dominant component during regular operations of the circuit. During normal mode of operation, the major leakage currents are subthreshold, gate, and junction BTBT leakage. The increase in different components of leakage with technology scaling has two major implications in logic design. First, leakage reduction techniques are becoming indispensable. Moreover, different leakage mechanisms are becoming equally important with device scaling. Hence, the relative magnitudes of each of the leakage components should be considered in any low-leakage logic design.

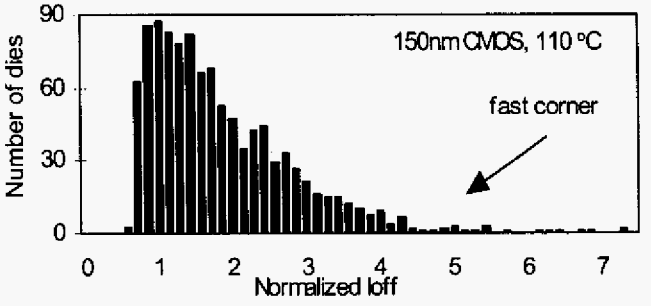

Figure 1. Leakage variation across dies in 150nm technology, (source Intel).

Second, controlling the variation in device parameters during fabrication is becoming a great challenge for scaled technologies. As the delay and the various components of leakage in a device depend on the transistor geometry (gate length, oxide thickness, width, the doping profile and "halo" doping concentration, etc.), the flat-band voltage, and the supply voltage, any statistical variation in each of these parameters results in a large variation in different components of leakage current and significant spread in delay. It has been shown that there can be 20X variation in leakage current in 150nm technology [2] (Figure 1). Hence, any low leakage design needs to consider the spread of leakage and delay both at the circuit and device design phase to minimize overall leakage, while maintaining yield with respect to a target delay under process variation.

Dual-V_t design technique has proven to be extremely effective in reducing sub-threshold leakage in both active and standby modes of operation of a circuit in submicron technologies. However, with the emerging issues related to technology scaling as

ISLPED'05, August 8–10, 2005, San Diego, California, USA.

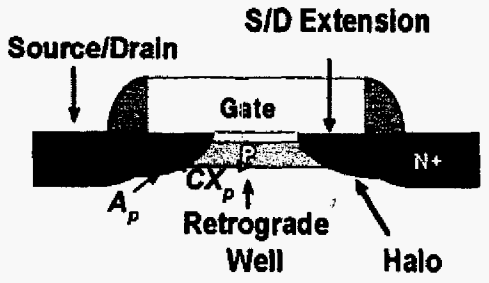

Figure 2. Nano-scaled n-channel device with halo doping.

mentioned above, the effectiveness of conventional dual-V_t design technique [3, 4] may degrade in nano-scale technologies. In this paper, we explore issues related to dual-V_t design in nano-scale technologies and propose device aware solutions. We also show that non-scalability of present way of realizing high-V_t (by changing halo doping) will result in negligible leakage savings and different process options such as metal gate work function engineering might be indispensable in future technologies.

2. EFEECTIVENESS OF DUAL-V_t DESIGN ACROSS DIFFERENT TECHNOLOGY GENERATIOS

2.1 Effectiveness with respect to Leakage Savings

Scaled devices require the use of higher substrate doping and the application of the "halo" profiles to reduce short channel effect. In ultra-scaled technologies, the high halo doping supersedes any change in base channel doping or threshold voltage implants, which were used traditionally to achieve high-V_t devices. In nano-scaled bulk Silicon technologies, high-V_t devices are obtained by changing the peak halo density and its location. In n-channel device the strength of the halo can be increased by: (a) increasing the peak halo doping A_p, (b) moving the position of the lateral peak of the halo (Cx_p) close to the center of the channel and (c) moving the position of the vertical peak of the halo (Cy_p) away from the bottom junction and towards the surface (Figure 2). An increase in the strength of the 'halo' reduces subthreshold leakage and improves short channel effects, however, it increases the junction BTBT due to high electric field across p-n junctions (note that gate leakage is insensitive to halo doping profile). It also increases the junction capacitance. To investigate the effectiveness of dual-V_t design with technology scaling and to achieve optimum low/high-V_t devices, NMOS transistors were designed based on the doping profile and device structure given in [5] and the design guidelines given in 2001 and 2003 ITRS Roadmap for effective gate lengths of 90nm, 50nm and 25nm. The devices were simulated using MEDICI device simulator.

The peak halo density (A_p) along with halo location (Cx_p, Cy_p) was varied to achieve optimum low/high-V_t devices. The oxide thickness, source/drain junction doping, base channel doping and all other device parameters were kept fixed based on ITRS Roadmap and device structure given in [5]. Device optimization was performed by varying halo doping profile while keeping the subthreshold leakage fixed at a desired value. The goal of the optimization was to maximize I_{on}/I_{off}, while maintaining the subthreshold slope within 120mV/decade with reasonable V_t roll-off and DIBL. Here, I_{off} consists of all components of leakage (gate, subthreshold and junction BTBT leakage). Different subthreshold leakage corresponds to devices having different V_t's. Since gate leakage is almost insensitive to change in halo doping profile, by maximizing I_{on}/I_{off} we achieved an optimum device with minimum junction BTBT and highest performance for a given subthreshold leakage (in other words for a given V_t). In this paper we use these devices to show our results on 90nm, 50nm and 25nm effective gate length technologies.

Figure 3 plots the different components of leakage in our optimized low/high-V_t NMOS devices in off state for 90nm, 50nm and 25nm devices at 100°C. It can be observed from the figure that increasing the V_t of the device reduces the subthreshold leakage exponentially, however, it also increases the junction BTBT leakage. The gate leakage is almost insensitive to the change in V_t. In reality, during inversion (on state) an increase in effective channel doping increases the band-bending, thereby increases the gate to channel leakage, but at the same time it also decreases the amount of inversion charge available for tunneling (at same $V_{GS}=V_{DD}$) thereby, decreasing the leakage current. We observed that, the second effect prevails over the first and the gate tunneling current decreases at high-V_t. However, decrease in gate-

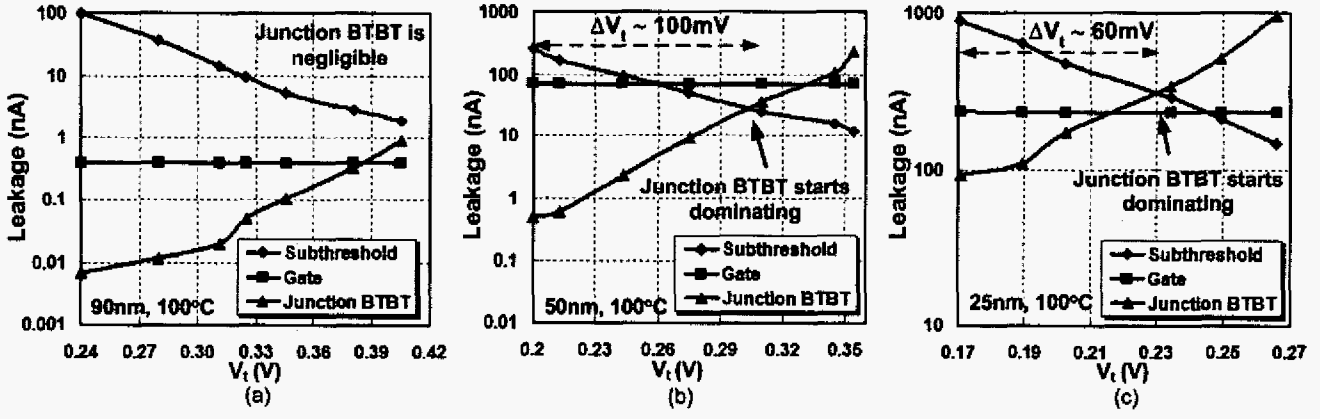

Figure 3. Simulation results of low/high-V_t optimum n-channel devices leakage components a) 90nm, V_{DD} = 1.5V b) 50nm, V_{DD} = 1.2V c) 25nm, V_{DD} = 1.0V.

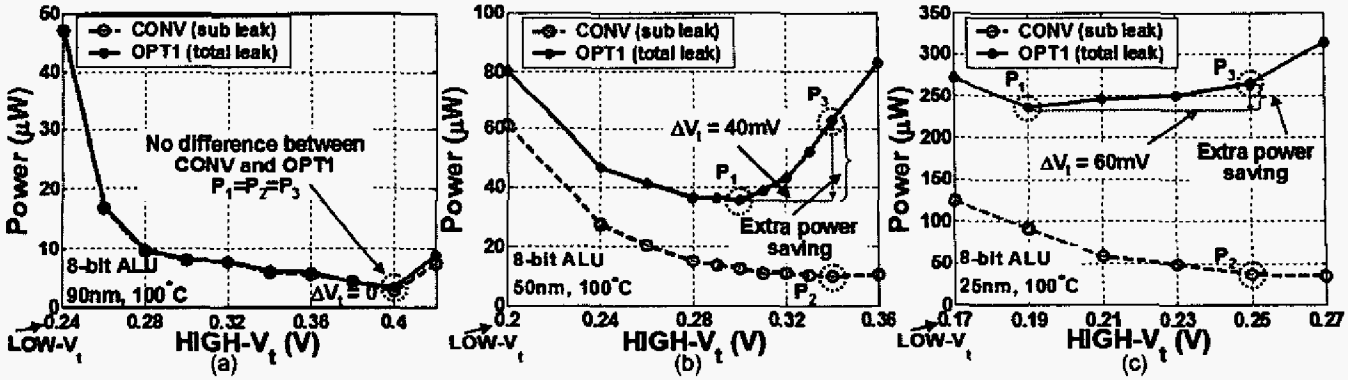

Figure 4. High-V_t assignment using CONV and OPT1 a) 90nm b) 50nm c) 25nm technology. P1: Leakage power using OPT1, P2: Expected leakage Power using CONV, P3: Actual leakage power using CONV.

leakage is negligible compared to increase in junction BTBT leakage. Hence, any reduction in subthreshold leakage because of high-V_t device in dual-V_t design will be at the expense of corresponding increase in junction BTBT leakage, which in the worst case might increase the total leakage. Since 90nm devices do not require strong halo concentration to maintain short channel effect and to meet the required subthreshold leakage, junction BTBT is almost negligible compared to the subthreshold leakage for a wide range of V_t's (Figure 3a). Hence, conventional dual-V_t designs that did not consider junction BTBT while assigning high-V_t, was extremely effective in saving leakage in submicron technologies. However, in a 50nm device, the junction BTBT leakage increases significantly with small change in V_t and becomes comparable to subthreshold leakage at V_t = 0.3V, which is only 100mV higher than the low-V_t (Figure 3b). This difference between low and high-V_t gets smaller (only 60mV) as we go to 25nm technology (Figure 3c). Since the relative magnitudes of different leakage components vary across devices having different V_t's, considering only subthreshold leakage in dual-V_t optimization will overestimate the leakage savings and in the worst case might increase the total leakage. The selection of high-V_t device in nano-scale dual-V_t designs should consider this tradeoff so that total leakage savings is maximized.

An 8-bit ALU is simulated using our optimized devices to estimate the leakage savings achieved by dual-V_t design in scaled technologies. Figure 4 compares the optimum high-V_t and the leakage savings achieved by conventional dual-V_t design (CONV) and OPT1 (our methodology). Here CONV only considers subthreshold leakage as the optimization criteria, while OPT1 considers all components of leakage. For 90nm technology both design techniques select the same optimum high-V_t and results in around 90% leakage savings (Figure 4a). However, for 50nm technology, optimum high-V_t's selected by CONV and OPT1 differ by 40mV (Figure 4b). Figure 5 analyzes the different power components in CONV for the 50nm node. It shows that even though the subthreshold leakage power is minimum at V_t = 0.34V, the total leakage minima occurs at a smaller V_t due to increase in junction BTBT. The gate leakage and dynamic power do not change significantly across different V_t's and depend on the size of logic gates. If we include junction BTBT and gate leakage power at optimum V_t point (P2) in CONV curve (Figure 4b), the total power (P3) actually exceeds the minimum power (P1) achieved by OPT1 by 35%. Hence, OPT1 achieves more leakage power savings compared to conventional approaches. It also shows that CONV overestimates the total power saving by 65% and only saves 20% of total leakage. Even though OPT1 results in higher junction BTBT leakage compared to low-V_t design, it saves more than 55% of total leakage. However, in 25nm technology, due to significant increase in junction BTBT, dual-V_t design using CONV results in negligible leakage saving, while OPT1 results in only 14% leakage saving (Figure 4c). Moreover, the difference between low-V_t and optimum high-V_t for OPT1 is only 20mV. Such dual-V_t's will be difficult to fabricate accurately considering the large process variation in nano-scaled technologies.

It is evident from the above results that dual-V_t designs should consider each component of leakage while optimizing circuits to reduce total leakage power. Since, increasing peak halo doping to realize high-V_t devices increases junction BTBT leakage, resulting in negligible leakage savings, a different design option, to realize high-V_t's needs to be explored to maintain the effectiveness of dual-V_t design in nano-scale technologies.

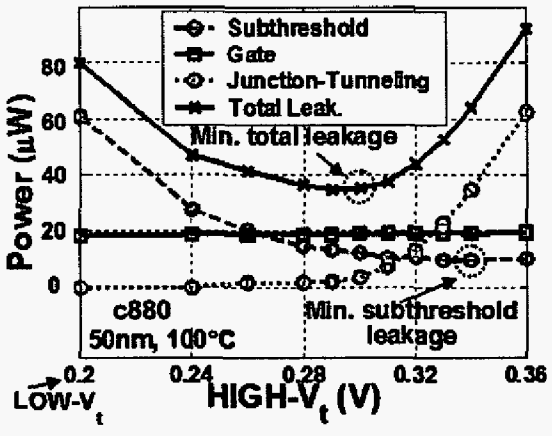

Figure 5. 50nm technology leakage power components in CONV.

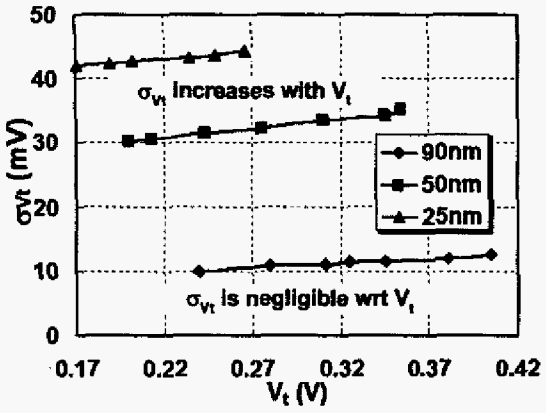

Figure 6. σ_{Vt} due to random dopant fluctuation vs. V_t.

2.2 Yield Loss

It has been observed that as the number of critical paths on a die increases, within-die delay variation causes both mean and standard deviation of the die frequency distribution to become smaller, resulting in reduced performance [6]. Since the idea behind dual-V_t design is to utilize the slack between off-critical and critical paths for high V_t assignment, in effect, it increases the number of critical paths in a circuit. This, in turn, increases the mean of the circuit delay distribution. Since circuits are designed to meet certain delay constraint, any increase in the mean of circuit delay distribution increases the number of dies failing to meet the delay boundary, and hence, results in reduced yield. Moreover, devices with different V_t's will have different process variation spread. A high-V_t device is expected to have large σ variation due to high halo doping concentration [1] (more random dopant fluctuation). Figure 6 plots the standard deviation of V_t (σ_{Vt}) due to random dopant fluctuation vs. V_t for 90nm, 50nm and 25nm optimized minimum width NMOS devices. σ_{Vt} depends on manufacturing process, doping profile and the transistor size and is given by [1]

$$\sigma_{V_t} \approx \frac{qT_{ox}}{\varepsilon_{ox}} \sqrt{\frac{N_a W_d}{3LW}} \qquad (1)$$

Where, N_a is the effective channel doping, W_d is the depletion region width, and T_{ox} is the oxide thickness. Since high-V_t devices have high effective channel doping, σ_{Vt} increases with V_t. Figure 6 shows that σ_{Vt} was negligible with respect to nominal V_t in 90nm devices, however, it becomes significant in 50nm and 25nm devices resulting in considerable spread in delay and leakage power. The use of high-V_t exacerbates the impact of process variation. σ_{Vt} for 25nm device verifies our conclusion that in nano-scale technologies, it would be difficult to fabricate exact low- and high-V_t devices with V_t difference of only 20mV, as required by dual-V_t optimization.

Figure 7 plots the circuit delay distributions of an 8-bit ALU obtained using statistical timing analysis tool [7] for the optimum dual-Vt design (high-V_t which achieves minimum power) using CONV, OPT1 and OPT2. Here CONV and OPT1 are as described earlier, and OPT2 takes circuit delay variation into account (95 percentile circuits delay of low-V_t circuit as a constraint in dual-V_t optimization) to ensure yield, while considering all components of leakage, but ignores any leakage variation. In this paper, we only consider the intrinsic fluctuation of the V_t of different transistors due to random dopant effect, which is the primary source of intra-die process variation [8]. For inter-die variation we consider variation in gate length (L_{gate}), usually considered to be the dominant source of inter-die variation. The standard deviation of V_t due to random dopant fluctuation is extracted from our optimized device (Figure 6), which depends on both V_t and width of the transistors. We assume 15% 3-sigma variation in L_{gate} for our analysis. Since in 90nm devices σ_{Vt} is negligible with respect to their V_t, in CONV, OPT1 and OPT2, 95% of the dies were able to meet the required delay constraint (95 percentile circuit delay of low-V_t circuit). However, for 50nm and 25nm technologies, CONV results in only 86% and 80% yield, while OPT1 results in 92% and 84% yield, respectively. Since OPT2 imposes yield constraint with respect to circuit delay variation while assigning high-V_ts, it is able to meet the required 95th percentile delay yield for both 50nm and 25nm technology. Hence, for nano-scale technologies, dual-V_t design should consider the delay distribution of circuit under process variation to ensure yield, while minimizing leakage. The leakage power saving achieved by OPT2 is 50% in 50nm technology. However, it is only 8% in 25nm technology. This shows that in aggressively scaled technology, dual-V_t optimization results in almost negligible power savings, if the yield constraint is forced. Hence, present way of realizing high-V_t devices, which results in higher process variation, may not be suitable in reducing leakage in nano-scaled technologies.

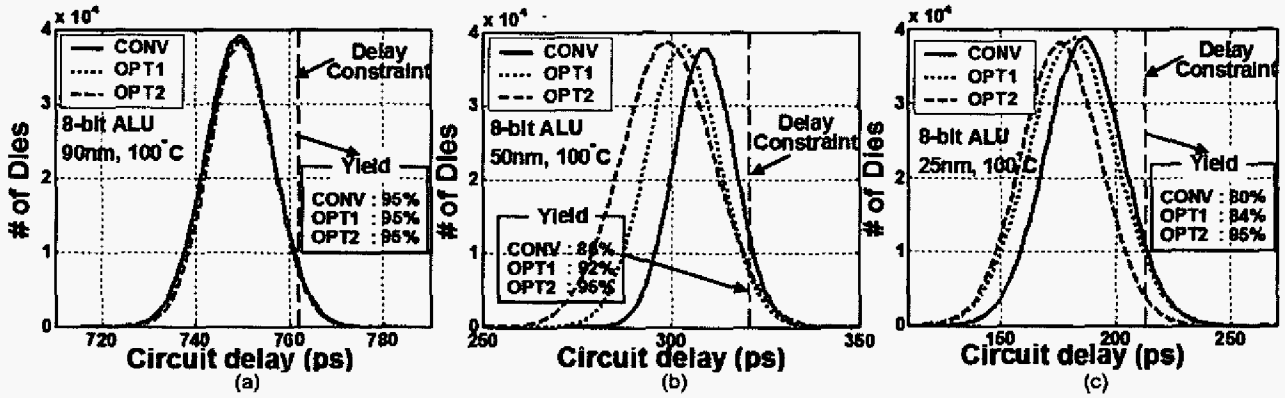

Figure 7. Circuit delay distribution and yield loss using CONV, OPT1 and OPT2 a) 90nm b) 50nm c) 25nm technology.

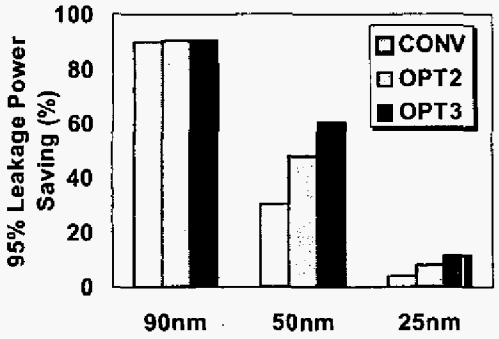

Figure 8. 95 percentile leakage power saving.

2.3 Leakage Distribution under Process Variation

Since circuit leakage follows statistical distribution under parameter variations, any dual-V_t design technique that considers either worst-case or best-case leakage will suffer from an overly pessimistic or optimistic approach. A good dual-V_t design should target probabilistic minimization of leakage considering the effect of process variation on the leakage of devices having different V_t's (high-V_t devices will have large σ).

Figure 8 compares the 95th percentile leakage power savings achieved by CONV, OPT2, OPT3 with respect to 95th percentile leakage power of a low-V_t design of a 8-bit ALU obtained using statistical leakage analysis [9] for 90nm, 50nm and 25nm technologies. Here CONV and OPT2 are the same as described earlier. OPT3 considers everything (all components of leakage and delay variation) and minimizes 95th percentile leakage of the circuit, while doing dual-V_t optimization. OPT3 results in best 95% percentile leakage power, while it ensures yield with respect to circuit delay for all technologies. As expected in 90nm technology, 95th percentile leakage savings are almost same for all the designs due to negligible intra-die process variation. However, in 50nm technology OPT3 results in 30% and 12% extra leakage power saving compared to CONV and OPT2, respectively. This shows the importance of considering leakage variation in dual-V_t optimization. In 25nm technology OPT3 results in 12% 95th percentile leakage saving, which is 3X and 1.5X times higher than the leakage savings achieved by CONV and OPT2, respectively. However, the total leakage power savings compared to low-V_t design itself is negligible.

2.4 Metal Gate and Work Function Engineering

As we expected, high-V_t accomplished by strengthening the halo doping concentration gives rise to a noticeable junction BTBT leakage. This becomes more evident in future nano-scale technologies where a higher baseline halo concentration is needed to suppress the worsening of V_t roll-off and DIBL with device scaling. In technologies where one cannot afford a higher halo doping, high-V_t devices can be realized by using metal gates -- materials with higher work functions -- without impacting the junction BTBT leakage and process variation [10]. Metal gates are being explored not only to have proper control on realizing devices having high-V_t, but also to achieve high performance while maintaining short channel effect. Aggressive scaling of gate length and oxide thickness of devices exacerbates the problems of poly-Si gate depletion, high gate resistance and boron penetration from the p+-doped poly-Si gate into the channel region. The poly depletion increases the effective oxide thickness which in turn reduces the gate capacitance in the inversion regime and hence, the inversion charge density, leading to a lower gate over-drive and thus degrading the device performance. Moreover, poly-Si has been reported to be incompatible with a number of high-k gate-dielectric materials, which are required to maintain reasonable gate leakage.

To show our results, we first designed an optimum low-V_t 25nm device, by varying metal gate work function along with T_{ox}, peak halo density (A_p) and halo location (Cx_p, Cy_p), which meets the ITRS roadmap. The devices having different-V_t's are then obtained by changing the gate work function. Figure 9a plots different leakage components in our optimized low/high-V_t metal gate NMOS devices for 25nm technology at 100°C. It can be observed from the figure that the subthreshold leakage dominates the total leakage in low-V_t devices. Increasing the V_t (by changing the work function) of the device reduces the

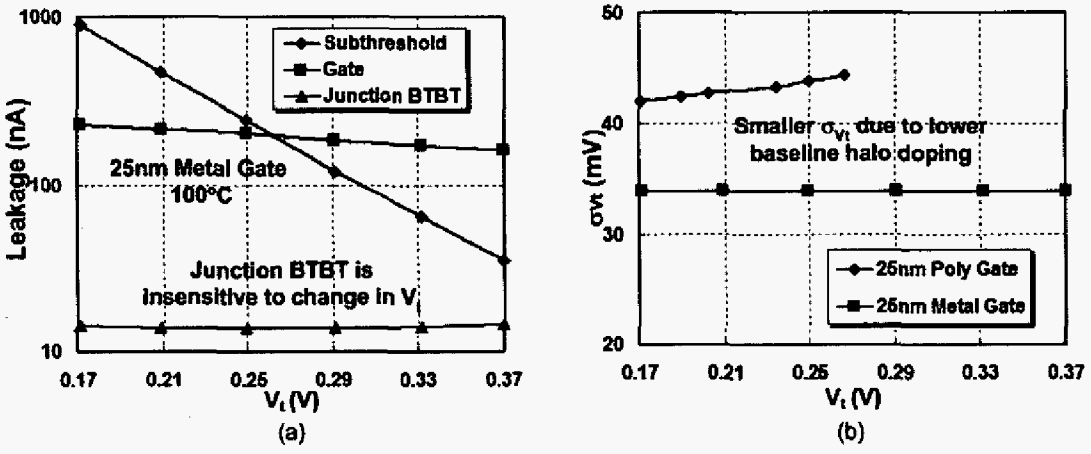

Figure 9. Simulation results of 25nm low/high-V_t optimum metal gate devices a) Leakage components b) σ_{Vt} due to random dopant fluctuation vs V_t.

subthreshold leakage exponentially. It also decreases the gate leakage due to reduction in both the oxide field and the inversion charge available for tunneling (increasing V_t) [11]. The junction BTBT leakage is almost insensitive to the change in V_t. Since metal gate devices require lower baseline halo concentration to maintain SCE, it has lesser junction BTBT and smaller σ_{Vt} (due to random dopant fluctuation, Figure 9b) compared to poly-Si gate devices. Moreover, they are insensitive to change in V_t. A dual-V_t optimization using OPT3 results in 55% reduction in leakage (optimum high-V_t = 0.29V, 120mV higher than low V_t), while ensuring yield for the 8-bit ALU designed using metal gate devices.

We can conclude from above discussions, that metal gate work function engineering to realize high-V_t devices is suitable for dual-V_t 25nm technology, while achieving high performance and target yield. The most desired metal gates should possess work functions close to Si band edges for CMOSFETS. More importantly, these metal gates should be thermally stable to employ a convenient process flow for fabrication. However, it is extremely challenging to identify two thermally stable metal gates with the correct work functions. Furthermore, the method of preparing the metal gates is critical due to process induced damages [12] and Fermi level pinning. Many researchers have proposed different metal gates and fabrication process to achieve these tasks [10-12] and significant research is still under way.

3. CONCLUSIONS

In this paper, we show that in nano-scale regime, conventional dual-V_t design suffers from yield loss due to process variation and vastly overestimates leakage savings since it does not consider junction BTBT leakage into account. Our analysis shows the importance of considering device based analysis while designing low power schemes like dual-V_t. It also shows, that in scaled technology, statistical information of both leakage and delay helps in minimizing total leakage while ensuring yield with respect to target delay in dual-V_t designs. However, non-scalability of the present way of realizing high-V_t requires the use of different process options such as metal gate work function engineering in future technologies.

Acknowledgement: This research was sponsored in part by Semiconductor Research Corporation under contract 1078.001 and by Intel Fellowship.

4. REFERENCES

[1] Y. Taur, and T.H. Ning, Fundamentals of Modern VLSI Devices, Cambridge University Press, New York, 1998.

[2] A. Agarwal, K. Roy, S. Hsu, R. K. Krishnamurthy, and S. Borkar, "A 90 nm 6.5GHz 128x64b 4-read 4-write ported Parameter Variation Tolerant Register File," In IEEE Symposium on VLSI Circuits, June 2004, pp. 386-387.

[3] P. Pant et al., "Dual-threshold voltage assignment with transistor sizing for low power CMOS," In IEEE Transactions on VLSI Systems, April 2001, pp. 390-394.

[4] M. Ketkar et al, "Standby power optimization via transistor sizing and dual threshold voltage assignment," in Int. Conf. on Computer Aided Design, Nov 2002, pp. 375-378.

[5] "Well-Tempered" Bulk-Si NMOSFET Device, Microsystems Technology Laboratory, MIT, Available: http://www-mtl.mit.edu/Well/

[6] K. A. Bowman, et al., "Impact of die-to-die and within die parameter fluctions on the maximum clock frequency distribution for gigascale integration," IEEE Journal of Solid State Circuits, Feb 2002.

[7] K. Kang, B. C. Paul, and K. Roy, "Statistical Timing Analysis using Levelized Covariance Propagation" in Proc. of DATE, 2005.

[8] X. Tang, V. De, and J.D. Meindl, "Intrinsic MOSFET parameter fluctuations due to random dopant placement," IEEE Transaction on VLSI System, Dec. 1997, pp. 369-376.

[9] R. Rao et al., "Statistical estimation of leakage current considering inter- and intra-die process variation," In Int. Symp. on Low Power Elect. and Design ISLPED 2003.

[10] D.-G. Park et al., "Thermally robust dual-work function ALD-MN MOSFET using conventional CMOS process flow," In IEEE VLSI Technology Symposium, June 2004.

[11] Y.-T. Hou et al., "Metal gate work function engineering on gate leakage of MOSFET," In IEEE Transaction on Electron Devices, Nov 2004.

[12] H. Y. Yu et al., "Fermi pinning-induced thermal instability of metal gate workfunctions," In IEEE Electron Device Letters, May 2004, pp. 123-125.