Device-Aware Yield-Centric Dual-V_t and Transistor Sizing 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

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

***Abstract*— Dual-V_t design technique has proven to be extremely effective in reducing sub-threshold leakage in both active and standby mode of operation of a circuit in submicron technologies. However, aggressive scaling of technology results in different leakage components (subthreshold, gate and junction tunneling) to become significant portion of total power dissipation in CMOS circuits. High-V_t devices are expected to have high junction tunneling current (due to stronger halo doping) compared to low-V_t devices, which in the worst case can increase the total leakage in dual-V_t design. Moreover, process parameter variations (and in turn V_t variations) are expected to be significantly high in sub-50 nm technology regime, which can severely affect the yield. In this paper, we propose a device aware simultaneous sizing and dual-V_t design methodology that considers each component of leakage and the impact of process variation (on both delay and leakage power) to minimize the total leakage while ensuring a target yield. Our results show, conventional dual-V_t design can overestimate leakage savings by 36% while incurring 17% average yield loss in 50nm predictive technology. The proposed scheme results in 10-20% extra leakage power savings compared to conventional dual-V_t design, while ensuring target yield. This paper also shows that non-scalability of the present way of realizing high-Vt devices results in negligible power savings beyond 25nm technology. Hence, different dual- V_t process options, such as metal gate work function engineering, are required to realize high-performance and low-leakage dual-V_t designs in future technologies.**

I. INTRODUCTION

CMOS devices are being scaled down aggressively in each technology generation to achieve higher integration density, while the supply voltage is scaled to achieve lower switching energy per device. However, to achieve high performance, there is a need for commensurate scaling of the transistor threshold voltage (V_t), which in turn increases the subthreshold leakage exponentially [1]. This aggressive scaling of the devices not only increases subthreshold leakage but also has other negative impacts such as increased drain induced barrier lowering (DIBL), V_t roll-off, reduced on-current to off current ratio, and increased source-drain resistance [2]. To avoid the short channel effects, 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. However, low oxide thickness gives rise to high electric field, resulting in considerable direct tunneling current (gate leakage, Fig. 1). Higher doping results in high electric field across the p-n junctions (source-substrate or drain-substrate), which causes significant junction (source/substrate & drain/substrate) band to band tunneling (BTBT leakage) of electrons from the valence band of the p-region to the conduction band of the n-region (Fig. 1). There is another leakage component called gate induced drain leakage (GIDL)) which is also a product of small transistor geometries and may not be a dominant component during regular operations of the circuit. During normal mode of operation, the major leakage currents are gate, junction BTBT and subthreshold leakage. The increase in different leakage components with technology scaling has two major implications in logic design. First, leakage reduction techniques are becoming indispensable in future design. Moreover, different

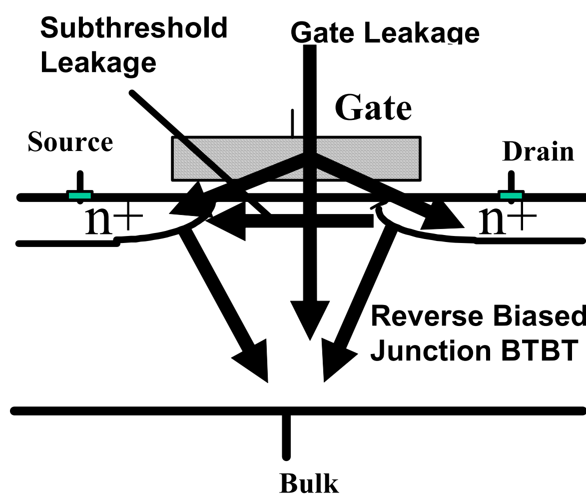


Fig. 1. Major leakage components in a transistor.

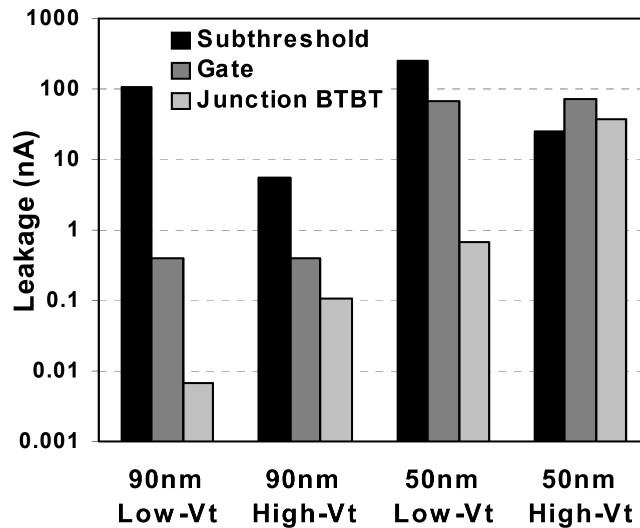


Fig. 2. Leakage components in 90nm and 50nm low and high-V_t.

leakage mechanisms are becoming equally important with device scaling. Hence, the relative magnitudes of each of the leakage components play a major role in low-leakage logic design.

Furthermore, controlling the variation in device parameters during fabrication is becoming a great challenge for scaled technologies. The delay and leakage currents 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 leakage components and significant spread in delay. Among the statistical variations, the random placement of dopants is of great concern [3] because it is independent of transistor spatial location and causes threshold voltage mismatch between transistors even though they may be close to each other (intra-die variation) resulting in significant leakage and delay variation of logic gates and circuits. Hence, any low leakage design needs to consider the spread of leakage and delay, both at 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 mode of operation of a circuit in submicron technologies. However, with the emerging issues related to technology scaling, the effectiveness of conventional dual-V_t design technique [4-6] may be degrading in nano-scale technologies. The issues related to dual-V_t design in nano-scale technologies are:

1) Scaled devices require the use of higher substrate doping and the application of the "halo" profiles to reduce the short channel effect. The high halo doping supercedes any change in base channel doping or threshold voltage

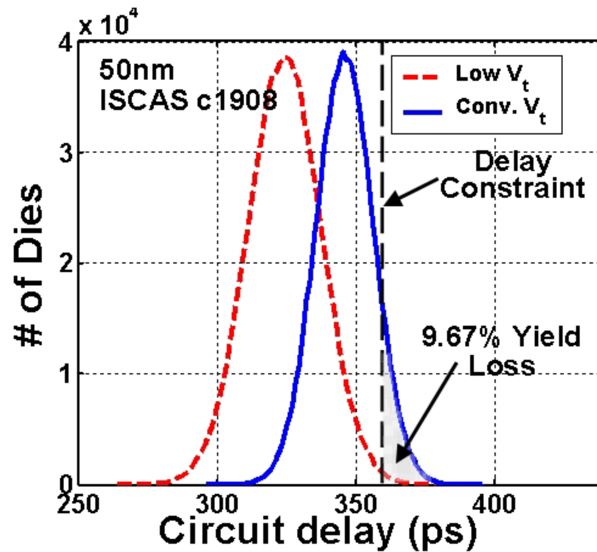


Fig. 3. Yield loss due to dual-V_t design in 50nm technology.

implants, which were used traditionally to achieve high-V_t devices. In nano-scale technologies, high-V_t devices can be obtained by increasing the peak halo doping. This higher halo doping reduces the subthreshold leakage exponentially, however, it results in significant junction BTBT current (note gate leakage is insensitive to halo doping profile, Fig. 2). 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 the relative magnitudes of different leakage components vary across different V_t devices, the selection of high-V_t device in a dual- V_t design should consider this tradeoff. A device aware dual-V_t design, which investigates different device design options for realizing the optimum low/high-V_t devices, is required so that the leakage savings can be amplified.

2) 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 [8]. 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 resulting in reduced yield. Fig. 3 plots the circuit delay distributions of a low-V_t and a conventionally optimized (for low leakage) dual-V_t circuit. We can observe that after V_t assignment, more number of dies may fail to meet the required delay constraint resulting in low yield. Moreover, different V_t devices will have different process variation spread. A

high-V_t device is expected to have large σ variation due to high halo doping concentration [9] (more random dopant fluctuation). Hence, a device aware dual-V_t design, which considers the delay distribution of circuit under process variation, is required to minimize leakage, while ensuring yield.

3) 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 different V_t devices (high-V_t devices will have large σ).

All previously proposed dual-V_t design techniques either ignore the effect of process variation [4-7] or do not consider all leakage components while selecting high-V_t devices. Since both process variation and relative magnitude of different leakage components strongly depend on the choice of low/high V_t devices, *we propose a device aware yield-centric dual-V_t design methodology, which will consider each component of leakage and the impact of process variation (on both delay and leakage power) to minimize the total leakage while ensuring a target yield.* We also analyze the effectiveness of dual-Vt design with technology scaling. Our results show that non-scalability of present way of realizing high-Vt devices results in negligible power savings beyond 25nm technology even in our proposed device aware dual-V_t design. Hence, 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 bulk technologies.

The rest of the paper is organized as follows. In section 2, we show our device level analysis for 90nm, 50nm, and 25nm dual-Vt technologies. Section 3 explains the statistical leakage and delay analysis and presents our proposed device-aware yield-centric dual-Vt design under parameter variations. Section 4 presents the experimental results on a set of ISCAS85 benchmark circuits. Section 5 describes the metal gate work function engineering, and shows its effectiveness in saving leakage power, when used as an option for designing dual-Vt devices. In section 5, we draw our conclusions.

II. DEVICE LEVEL ANALYSIS

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

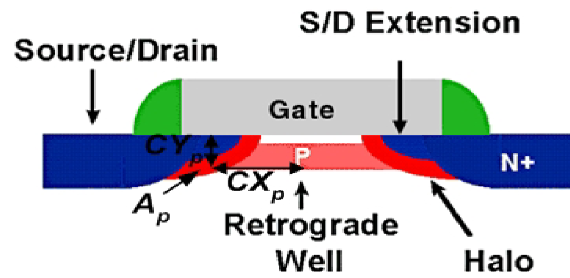


Fig. 4. Nano-scaled n-channel device with halo doping

of the vertical peak of the halo (Cy_p) away from the bottom junction and towards the surface (Fig. 4). 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. It also increases the V_t variability (σ) due to random dopant fluctuation and the junction capacitance. To investigate 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 [10] and the design guideline given in 2001 & 2003 ITRS Roadmap for effective gate length of 90nm, 50nm and 25nm. The devices were simulated using MEDICI device simulator [11].

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 [10]. Device optimization was performed by varying halo doping profile while keeping the subthreshold leakage fixed to 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

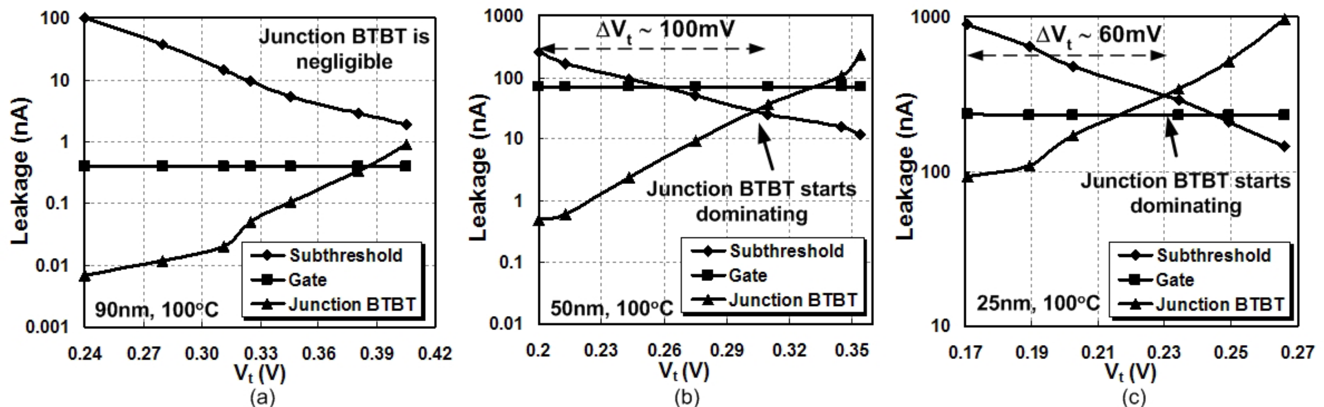


Fig. 5. Simulation results of low/high-Vt optimum n-channel devices leakage components a) 90nm, V_{DD} = 1.5V b) 50nm, V_{DD} = 1.2V c) 25nm, V_{DD} = 1.0V.

subthreshold leakage devices correspond to different-V_t devices. 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.

Fig. 5 plots the different leakage components in our optimized low/high V_t NMOS devices 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-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, the junction BTBT is almost negligible as compared to the subthreshold leakage for a wide range of V_t (Fig. 5a). 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 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 (Fig. 5b). This difference between low and high-V_t gets smaller (only 60mV) as we go to 25nm technology (Fig. 5c). Hence, 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, therefore, overestimate the leakage savings and in the worst case might increase the total leakage. The gate leakage is almost insensitive to the change in V_t.

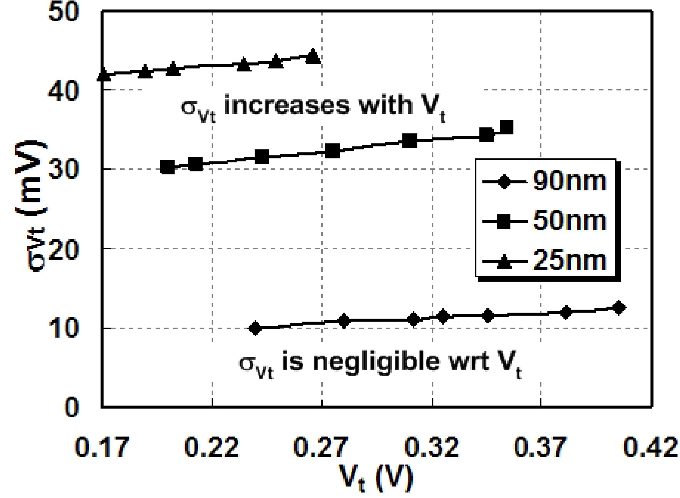


Fig. 6. σ_{Vt} due to random dopant fluctuation vs V_t.

Fig. 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 [9]:

$$\sigma_{Vt} = \frac{qT_{ox}}{\varepsilon_{ox}}\sqrt{\frac{N_a W_d}{3LW}} \tag{1}$$

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. Fig. 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. Since σ_{Vt} is inversely proportional to the square root of width of transistor, higher width devices have smaller V_t variation with respect to random dopant fluctuation.

III. STATISTICAL CIRCUIT ANALYSIS

In nano-scaled CMOS devices, the random variations in the number and the placement of dopant atoms in the channel region cause random variations in the transistor threshold voltage (V_t). Moreover, the delay and leakage distribution of a circuit strongly depends on the device geometry (channel length, width, oxide thickness etc.) and doping profile. Although Monte-Carlo simulation of gates is accurate (e.g. using circuit simulator like SPICE during circuit design and device simulator like MEDICI during device design) in estimating the delay and leakage

distributions, it considerably increases the design time. This is also computationally expensive particularly if estimation is required at the device design phase. Hence, statistical modeling and analysis of delay and leakage of logic gates are necessary both at the circuit and device design phase for low power dual-V_t design. This section describes our semi-analytical method to estimate both leakage and delay distribution in a circuit using the feedback from device simulations to improve the accuracy and efficiency of dual-V_t design.

In this work, for intra-die variation, we 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 [3]. For inter-die variation we consider variation in gate length (L_{gate}), usually considered to be the dominant source of inter-die variation. It should be noted that any other variations can easily be incorporated into our model. We assume that random dopant fluctuation is independent of L_{gate} variation. While there is a minor dependency between L_{gate} and random dopant fluctuation (1), the error introduced as a result of the independence assumption was found to be negligible. The standard deviation of V_t due to random dopant fluctuation is extracted from our optimized device using (1) (Fig. 6), which depends on both V_t and width of the transistors. We assume 15% 3-sigma variation in L_{gate} for our analysis.

A. Statistical Leakage Power Estimation

First, the different components of leakage (subthreshold, gate and junction BTBT leakage) of a device is modeled using the device geometry, 2-D doping profile and the operating temperature based on analytical models described in [12]. These analytical models are calibrated and verified by device simulation across different biasing condition and device/circuit parameters. The leakage models are used to estimate total subthreshold, gate and junction BTBT leakage of a circuit in our dual-V_t design. Second, the sensitivity of different parameters (σ_{Vt}, gate length variation) on leakage is extracted from device level analysis for different V_t devices. Finally, the developed model and extracted sensitivity are used to estimate the total leakage and its distribution in a circuit. The subthreshold leakage dependency on L_{gate} and V_t variation is modeled as an exponential decay model (e.g. $I_0e^{-K\Delta L_{gate}}$). Since the junction BTBT and the gate leakage are almost insensitive to random dopant fluctuation and only linearly dependent on L_{gate} variation, we neglect any variation for these two leakage components.

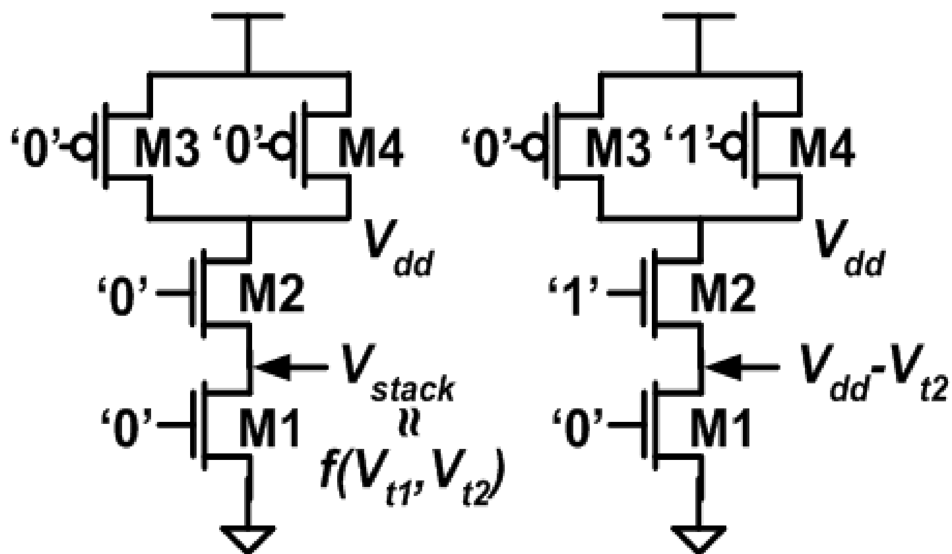


Fig. 7. NAND gate leakage dependency with respect to input vector, stacking and body effect.

Since, the V_t variation is random in nature, it is assumed to be independent for all transistors in a circuit. Hence, the subthreshold leakage in a logic gate depends on the V_t variation in different transistors in that logic gate and the input vector. As we can observe from Fig. 7, the subthreshold leakage of a 2-input NAND gate, for input vectors "00" and "01", depends on threshold voltage of both M1 and M2. This can be modeled as:

$$I_{nand00} = I_{00} e^{-K_1 \Delta Vt_1 - K_2 \Delta Vt_2} \qquad I_{nand01} = I_{01} e^{-K_1' \Delta Vt_1 - K_2' \Delta Vt_2}$$

The subthreshold leakage for input vector "10" and "11" can easily be written as:

$$I_{nand10} = I_{10} e^{-K_2'' \Delta Vt_2} \qquad I_{nand11} = I_{11} e^{-K_3 \Delta Vt_3} + I_{11} e^{-K_4 \Delta Vt_4}$$

Hence the total leakage in a NAND gate is:

$$I_{nand} = p_1 I_{nand00} + p_2 I_{nand01} + p_3 I_{nand10} + p_4 I_{nand11}$$

where p_1, p_2, p_3 and p_4 are the signal probability (input probability of 00, 01, 10 and 11, respectively) of input vectors at the inputs of the NAND gate. These probabilities are calculated using signal probability *P(k)* of a node *k* in a circuit, which is the probability that node *k* is logical one. These signal probabilities can be propagated through basic gates based on simple rules of probability and logic function of the gate [13]. Hence, the PDF of total leakage in a NAND gate is sum of correlated lognormals, which can be well approximated by another lognormal using Wilkinson's method [14]. Similarly, the PDF of leakage distribution in different types of gates (NOR, Inverter) can

be approximated as a lognormal based on their input probabilities.

The PDF of total leakage of a circuit due to intra-die V_t variation can be written as a sum of independent lognormal distribution associated with different logic gates. This can be again approximated as lognormal using Wilkinson approximation.

$$I_{intra} = I_1 + I_2 + ... = e^{Y_1} + e^{Y_2} + ... = e^{i_{intra}} \quad (2)$$

Finally, the total leakage distribution of a circuit considering both inter (L_{gate} variation) and intra die variation can be written as the product of two lognormal distributions, which itself can be represented as a lognormal:

$$I_{total} = e^{-K_L \Delta L_{gate}} e^{i_{intra}} = \left(\frac{1}{2x\sqrt{2\pi\sigma}} \right) e^{\left(\frac{-(\ln(x)-\mu)^2}{2\sigma^2} \right)} \quad (3)$$

where μ and σ are the parameters of final lognormal distribution. All sensitivities (e.g K_L, K_1, K_2 etc.) and σ_{Vt} associated with all the transistors are calculated using the developed leakage models and device simulations. The percentage point (θ) function of a lognormal is defined as:

$$P\{I_{total} \le i_\theta\} = \theta; \quad \text{where } i_\theta = e^{(\mu + \sigma\Phi^{-1}(\theta))} \quad (4)$$

Here, Φ is the CDF of a standard normal distribution. This can be used to estimate the confidence point of leakage distribution.

B. Statistical Delay Model

For optimizing dual-Vt designs, we have employed the statistical static timing analysis (SSTA) algorithm proposed in [15], where delay distribution of a circuit is calculated using the Levelized Covariance Propagation (LCP) technique. In contrast to the conventional block-based SSTA algorithm where signals are propagated through each logical gate in a breadth-first search (BFS) order, this algorithm groups the number of signals in a single logic level to a statistical data structure. This statistical data structure, namely, Levelized Covariance (LC), is propagated through the target circuit in a logical order. As shown in Fig. 8, LC_i of a single logic level serves as an input condition to the next logic to compute the LC_{i+1} of the next logic level. This LC structure includes the mean and standard-deviation of signal arrival time and the covariance (correlation due to spatial correlation of process

parameters and reconvergent paths) among the output signals in a single logic level. To compute the mean and standard deviation of gate delay, we have used a first order Taylor's expansion as:

$$D = D_0 + \sum_{i}^{n} s_i X_{Vt,i} + s_L X_L \quad (5)$$

where D_0 represents nominal value of the gate delay without any variation and s_i's are the sensitivities of the gate delay to the variation in threshold voltages of each transistor in a particular logic gate. s_L is the sensitivity of the gate delay to the variation in channel length of transistors in a particular logic gate. $X_{vt,i}$ and X_L are Gaussian random variables representing variation in transistor threshold voltage and channel length, respectively. $X_{vt,i}$ is chosen to be a completely independent random variable with respect to all other random variables. On the other hand, X_L shows a systematic variation due to the spatial correlation of process parameters. In [15], rectangular grid-based spatial correlation model [16] is applied to compute the correlation among different X_L's. Nominal delay D_0 is modeled using analytical expression presented by Sakurai et al. [17], which represents the gate delay in terms of several device parameters extracted from the I-V characteristics. In our case, the I-V characteristics are obtained from the set of both high-V_t and low-V_t devices generated using MEDICI device simulator. Sensitivity values (s_i, s_L) used in Eq. (5) are also obtained from the MEDICI device simulation. It was shown in [15] that, using this technique, the effect of both inter- and intra-die variation can be taken into account. The simulation results on several ISCAS85 benchmarks show

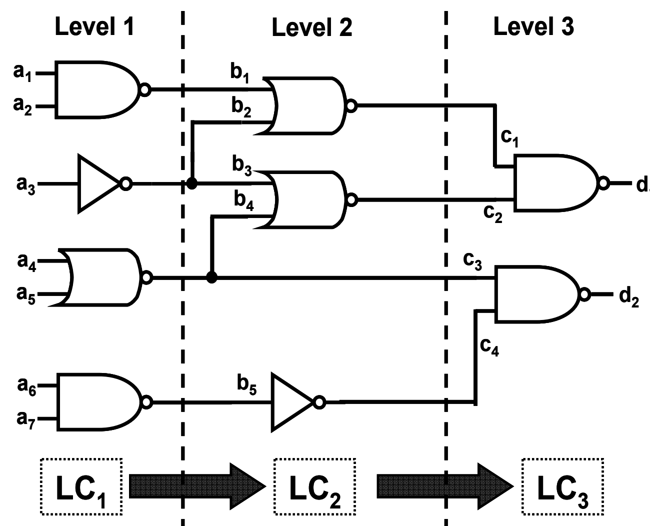


Fig. 8. Levelized Covariance Propagation (LCP) technique.

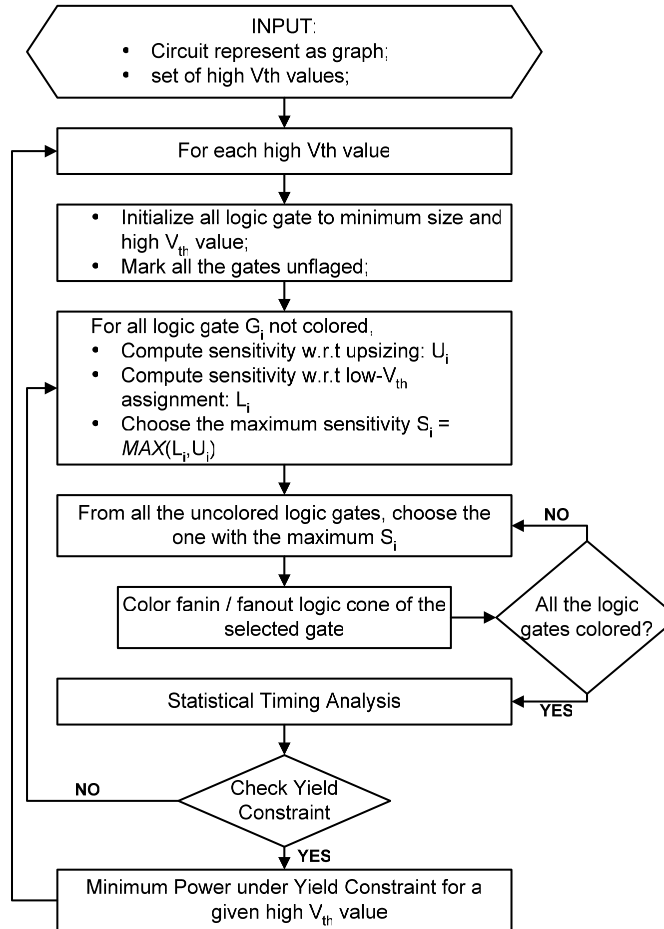


Fig. 9. Flow chart for the simultaneous gate sizing and dual-V_t design algorithm.

average error of 0.21% and 1.07% compared to the Monte-Carlo analysis for mean and standard deviation of delay, respectively.

C. Simultaneous Gate Sizing and Dual-Vt Assignment Algorithm

In this section, we describe the simultaneous gate sizing and dual-V_t assignment algorithm to minimize total power (dynamic power and leakage power) of the circuit while meeting a target yield (with respect to a given delay constraint). The proposed algorithm is iterative, where the choice of transistor size or V_t-assignment in each iteration is based on sensitivity analysis of the logic gates. The basic steps of the algorithm are described in Fig. 9.

The algorithm starts by assigning high-V_t and minimum size to all logic gates in the given circuit, which corresponds to the minimum power / maximum delay configuration of the circuit. Then the algorithm proceeds to selectively choose logic gates for low-V_t assignment or up-sizing in multiple iterations. We use a flag (i.e., color) to indicate if a particular logic gate is either assigned high-V_t value or up-sized in a particular iteration. In each iteration, we

compute sensitivity of each logic gate with respect to the change in circuit delay for unit change in circuit power as follows:

$$S_i = MAX\left(\left.\frac{\Delta D_{ckt}}{\Delta P_{ckt}}\right|_{upsizing}, \left.\frac{\Delta D_{ckt}}{\Delta P_{ckt}}\right|_{low-Vt} \right) \tag{6}$$

where ΔD_{ckt} and ΔP_{ckt} represent the change in circuit delay or power by upsizing or low-V_t assignment, respectively. ΔD_{ckt} can be calculated using the statistical timing analysis method proposed in the previous section. ΔP_{ckt} represents the change in total power of the circuit for a sizing or low-V_t assignment operation. Note that ΔP_{ckt} includes the dynamic power and all leakage components (subthreshold, junction-tunneling and gate leakage). To consider the impact of process variation on leakage, we have characterized the leakage values of the cells for the 95 percentile value from the leakage distribution. The leakage and delay distribution are generated using the algorithm presented in section 3.1 and 3.2, respectively. The necessary parameters used in delay and power models are extracted from device simulation results for high / low-V_t devices.

In each iteration, we rank all the logic gates in descending order of their sensitivity (S_i) values. Then we choose the logic gates in order of their sensitivities and assign them with the best size-factor / V_t-value (with highest sensitivity S_i). When the selection of size-factor / V_t-value is done, all the logic gates in its fan-in and fan-out cone are colored, so that they are not considered for sizing / V_t-allocation in the current iteration. This helps us to improve the runtime of the algorithm, by minimizing expensive statistical timing analysis runs. When all gates are colored, we run the statistical timing analysis and check for the yield constraint. If the constraint is satisfied, the algorithm terminates for a fixed V_t value, else it goes back to the initialization step and proceeds to the next iteration. The algorithm for Vt assignment/sizing for a fixed V_t value terminates before the yield constraint is satisfied after an iteration failing to improve yield by a threshold margin.

The proposed algorithm works on a greedy heuristic of sensitivity based sizing / V_t-value selection. The V_t assignment/sizing algorithm is effective in terms of reducing total power while satisfying yield constraint since, at each iteration, it selects only the most sensitive logic gates that can meet the yield bound with minimum power increase. The effectiveness of the algorithm largely depends on the accuracy and efficiency of calculating sensitivity values at each iteration. In our experiments, we have used a simplified version of timing and power analysis to re-

compute sensitivity of logic gates with respect to both upsizing and low-V_t assignment at the beginning of each iteration. The complexity of the algorithm (*O(A)*) is:

$$O(A) = r \times (N \times (O(S_t) + O(S_p)) + O(T)) \quad (7)$$

where *r* is the number of iterations required; *N* is the number of logic gates; $O(S_t)$ and $O(S_p)$ are the complexity timing analysis and power (dynamic and leakage) analysis, respectively. *O(T)* is the complexity of statistical timing analysis and yield computation. Note that we run this algorithm over a set of pre-selected high V_t values. Hence, the overall complexity of the algorithm in Fig. 9 will be: $K \times O(A)$, where K is the number of pre-selected high V_t values.

IV. RESULTS

In this section, we compare the leakage savings and yield improvement achieved by traditional static dual-V_t design (CONV), which only considers subthreshold leakage as the optimization criteria, with our proposed device aware dual-V_t scheme considering process variation and all components of leakage. To show different tradeoffs associated with leakage savings and yield improvement, we categorized our algorithm into three progressive optimizations. First, optimization (OPT1) considers all components of leakage but ignores any variation in delay and leakage. The second optimization (OPT2) takes circuit delay variation into account to ensure yield, while considering all components of leakage, but ignores any leakage variation. The third optimization (OPT3) considers all the above parameters and also minimizes 95 percentile leakage of the circuit. We show all our results for ISCAS benchmarks using our optimized devices for 90nm, 50nm and 25nm technologies. We first size all ISCAS benchmark circuits using low-V_t transistors for a given delay constraint for minimum dynamic power. We estimate the total dynamic and leakage power of these optimally sized circuits considering all components of leakage. We also measure the 95th percentile delay of these circuits. We use these power and delay values as the basis for showing our results for power savings and yield estimation.

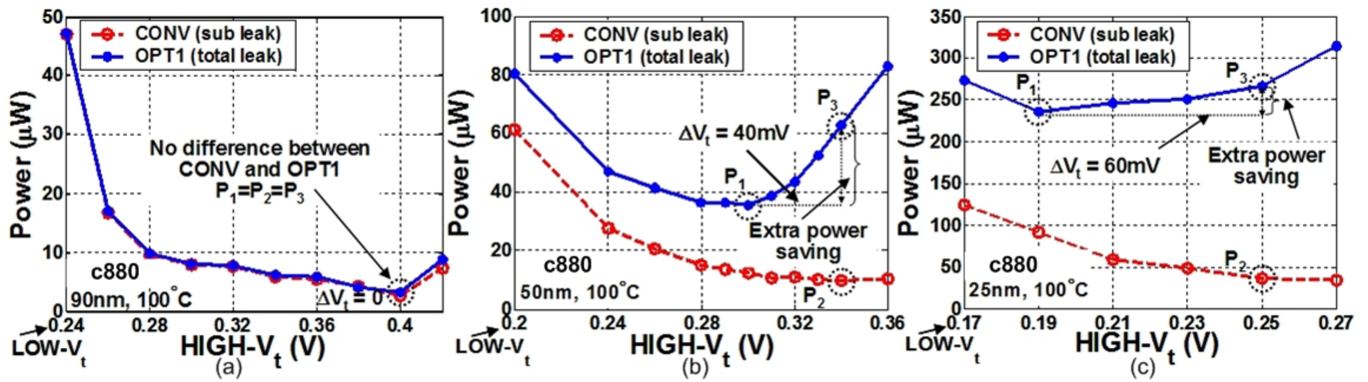


Fig. 10. High-Vt assignment using CONV and OPT1 a) 90nm b) 50nm c) 25nm technology. P1: Leakage power using OPT1, P2: Expected leakage Power using CONV, P3: Actual leakage power using CONV.

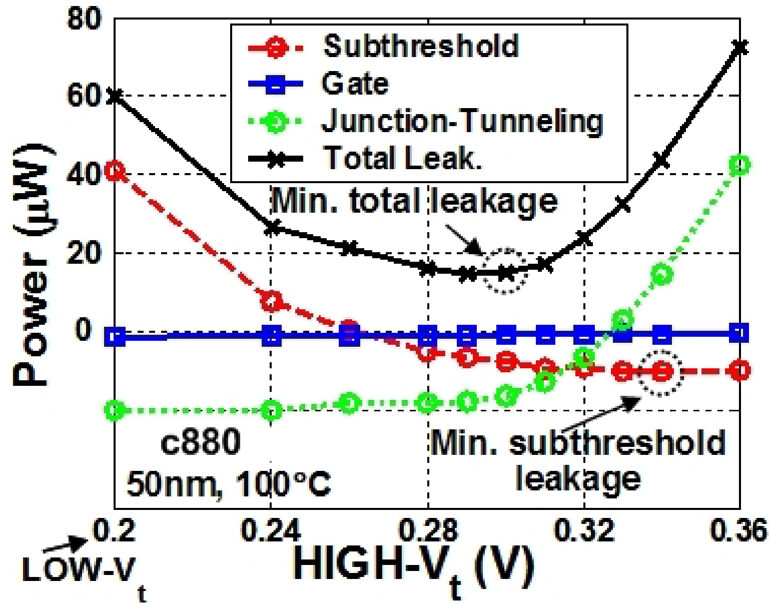
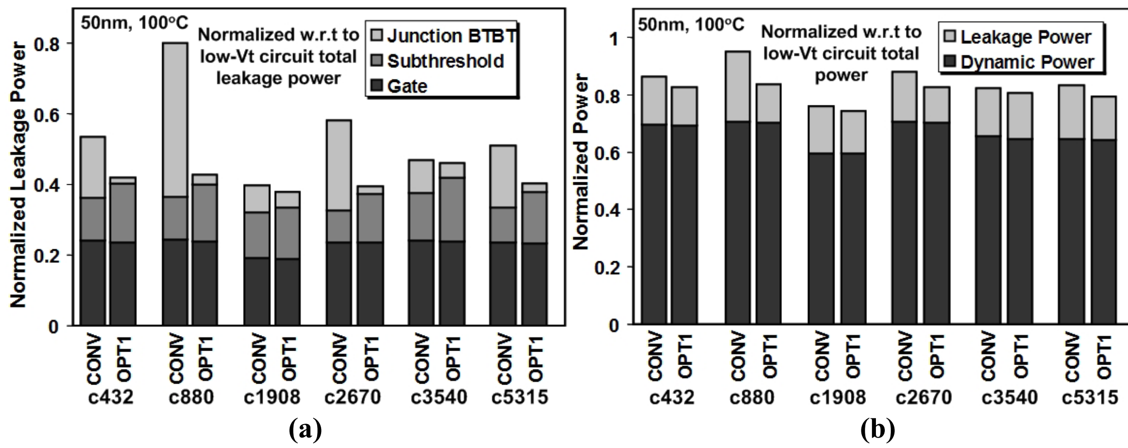


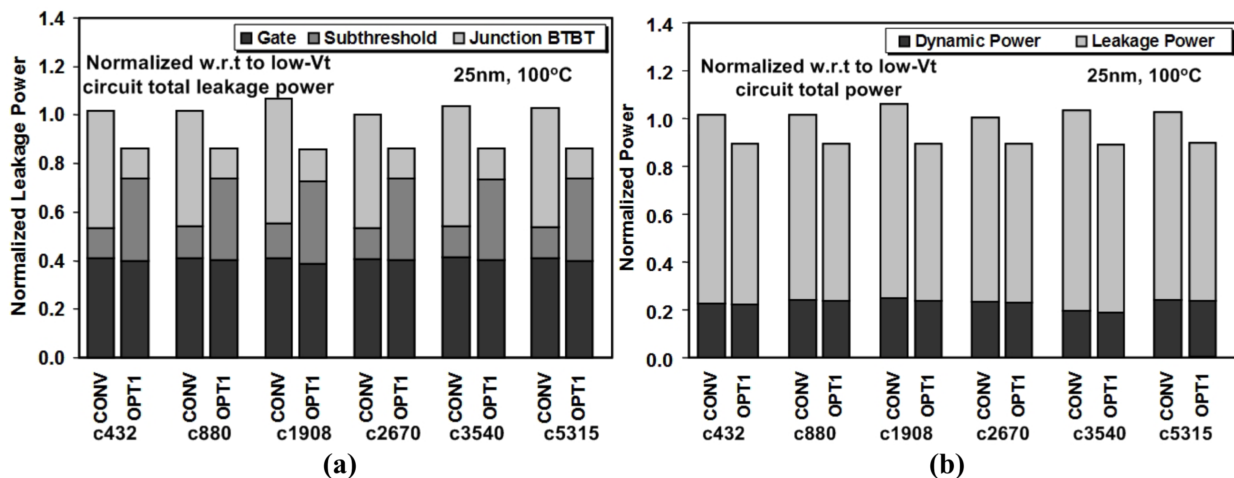
Fig. 11. 50nm technology leakage power components in CONV.

Fig. 10 compares the optimum high-V_t and the leakage savings achieved by conventional dual-V_t design (CONV) and OPT1 for ISCAS c880. For 90nm technology both design techniques select the same optimum high-V_t and results in around 90% leakage savings (Fig. 10a). However, for 50nm technology, optimum high-V_t's selected by CONV and OPT1 differ by 40mV (Fig. 10b). Fig. 11 analyzes the different power components in CONV for the 50nm node. It shows that even though subthreshold leakage power is minimum at V_t = 0.34V, the total leakage minima occurs at a smaller V_t value 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 (Fig. 10b), the total power (P3) actually exceeds the minimum power (P1) achieved by OPT1 by 37%. Hence, OPT1 achieves more leakage power savings compared to conventional approaches. It also shows that CONV overestimates the total power saving by 63% and only saves 17%



(a) **(b)**

Fig. 12. (a) Minimum total leakage using CONV and OPT1 (b) Minimum total power using CONV and OPT1 in 50nm technology.



(a) **(b)**

Fig. 13. a) Minimum total leakage using CONV and OPT1 b) Minimum total power using CONV and OPT1 in 25nm technology.

of total leakage. Even though OPT1 results in higher junction BTBT leakage compared to low-V_t design, it saves more than 54% of total leakage.

Fig. 12 plots the minimum total leakage and minimum total power achieved by both design methods for different benchmarks in 50nm technology. Total leakage and total power shown are normalized with respect to total leakage and total power of an optimally sized low-V_t design, respectively. Table I shows the expected (P2, considering only subthreshold leakage) and actual leakage power savings (P3, considering all components of leakage) in CONV and total leakage power savings (P1) achieved by OPT1 across different benchmarks, while considering the dynamic power overhead due to sizing. Our device aware scheme results in average 14% and a maximum of 37% more leakage power savings compared to conventional scheme. Conventional designs overestimate the leakage savings by 36% (average). Considering all components of leakage power results in around 20mV smaller optimum V_t compared to considering subthreshold leakage in optimization.

TABLE I
% LEAKAGE SAVINGS USING CONV AND OPT1 IN 50NM TECHNOLOGY

		c432	c880	c1908	c2670	c3540	c5315
CONV	Expected (P2)	79.6	80.2	82.0	86.8	77.9	84.5
	Actual (P3)	43.1	17.0	58.7	40.6	49.7	46.8
OPT1	Actual (P1)	54.8	54.1	63.1	57.8	54.1	57.2

TABLE II
% LEAKAGE SAVINGS USING CONV AND OPT1 IN 25NM TECHNOLOGY

		c432	c880	c1908	c2670	c3540	c5315
CONV	Expected (P2)	71.9	70.4	66.5	72.1	70.8	70.8
	Actual (P3)	-2.1	0.2	-7.9	-0.3	-4.1	-3.4
OPT1	Actual (P1)	13.7	13.9	14.1	13.7	13.7	13.8

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 (Fig. 10c). 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. Fig. 13 plots the minimum total leakage and minimum total power achieved by both CONV and OPT1 for different benchmarks in 25nm technology. Table II shows the respective leakage savings as discussed above for 25nm technology. It can be observed from the figure that CONV results in negligible power savings and for some benchmarks it actually increases the total power compared to low-Vt design. Our device aware scheme OPT1 results in average 13.8% and a maximum of 14.1% leakage power savings. The difference between low-Vt and optimum high-Vt for OPT1 varies from 20-30mV across benchmarks. It is evident from the above results that dual-V_t designs should consider each component of leakage while optimizing circuit to reduce total leakage power. Since, increasing peak halo doping to realize high-V_t increases junction BTBT leakage results in negligible leakage savings in aggressively scaled technologies, 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.

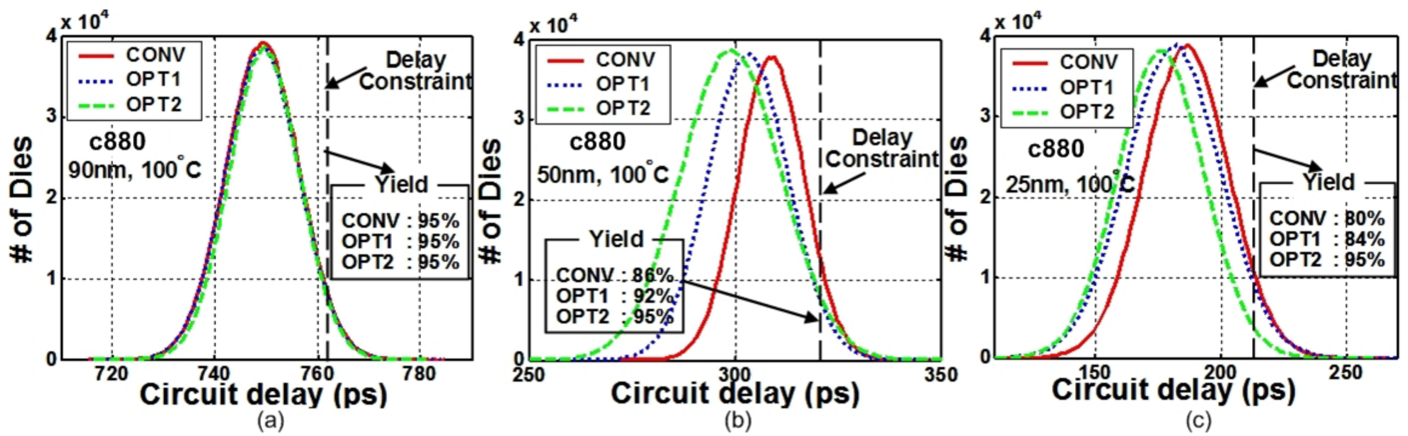


Fig. 14. Circuit delay distribution and yield loss using CONV, OPT1 and OPT2 a) 90nm b) 50nm c) 25nm technology.

Table III

% Yield using CONV and OPT1 in 50nm Technology

	c432	c880	c1908	c2670	c3540	c5315
CONV	90	88	79	41	94	81
OPT1	85	94	79	48	93	79

Fig. 14 plots the circuit delay distributions obtained using our statistical timing analysis tool (section 3) for the optimum dual-Vt circuits (high-Vt which achieves minimum power) using CONV, OPT1 and OPT2 in ISCAS c880. 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_t, it is able to meet the required 95th percentile delay yield for both 50nm and 25nm technology. Table III shows the yield of different benchmarks using CONV and OPT1 for 50nm technology. The CONV and OPT1 result in average 17% and 16% and maximum of 59% and 52% yield loss in 50nm technology, respectively. The yield loss for 25nm is observed to be higher and OPT1 results in better yield on an average than CONV. Since c2670 has large number of primary inputs (233) and a large number of critical paths, which in turn results in large mean shift [8], it has maximum yield loss.

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 55% (average)

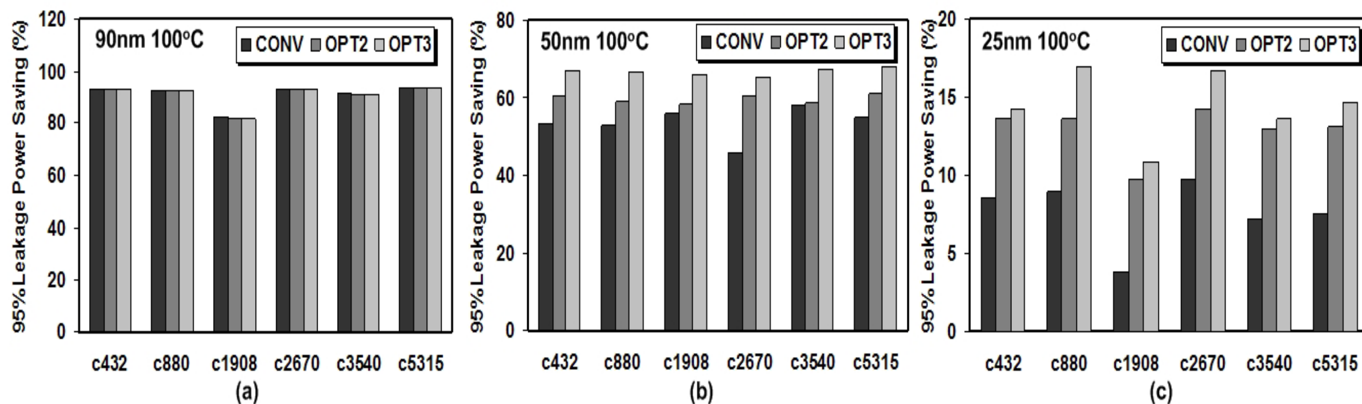


Fig. 15. 95 percentile leakage power saving using CONV, OPT2 and OPT3 a) 90nm b) 50nm c) 25nm technology.

in 50nm technology. However, it is only 11% (average) in 25nm technology. This shows that in aggressively scaled technologies, 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.

Fig. 15 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 for different benchmarks. OPT3 results in best 95% percentile leakage power, while it ensures yield with respect 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 average 13.4% and 7.3% extra leakage power saving compared to CONV and OPT2, respectively, and saves 67% leakage power compared to all low-Vt design. This shows the importance of considering leakage variation in dual-V_t optimization. In 25nm technology, OPT3 results in average 14% 95th percentile leakage saving compared to low-Vt design, which is 2X higher than the leakage savings achieved by CONV. However, the total leakage power savings compared to low-V_t design itself is negligible.

V. 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

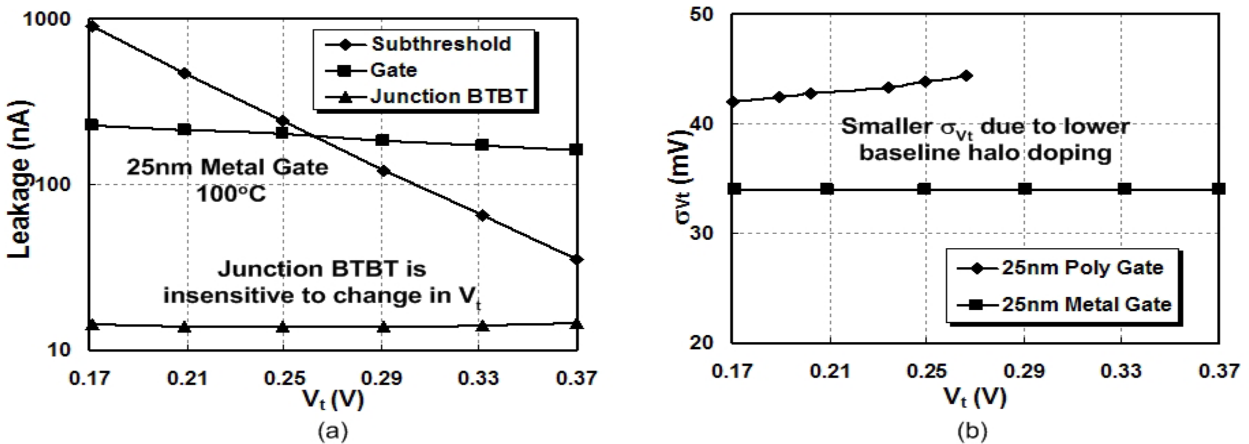


Fig. 16. Simulation results of 25nm low/high-Vt optimum metal gate devices a) Leakage components b) σ_{Vt} due to random dopant fluctuation vs. V_t.

work functions -- without impacting the junction BTBT leakage and process variation [18]. Recently, 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 [18]. The poly depletion increases the effective oxide thickness which in turn reduces the gate capacitance in the inversion region 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. Fig. 16a 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 subthreshold leakage dominates the total leakage in low-V_t devices. Increasing the V_t (by changing the work function) of the device reduces 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) [19]. 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, Fig. 16b) compared to poly-Si gate devices. Moreover, they are insensitive to change in V_t. A dual-V_t optimization using OPT3 results in average 44% reduction in leakage (optimum high-V_t = 0.29V, 120mV higher than low V_t), while

ensuring yield for all the ISCAS benchmarks 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 [20] and Fermi level pinning. Many researchers have proposed different metal gates and fabrication process to achieve these tasks [18-20] and significant research is still under way.

VI. CONCLUSION

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 a low power schemes like dual-V_t. It also shows, that in nano-scale 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. Our proposed device and process variation aware simultaneous sizing and dual-V_t design methodology results in 10-20% extra leakage power savings compared to conventional dual- V_t design, while maintaining yield in 50nm technology. However, non-scalability of the present way of realizing high-V_t devices results in negligible power savings beyond 25nm technology even in our proposed device aware dual-V_t design. We show that the use of different process options such as metal gate work function engineering to realize high-V_t devices will be helpful in achieving high-performance, low-leakage dual-V_t designs in future technologies.

ACKNOWLEDGMENT

The work was sponsored in part by Semiconductor Research Corporation (SRC) and by Gigascale Systems Research Center (GSRC).

REFERENCES

[1] S. Borkar, “Design challenges of technology scaling”, in *IEEE Micro*, 19(4), 23, 1999, pp. 23-29.

[2] Simon M. Sze, “Physics of Semiconductor Devices”, Wiley-Interscience, 1990.

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

[4] P. Pant, R. K. Roy and A. Chattejee, “Dual-threshold voltage assignment with transistor sizing for low power CMOS”, in *IEEE Transactions on VLSI Systems*, April 2001, pp. 390-394.

[5] T. Karnik, Y. Yibin, J. Tschanz, W. Liqiong, S. Burns, V. Govindarajulu, V. De and S. Borkar, “Total power optimization by simultaneous dual-Vt allocation and device sizing in high performance microprocessors”, in *Proceedings of 39th Design Automation Conference,* June 2002, pp. 486-491.

[6] M. Ketkar and S. S. Sapatnekar, “Standby power optimization via transistor sizing and dual threshold voltage assignment”, in *International Conference on Computer Aided Design*, Nov 2002, pp. 375-378.

[7] A. Srivastava, D. Sylvester, D. Blaauw, ”Power minimization using simultaneous gate sizing, dual-Vdd and dual-Vth assignment”, in *Proceedings of 41st Design Automation Conference*, June 2004, pp. 783-787.

[8] K. A. Bowman, S. G. Duvall and J. D. Meindl, “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, pp. 183-190.

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

[10] “Well-Tempered” Bulk-Si NMOSFET Device Home Page, Microsystems Technology Laboratory, MIT, Available: http://www-mtl.mit.edu/Well

[11] MEDICI: Two-dimensional semiconductor device simulation program, *AVANT! Corp.,* Fremont, CA, 2000.

[12] K. Roy, S. Mukhopadhyay and H. Mahmoodi, “Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits”, *Proceedings of the IEEE*, 2003, pp. 305-327.

[13] K. Roy and S. Prasad, “Low Power CMOS VLSI Circuit Design”, Wiley-Interscience, 2000.

[14] S.C. Schwartz et al., “On the distribution function and moments of power sums with lognormal components”, *Bell Systems Technical Journal,* vol. 61, Sept. 1982, pp. 1441-1462.

[15] K. Kang, B. C. Paul, and K. Roy, “*Statistical Timing Analysis using Levelized Covariance Propagation*” in *Proceedings of Design, Automation and Test in Europe*, March 2005. pp. 764 – 769.

[16] A. Agarwal, D. Blaauw, V. Zolotov, S. Sundareswaran, M. Zhao, K. Gala, and R. Panda, “*Path-Based Statistical Timing Analysis Considering Inter- and Intra-Die Correlations*” in *Proceedings of TAU*, 2002.

[17] T. Sakurai and R. Newton, “*Delay analysis of series-connected MOSFET circuits*” in *IEEE Journal of Solid State Circuits*, 2001, pp. 122-131.

[18] 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.

[19] Y.-T. Hou, M. Li, T. Low and K. Dim-Lee, “Metal gate work function engineering on gate leakage of MOSFET,” in *IEEE Transaction on Electron Devices*, Nov 2004.

[20] H. Y. Yu, C. Ren, Y. C. Yeo, J.F Kang, X.P Wang, H.H.H Ma, M. F. Li; D.S.H Chan, D. –L. Kwong, “Fermi pinning-induced thermal instability of metal gate workfunctions” in *IEEE Electron Device Letters*, May 2004, pp. 123-125.