Nanometer Device Scaling in Subthreshold Circuits

Scott Hanson, Mingoo Seok, Dennis Sylvester, David Blaauw
University of Michigan, Ann Arbor, MI
hansons@umich.edu

ABSTRACT

Subthreshold circuit design is a strong candidate for use in future low power applications. It is not clear, however, that device scaling to 45nm and beyond will be beneficial in subthreshold circuits. We investigate the implications of device scaling on subthreshold circuits and find that the slow scaling of gate oxide thickness leads to a 60% reduction in I_{on}/I_{off} between the 90nm and 32nm device generations. We highlight the effects of this device degradation on noise margins, delay, and energy. We subsequently propose an alternative scaling strategy and demonstrate significant improvements in noise margins, delay, and energy in sub-V_{th} circuits.

Categories and Subject Descriptors
B.7 [**Integrated Circuits**]: General
General Terms: Design
Keywords: Subthreshold circuits, device scaling, ultra-low power

1. INTRODUCTION

Subthreshold (sub-V_{th}) design techniques and strategies have advanced rapidly in recent years. A wide range of applications, from radio frequency identification (RFID) tags to cellular phones, demand minute energy budgets and have driven researchers to investigate sub-V_{th} circuits. Though sub-V_{th} design has not yet gained widespread commercial adoption, recent work has shown that the potential benefits of sub-V_{th} circuits are substantial. Sub-V_{th} processors have been demonstrated with the supply voltage (V_{dd}) as low as 180mV [1] and with energy consumption of only 2.6pJ/instruction [2]. The high energy efficiency achieved in the sub-V_{th} regime comes at the price of severely degraded performance. The speed of sub-V_{th} circuits, which is exponentially dependent upon V_{th} and V_{dd}, has generally been reported in the kHz and low MHz range [1][2]. Furthermore, timing variability grows dramatically as V_{dd} reduces, forcing the adoption of pessimistic design practices and large timing margins.

The poor energy-performance trade-off in the sub-V_{th} regime has left many designers looking forward to future scaled devices. The scaling of transistor dimensions and electrical characteristics has been primarily responsible for performance improvements in standard super-threshold (super-V_{th}) MOSFETs over the past several decades. The International Technology Roadmap for Semiconductors calls for annual frequency improvements of 14% in low operating power circuits and 17% in high-performance circuits operating in the super-V_{th} region. If device scaling yields similar benefits at low voltages, then designs requiring MHz-class and GHz-class processors may be able to achieve high energy efficiency by leveraging sub-V_{th} circuits. However, device scaling is generally driven by the needs of high-performance applications. The focus of high-performance scaling has been gate length reduction, and more recently, leakage management. It is not clear that these goals align precisely with the needs of sub-V_{th} circuits.

Sub-V_{th} device optimizations were considered in [3][4], and it was shown that the optimal sub-V_{th} device should minimize inverse subthreshold slope. Additionally, the use of drain-source underlap was suggested for sub-V_{th} devices in [5]. The use of ultra-thin body FinFETs in sub-V_{th} logic was advocated for improved channel control and variability characteristics in [4][6]. However, no study has yet suggested how the scaling of physical dimensions and electrical parameters will affect sub-V_{th} circuits. In this paper, we study the evolution of static noise margins (SNM), performance, and energy in sub-V_{th} circuits as devices scale deep into the nanometer regime. We place a strong emphasis on understanding the consequences of traditional performance-driven scaling and also propose an improved scaling strategy targeting the needs of sub-V_{th} circuits.

We first use realistic two-dimensional device models (in MEDICI) scaled from the 90nm technology node down to the 32nm technology node to quantify the device-level and gate-level implications of performance-driven device scaling. We show that the slow scaling of gate oxide relative to the channel length leads to a 60% reduction in I_{on}/I_{off} between the 90nm and 32nm nodes, which results in SNM degradation of more than 10% between the 90nm and 32nm nodes in a CMOS inverter. We propose a modified scaling strategy that uses increased channel lengths and reduced doping to improve inverse subthreshold slope. We develop new delay and energy metrics that effectively capture the important effects of device scaling, and we use those to drive device optimization. We find that noise margins improve by 19% and energy improves by 23% in 32nm sub-V_{th} circuits when applying our modified device scaling strategy. Our proposed strategy also uses tight control of inverse subthreshold slope and off-current to reduce delay by 18% per generation. Our approach is particularly attractive since it requires only simple modifications to existing device technologies.

The remainder of this paper is organized as follows. In Section 2, we describe the implications of performance-driven scaling in the sub-V_{th} regime. In Section 3, we propose an alternative scaling strategy driven by the needs of sub-V_{th} circuits and compare it to a super-V_{th} scaling strategy. Finally, in Section 4, we conclude the paper.

2. SUPER-V_{TH} SCALING

In this section, we first describe the theory behind device scaling and then use two dimensional device simulations to understand the effects of super-V_{th} scaling strategies on device and circuit behavior in the sub-V_{th} regime.

2.1 Scaling Theory

Device scaling is based upon simple principles; by reducing the sizes of devices and interconnect (and therefore capacitance), performance, density and power can be improved. In general, we

*This work was supported in part by the NSF, SRC, and Intel.

DAC 2007, June 4–8, 2007, San Diego, California, USA
Copyright 2007 ACM 978-1-59593-627-1/07/0006…5.00

describe scaling by referring to several key device parameters, shown in Fig. 1. The scaling of transistor dimensions was first conceived as constant-field scaling in [7], where the maximum electric field in the channel is maintained across technology generations. An updated version of scaling, called generalized scaling [8], is highlighted in Table 1 and is a more realistic representation of modern scaling. The scaling of physical dimensions like gate length (L_{eff}), gate width (W), gate oxide thickness (T_{ox}), and wire dimensions are controlled by a factor, α. In contrast to constant field scaling [7], the maximum electric field in the channel is allowed to increase by a factor, ε, each technology generation. As a result, channel doping increases by the factor εα. Ideally, circuit area, delay, and power scale according to the values in Table 1 [9]. However, device scaling has not followed generalized scaling precisely; rather, L_{eff} has been scaled more aggressively than T_{ox}, V_{dd}, and V_{th} [10]. Furthermore, scaling has become an exercise in strain engineering, experimentation with new gate oxide materials, and novel device design [9]. As we will see in subsequent sections, the slow scaling of T_{ox} relative to L_{eff} is particularly problematic in the sub-V_{th} regime since the gate is losing control of the channel.

Table 1: Generalized scaling [8,9]

Parameter	Scaling Factor
Physical Dimensions (L_{poly}, T_{ox}, etc)	$1/\alpha$
N_{ch}	$\varepsilon\alpha$
V_{dd}	ε/α
Area	$1/\alpha^2$
Delay	$1/\alpha$
Power	ε^2/α^2

2.2 A Super-V_{th} Scaling Model

We now describe a simple but accurate bulk transistor model, illustrated in Fig. 1(a), which captures the important effects of conventional super-V_{th} scaling. Our text and figures will focus on the NFET device for the remainder of this paper, but we use an analogous methodology to describe the PFET device. The device model has four key scaling parameters: physical gate length (L_{poly}), gate oxide thickness (T_{ox}), substrate doping (N_{sub}), and peak halo doping ($N_{p,halo}$). These parameters receive special attention because they are most important when determining key device characteristics like V_{th}, on-current, off-current, and gate capacitance. In addition to these four parameters, we specify V_{dd} as an additional knob for adjusting performance. All physical dimensions other than T_{ox} (source/drain junction depth, lateral source/drain diffusion, halo dimensions, etc.) scale in proportion to L_{poly}.

Note that halo doping regions are located near the source and drain edges. Halo doping is used to control V_{th} roll-off observed at short channels and large drain biases, and has become indispensable for super-V_{th} devices. The V_{th} of a short channel device with halo doping may be represented as the sum of three components: intrinsic (long channel) threshold voltage (V_{th0}), roll-off due to short channel effects and DIBL ($\Delta V_{th,SCE}$), and roll-up due to halo doping ($\Delta V_{th,halo}$) [11]. In a well optimized device, the halo regions increase the effective channel doping at short channel lengths such that $-\Delta V_{th,SCE}=\Delta V_{th,halo}$, and V_{th} remains flat as a function of both L_{poly} and V_{ds}. We model the halo regions as a pair of two dimensional Gaussian distributions superimposed on a uniformly doped substrate similar to [3][12]. The doping contours of a representative 90nm device are shown for illustrative purposes in Fig. 1(b). The net halo doping, N_{halo}, is the sum of N_{sub} and $N_{p,halo}$.

For our purposes, describing a device at a particular technology node only requires that the four key parameters and V_{dd} are specified. We use the iterative process in Fig. 1(c) to optimize device parameters at a given technology node. L_{poly} and T_{ox} are first determined based upon published industry data. V_{dd} and V_{th} (through N_{sub} and $N_{p,halo}$) are then chosen to optimize delay under leakage constraints. We describe the selection of each parameter in the remainder of this section.

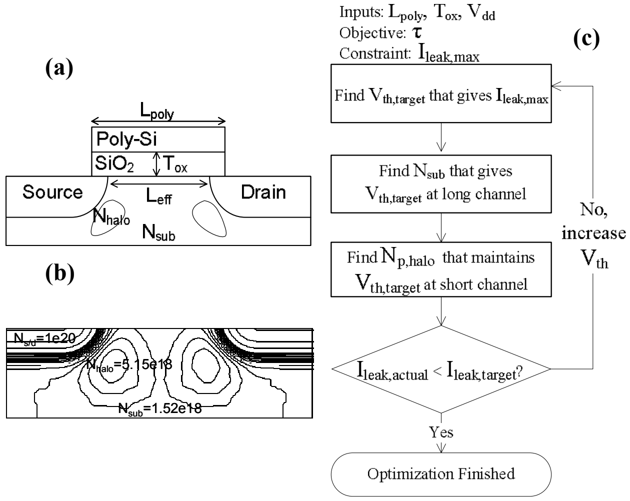
(a)
L_poly
Poly-Si
SiO2
T_ox
Source
L_eff
Drain
N_halo
N_sub
(b)
N_s/d=1e20
N_halo=5.15e18
N_sub=1.52e18
Inputs: L_poly, T_ox, V_dd
Objective: τ
Constraint: I_leak,max
(c)
Find V_th,target that gives I_leak,max
Find N_sub that gives V_th,target at long channel
Find N_p,halo that maintains V_th,target at short channel
I_leak,actual < I_leak,target?
No, increase V_th
Yes
Optimization Finished

Figure 1: (a) A device cross-section showing scaling parameters (b) Doping profile for a 90nm NFET (c) The iterative process used to select N_{sub} and $N_{p,halo}$ given a delay (τ) objective and a leakage ($I_{leak,target}$) constraint

The aggressive scaling of L_{poly} has been one of the primary drivers of performance improvement in MOSFETs. Note that L_{poly} represents the length of the bottom of the poly-Si gate after etching. For example, a gate with a designed length of 90nm might have L_{poly}=65nm after etching. Throughout this paper, we assume that the minimum L_{poly} is reduced by 30% per generation, which agrees well with recent L_{poly} scaling trends.

Selecting a realistic value for T_{ox} plays a critical role in determining the sub-V_{th} characteristics of a device. As suggested in the previous section, T_{ox} has actually scaled more slowly than L_{poly} due to oxide reliability and gate leakage concerns. A survey of recent industrial publications in [13] shows that T_{ox} has been reduced by ~10% per generation below the 130nm technology node. In this paper, we make the simple assumption that T_{ox} reduces by 10% per generation. Note that the oxide scaling problem may be even worse than our assumption of 10%. High-κ dielectrics may be the only solution since conventional gate stacks may be limited to a minimum of ~1nm thickness [20].

With L_{poly} and T_{ox} fixed for each generation, the remaining three parameters (N_{sub}, $N_{p,halo}$, V_{dd}) may be tuned to match delay and leakage requirements. As shown in Fig. 1(c), our optimization uses delay (τ) as an objective and leakage ($I_{leak,max}$) as a constraint. Note that N_{sub} is treated as a function of the long channel device (where halo doping is largely unnecessary), and $N_{p,halo}$ is treated as a function of the short channel device. While the approach described in Fig. 1(c) may not converge on the optimal solution, it is a systematic, simple heuristic that produces realistic scaled devices.

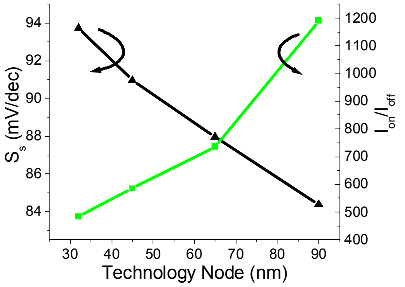

Figure 2: NFET inverse sub-V_{th} slope and on-current to off-current ratio

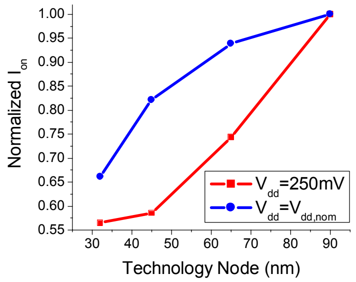

Figure 3: NFET on-current

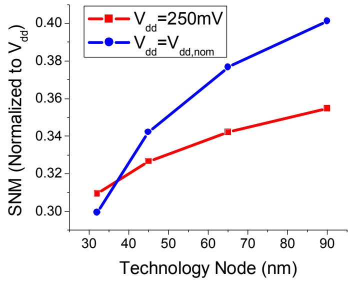

Figure 4: Simulated SNM for a scaled inverter

The selection of $I_{leak,max}$ is a complex topic since every new technology provides a range of devices optimized for different power-delay points. For example, the 65nm technology described in [14] offers low power and high power devices, with each device having 3 different V_{th} variants. The International Technology Roadmap for Semiconductors (ITRS) [15], which maps out near-term and long-term goals for the semiconductor industry, describes three different devices with different power-delay trade-offs: high performance, low operating power (LOP), and low standby power (LSTP). The LOP and LSTP devices are optimized in a similar manner, though the LSTP device has more stringent leakage constraints. In this paper, we use a super-V_{th} scaling strategy similar to that of the LSTP device. The ITRS predictions rely on the introduction of advanced technologies like high-κ gate stacks to meet stringent leakage constraints. Since we are studying the effects of current scaling trends (rather than projected scaling goals that require the introduction of advanced technologies), we relax leakage constraints slightly. We set a maximum leakage current of 100pA/µm at the 90nm node and allow leakage to grow by 25% each generation. We reduce V_{dd} regularly at each generation to control dynamic energy, and we optimize the device for minimum delay under the leakage constraint. Table 2 shows values for the NFET model parameters generated for the 90nm through 32nm nodes using the scaling approach described in this section. Throughout this paper, we refer to the results in Table 2 as the "super-V_{th} scaling strategy."

The intrinsic delay of a device may be quantified as $\tau = C_g V_{dd}/I_{on}$ where C_g is the gate capacitance including gate/drain-source overlap and I_{on} is the drain current at $V_{gs}=V_{ds}=V_{dd}$. This metric, which has been shown to correlate well with CMOS gate delay [10], is shown for reference in Table 2.

2.3 Device and Circuit-Level Implications

The device models from the previous section have been simulated in MEDICI, a two-dimensional device simulator. In this section, we first examine the low-level behavior of these devices in the sub-V_{th} region. We then highlight the gate and circuit level implications of scaling and make comparisons between super-V_{th} and sub-V_{th} behavior. In particular, we focus on SNM, delay, and energy consumption in sub-V_{th} circuits.

Table 2: NFET parameters under super-V_{th} scaling

Node	90nm	65nm	45nm	32nm
L_{poly} (nm)	65	46	32	22
T_{ox} (nm)	2.10	1.89	1.70	1.53
N_{sub} (cm^{-3})	1.52e18	1.97e18	2.52e18	3.31e18
N_{halo} (cm^{-3})	3.63e18	5.17e18	7.83e18	12.0e18
V_{dd}	1.2	1.1	1.0	0.9
$V_{th,sat}$ (mV)	403	420	438	461
I_{off} (pA/µm)	100	125	156	195
C_gV_{dd}/I_{on} (ps)	1.3	0.97	0.75	0.62

2.3.1 Device-Level Behavior

The current in a sub-V_{th} circuit may be described by the well-known weak inversion current expression shown in Eq. 1 [19], where m is the subthreshold slope factor and C_{dep} is the depletion capacitance. Note the exponential dependence on m and V_{th}.

$$I_{sub} = \frac{W}{L_{eff}} \cdot \mu_{eff} \cdot C_d \cdot v_T^2 \cdot e^{\left(\frac{V_{gs}-V_{th}}{m \cdot v_T}\right)} \cdot \left(1 - e^{-\frac{V_{ds}}{v_T}}\right) \tag{1}$$

The inverse subthreshold slope (S_S), an excellent measure of channel control, may be expressed for short channel MOSFETs as [19]:

$$S_S = 2.3 \cdot v_T \cdot m \tag{2a}$$

$$S_S = 2.3 \cdot v_T \cdot \left(1 + \frac{3 \cdot T_{ox}}{W_{dep}}\right)\left(1 + \frac{11 T_{ox}}{W_{dep}} e^{-\frac{\pi \cdot L_{eff}}{2(W_{dep}+3T_{ox})}}\right) \tag{2b}$$

where $W_{dep} \propto 1/\sqrt{N_{eff}}$ is the depletion width with effective channel doping, N_{eff}. The value of S_S which is theoretically limited to values larger than ~60mV/dec at T=300K, should be as small as possible to ensure the steepest sub-V_{th} characteristic. As shown in Eq. 2(b), the final exponential term forces S_S to increase as L_{poly} (and consequently L_{eff}) reduces relative to T_{ox} and W_{dep}. Figure 2 shows the simulated S_S for an NFET device at different technology nodes. Between the 90nm and 32nm nodes, S_S degrades by 11%, which corresponds to a 60% reduction in the on-current to off-current ratio (I_{on}/I_{off}) at V_{dd}=250mV. I_{on} is measured at $V_{gs}=V_{ds}=V_{dd}$. Note in Table 2 that all devices have V_{th}>400mV, so V_{dd}=250mV is well within the sub-V_{th} regime. We will show later in this section that the dramatic reduction in I_{on}/I_{off} leads to serious problems for noise margins and energy efficiency. Figure 3 highlights the behavior of I_{on} at both nominal V_{dd} (with values taken from Table 2) and V_{dd}=250mV. Under our leakage constrained scaling scenario, I_{on} reduces between technology generations in the super-V_{th} region. Note that our choice of leakage constraint (100pA plus 25% per generation) affects this outcome. A more aggressive technology, especially one leveraging strain in the channel, would likely achieve increased drain current with scaling. However, in this study, we are concerned with low power devices. Note that the reduction in current is more dramatic for the device measured in the sub-V_{th} region. This loss of drain current has important delay implications that will be discussed later in this section.

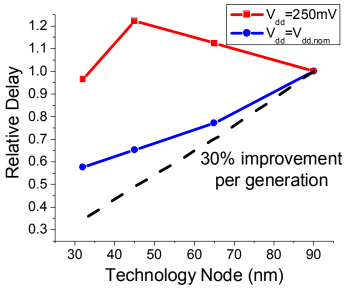

Figure 5: Simulated delay for a scaled inverter

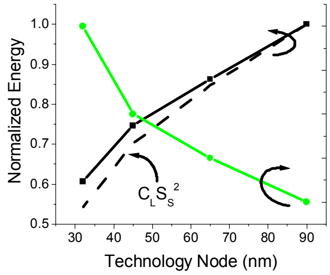

Figure 6: Simulated energy/cycle and V_{min} for a chain of 30 inverters with α=0.1

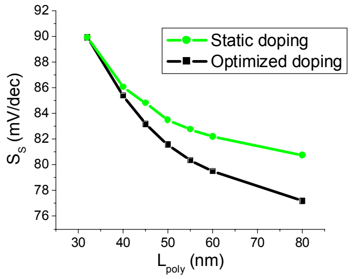

Figure 7: S_S as a function of gate length for a 45nm device

2.3.2 Static Noise Margins

Consider the static noise margins (SNM) of a CMOS inverter. The voltage transfer characteristic of a sub-V$_{\text{th}}$ inverter is computed by equating drain current (Eq. 1) through NFET and PFET devices, as shown in Eq. 3(a). $I_{o,N}$ and $I_{o,P}$ are the NFET and PFET currents at $V_{gs}=V_{th}$ with $V_{ds}>>v_T$. V_{in} and V_{out} are the voltages at the input and output of the inverter. We can relate V_{in} and V_{out} using Eq. 3(b). We can further simplify the expression by assuming $I_{o,N}=I_{o,P}$, $V_{th,N}=V_{th,P}=V_{th}$ and $m_N=m_P=m$ (Eq. 3(c)).

$$I_{o,N}\cdot e^{\frac{V_{in}-V_{th}}{m\cdot v_T}}\left(1-e^{-\frac{V_{out}}{v_T}}\right)=I_{o,P}\cdot e^{\frac{V_{dd}-V_{in}-V_{th}}{m\cdot v_T}}\left(1-e^{-\frac{V_{dd}-V_{out}}{v_T}}\right) \tag{3a}$$

$$V_{in}=\frac{m_n\left(V_{dd}-V_{th,p}\right)+m_pV_{th,n}+m_nm_pv_T\ln\left(\frac{I_{o,P}}{I_{o,N}}\cdot\frac{1-e^{-\frac{V_{dd}-V_{out}}{v_T}}}{1-e^{-\frac{V_{out}}{v_T}}}\right)}{m_n+m_p} \tag{3b}$$

$$V_{in}=\frac{V_{dd}}{2}+\frac{m\cdot v_T}{2}\ln\left(\frac{1-e^{-\frac{V_{dd}-V_{out}}{v_T}}}{1-e^{-\frac{V_{out}}{v_T}}}\right) \tag{3c}$$

The important role of S_S (through m) in determining the voltage transfer characteristic (and consequently SNM) is obvious, particularly in Eq. 3(c). Figure 4 shows the evolution of SNM for a CMOS inverter simulated at nominal V_{dd} (Table 2) and V_{dd}=250mV. We define SNM at the points where the gain in the voltage transfer characteristic equals negative one. The increase in S_S with scaling results in SNM degradation of more than 10% between the 90nm and 32nm nodes. This is a serious concern for sub-V$_{\text{th}}$ designers since absolute noise margins are already dramatically reduced compared to high voltage operation. It is particularly concerning for SRAM, where noise margins are paramount and a small I_{on}/I_{off} in sub-V$_{\text{th}}$ circuits already places tight limits on the maximum number of bits/line [16].

2.3.3 Delay

The delay of a CMOS gate may be expressed as:

$$t_p=\frac{k_d\cdot C_L\cdot V_{dd}}{I_{on}} \tag{4}$$

where C_L is the load capacitance and k_d is a fitting parameter. The sub-V$_{\text{th}}$ delay may be found by substituting Eq. 1 into Eq. 4:

$$t_p=\frac{k_d\cdot C_L\cdot V_{dd}}{I_{on}}=\frac{k_d\cdot C_L\cdot V_{dd}}{I_{o,N}\cdot e^{\frac{V_{dd}-V_{th}}{m\cdot v_T}}} \tag{5}$$

The V_{ds} dependence of I_{on} (shown in Eq. 1) has been ignored since it is negligible for $V_{gs}=V_{dd}>>v_T$. The delay expression is clearly dominated by an exponential dependence on V_{dd}, V_{th}, and m.

The simulated delay of a CMOS inverter with FO1 loading is shown in Fig. 5 at nominal V_{dd} (Table 2) and at 250mV. As expected, the delay at nominal V_{dd} improves with L_{poly}, though at a rate that is slower than the target of 30% per generation under generalized scaling (assuming 1/α=0.7). With the exception of the 32nm device, the delay actually increases with device scaling at V_{dd}=250mV due to strict leakage constraints during device optimization as well as degraded S_S. We must be careful in making any claims about delay trends in future sub-V$_{\text{th}}$ circuits, since sub-V$_{\text{th}}$ delay is exponentially sensitive to V_{th}. Even small changes to a super-V$_{\text{th}}$ device to control leakage and short channel effects may result in large fluctuations in sub-V$_{\text{th}}$ delay. It is likely that V_{th} scaling, not L_{poly} scaling, will control the performance of future sub-V$_{\text{th}}$ circuits. Strict attention to V_{th} selection will be an important part of any technology optimized for sub-V$_{\text{th}}$ use.

In sub-V$_{\text{th}}$ applications, V_{dd} is typically set at the energy optimal value, V_{min}, so the scaling of delay at $V_{dd}=V_{min}$ is of interest. The value of V_{min} was found in [17][18] to be proportional to S_S. If we ignore the dependence of V_{min} on the slope of the input waveform, then we can set $V_{dd}=V_{min}=K_{Vmin}\cdot S_S$ where K_{Vmin} is a parameter that depends only on the structure of the circuit (and not on scaling parameters) [17]. Using this new relation and by recognizing that $S_S=V_{dd}/log(I_{on}/I_{off})$, we can express Eq. 4 and Eq. 5 in terms of only scaling dependent parameters (Eq. 6). The simple expression in Eq. 6 suggests that we can predict the scaling behavior of sub-V$_{\text{th}}$ delay simply by understanding the scaling of C_L, S_S, and I_{off}. We develop a similar expression for energy in the next sub-section.

$$t_p=\frac{k_d\cdot C_L\cdot K_{V\min}\cdot S_S}{I_{off}\cdot 10^{\frac{K_{V\min}S_S}{S_S}}}\propto\frac{C_L\cdot S_S}{I_{off}} \tag{6}$$

2.3.4 Energy

The energy of a single inverter driving an identical inverter can nominally be separated into two components: dynamic (E_{dyn}) and leakage (E_{leak}).

$$E_{dyn}=C_L\cdot V_{dd}^2\cdot\alpha \tag{7a}$$

$$E_{leak}=I_{off}\cdot V_{dd}\cdot t_p=I_{off}\cdot V_{dd}\cdot\frac{k_d\cdot C_L\cdot V_{dd}}{I_{on}}=C_L\cdot V_{dd}^2\cdot k_d\cdot\frac{I_{off}}{I_{on}} \tag{7b}$$

The term α is the activity factor and all other terms are previously defined. If we again assume that operation only occurs at the energy optimal $V_{dd}=V_{min}$, then we can simplify Eq. 7(a) and Eq. 7(b) as follows:

$$E_{dyn}=C_L\cdot\left(K_{V\min}\cdot S_S\right)^2\cdot\alpha\propto C_L\cdot S_S^2 \tag{8a}$$

$$E_{leak}=C_L\cdot\left(K_{V\min}\cdot S_S\right)^2\cdot k_d\cdot 10^{-K_{V\min}}\propto C_L\cdot S_S^2 \tag{8b}$$

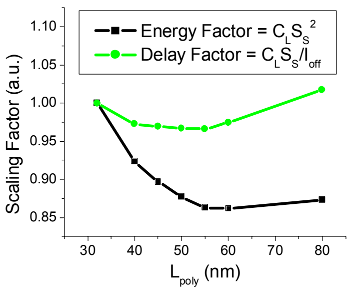

Figure 8: Energy and delay factors for a 45nm device

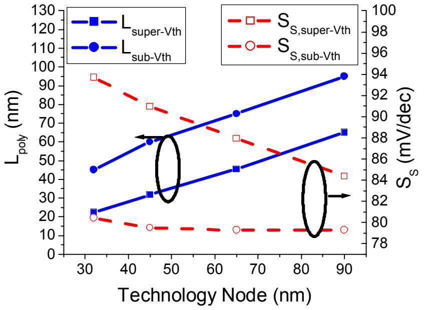

Figure 9: NFET L_{poly} and S_S and for sub-V_{th} and super-V_{th} scaling

Table 3: NFET parameters under sub-V_{th} scaling

Node	90nm	65nm	45nm	32nm
L_{poly} (nm)	95	75	60	45
T_{ox} (nm)	2.10	1.89	1.70	1.53
N_{sub} (cm^{-3})	1.61e18	1.99e18	2.53e18	3.19e18
N_{halo} (cm^{-3})	2.02e18	2.73e18	2.93e18	4.89e18
$C_L S_S^2$ (a.u.)	1	0.80	0.65	0.51
$C_L S_S$ (a.u.)	1	0.80	0.65	0.50

The only parameters that change as a result of device scaling are C_L and S_S. Equation 8 suggests the interesting result that dynamic energy and leakage energy in sub-V_{th} circuits have an identical dependence on scaling parameters and that the ratio E_{dyn}/E_{leak} is insensitive to scaling when operating at $V_{dd}=V_{min}$.

The simulated energy consumed per cycle by a chain of 30 inverters with α=0.1 and $V_{dd}=V_{min}$ is plotted in Fig. 6. There is a substantial energy reduction as devices are scaled from the 90nm to the 32nm node. However, note that V_{min} increases by 40mV for this simple circuit between the 90nm and 32nm nodes. Recall that V_{min} is proportional to S_S, so this trend is not surprising. It was shown in [6] that an increase in V_{min} is generally not beneficial for energy efficiency. An increase in V_{min} essentially equates to a dynamic energy ($C_L V_{dd}^2$) penalty. Ideally, a scaled sub-V_{th} device should experience a reduction in capacitance while maintaining V_{min}. The factor $C_L \cdot S_S^2$, which is also plotted in Fig. 6, matches very closely to the energy measurements, thus confirming the validity of Eq. 8.

3. SUB-V_{TH} SCALING

It became clear in the last section that the degradation of S_S with device scaling will be problematic for robust, energy efficient sub-V_{th} operation. Moreover, the scaling of L_{poly} to improve the delay characteristics of super-V_{th} devices is not relevant in sub-V_{th} circuits since delay is largely controlled by V_{th}. Ideally, we would like a sub-V_{th} transistor with a very small S_S to address noise margin and energy concerns. This device should be available in multiple well controlled thresholds in order to provide a wide range of performance points. In this section, we describe such a device and develop a scaling strategy for this device.

3.1 Sub-V_{th} Device Optimization

The degradation of S_S with scaling is driven by two related factors. The first factor has already been made clear: the ratio L_{eff}/T_{ox} reduces with each technology generation due to the slow scaling of T_{oz} and worsens the V_{th} roll-off problem. This suggests that longer channel lengths should be used to accommodate the gate oxide. The second factor causing S_S degradation, which was also covered in [3], is more subtle. To compensate for the V_{th} roll-off problem, the channel doping is effectively increased through aggressive use of halo doping. Recall that the depletion region width, W_{dep}, is inversely related to the channel doping and that, in general, S_S degrades as W_{dep} reduces (Eq. 2(b)). For long-channel devices, the halo doping is less critical and actually degrades S_S. Therefore, to fully optimize S_S with device scaling, it is not sufficient to simply lengthen L_{poly} without considering the doping. Instead, L_{poly} and doping must be optimized simultaneously. This notion is confirmed in Fig. 7, which shows S_S for a 45nm device with a fixed doping profile and for a 45nm device with a doping profile optimized for each value of L_{poly}.

Increasing L_{poly} and reducing doping improves S_S at the cost of increased gate capacitance. The cost of this optimization can be quantified in terms of energy and delay. Equation 6 shows us that sub-V_{th} delay is proportional to $C_L \cdot S_S/I_{off}$ at $V_{dd}=V_{min}$. Similarly, Eq. 8(a) and Eq. 8(b) show that energy in a sub-V_{th} circuit is proportional to $C_L \cdot S_S^2$. These expressions are useful since they are simple functions of device parameters and offer a quick estimation of energy and delay in a prospective technology. Figure 8 plots these energy and delay factors as functions of L_{poly} for the optimized 45nm device originally highlighted in Fig. 7. Both reach a minimum, suggesting that there is both a delay optimal and energy optimal L_{poly}. However, since the delay minimum is very shallow, we can select the energy minimal L_{poly} (60nm in Fig. 8) for a negligible penalty. Note that delay typically degrades as ~$1/L_{poly}$, but we are able to avoid this problem by also optimizing the doping.

3.2 A Sub-V_{th} Scaling Model

Given the important role that S_S plays in determining energy efficiency, performance, and noise margins, we propose a scaling strategy that reduces S_S by targeting the energy optimal L_{poly} at each technology node. The proposed strategy uses longer channel lengths that scale more slowly than the rate of 30% assumed in Section 2. As we will see, one consequence of this strategy is that S_S remains approximately constant with device scaling. For this study, we maintain a constant I_{off} of 100pA/µm across all device generations. Fixing I_{off} yields a more predictable delay scaling characteristic and avoids the problems illustrated in Fig. 5. Just as in super-V_{th} technologies, different performance levels can be targeted by offering multiple thresholds.

We begin with a 90nm device identical to the 90nm device in Section 2.2 but L_{poly} and doping have been optimized for minimum energy using Eq. 8(a) and Eq. 8(b). We again assume that T_{ox} reduces by 10% and all other physical dimensions, excluding L_{poly}, reduce by 30% each generation. We find the optimal L_{poly}, N_{sub}, and $N_{p,halo}$ at each generation as described in Section 3.1. The resulting NFET device parameters are listed in Table 3. Energy (Eq. 8) and delay (Eq. 6) factors are also listed in Table 3. Note that the delay factor simplifies to $C_L S_S$ since I_{off} is constant with scaling. A similar set of values is derived for PFET devices. We find that the energy optimal L_{poly} for the PFET device is almost identical to that of the NFET, so we use the L_{poly} values in Table 3 during PFET doping optimization. For the remainder of this paper, we refer to the results in Table 3 as the "sub-V_{th} scaling strategy."

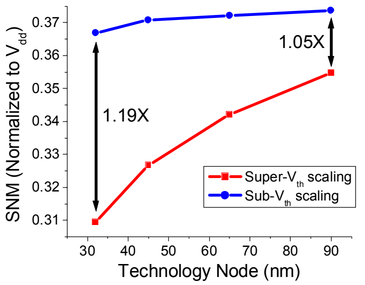

Figure 10: Simulated SNM for an inverter under super-V_{th} and sub-V_{th} scaling

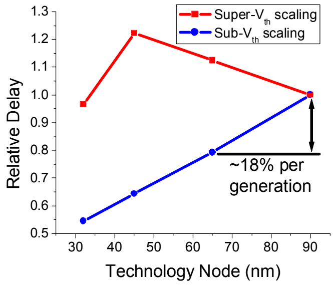

Figure 11: Simulated delay for an inverter at V_{dd}=250mV under super-V_{th} and sub-V_{th} scaling

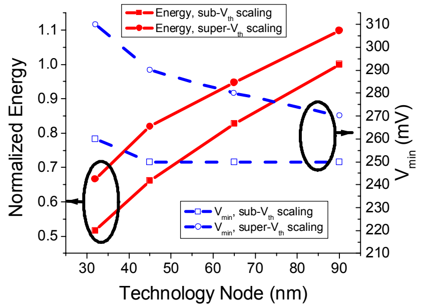

Figure 12: Simulated Energy and V_{min} under super-V_{th} scaling and sub-V_{th} scaling

3.3 Device and Circuit-Level Implications

The primary purpose of our revised scaling strategy is to maintain strong channel control, even at very small dimensions. Figure 9 shows how L_{poly} and S_S scale under our proposed scaling strategy and under the original super-V_{th} scaling strategy. L_{poly} is larger than in the super-V_{th} scaling scheme and also scales at a slower rate (20-25% per generation) than the L_{poly} in the super-V_{th} scaling scheme (30%). Note that S_S stays very close to ~80mV/dec under our proposed strategy, varying by only 1.2mV/dec between the 90nm and 32nm nodes. As a result, SNM remains nearly constant as well (Fig. 10). At the 32nm node, the optimized sub-V_{th} scaling strategy yields an SNM that is 19% larger than that observed under the super-V_{th} scaling strategy.

Normalized FO1 inverter delay is plotted in Figure 11 for both scaling scenarios. Delay reduces by ~18% per generation under our proposed strategy. Recall from Section 2.3.3 that the delay characteristic for the super-V_{th} scaling strategy is not monotonic due to the scaling of V_{th} and I_{off}. It is therefore not fair to directly compare the delay scaling of the two strategies. However, it is clear that the sub-V_{th} scaling strategy exerts much tighter control over I_{off} and S_S than the super-V_{th} strategy so the delay characteristic scales much more gracefully.

Figure 12 shows the simulated energy and V_{min} for a chain of 30 inverters under the conventional super-V_{th} scaling scheme and our proposed scheme. The proposed strategy consumes ~23% less energy than the super-V_{th} scaling strategy at the 32nm node (measured at V_{min}), with V_{min} changing by only 10mV between the 130nm and 32nm nodes. The relatively low V_{min} (which previous work has shown to be a strong function of S_S and leakage energy [17][18]) is responsible for this energy reduction.

4. CONCLUSION

Sub-V_{th} circuits are promising for future energy efficient applications. In this work we investigated the implications of device scaling on sub-V_{th} operation. In particular, we found that the slow scaling of gate oxide leads to 60% I_{on}/I_{off} degradation in the sub-V_{th} regime. We used MEDICI simulations of simple circuits to illustrate the energy, performance, and robustness characteristics of scaled sub-V_{th} devices. We proposed an alternative scaling strategy that uses larger gate lengths and reduced doping to achieve much improved inverse subthreshold slope. Our proposed strategy maintains an S_S~80mV/dec down to the 32nm node and offers a robust, energy efficient alternative to conventional devices. With very simple process modifications, sub-V_{th} circuits may be able to reliably scale deep into the nanometer regime.

5. REFERENCES

[1] A. Wang, A. Chandrakasan, "A 180mV FFT processor using subthreshold circuit techniques," *Int. Solid-State Circuits Conf.*, pp. 292-293, 2004.

[2] B. Zhai, et al., "A 2.60pJ/Inst Subthreshold Sensor Processor for Optimal Energy Efficiency," *Symp. on VLSI Circuits*, pp. 154-155, 2006.

[3] B.C. Paul, A. Raychowdhury, K. Roy, "Device optimization for digital subthreshold logic operation," *IEEE Trans. Elect. Devices*, pp. 237-247 (2005).

[4] J.J. Kim, K. Roy, "Double gate-MOSFET subthreshold circuit for ultralow power applications," *IEEE Trans. Elect. Devices*, pp. 1468-1474 (2004).

[5] B. Paul, A. Bansal, K. Roy, "Underlap DGMOS for digital-subthreshold operation," *IEEE Trans. Elect. Devices***,** pp. 910-913 (2006).

[6] S. Hanson, et al., "Ultralow-voltage, minimum-energy CMOS," *IBM J. Res. & Dev.*, pp. 469-490 (2006).

[7] R.H. Dennard, et al., "Design of ion-implanted MOSFET's with very small physical dimensions," *J. Solid-State Circuits*, pp. 256-268 (1974).

[8] G. Baccarani, M.R. Wordeman, R.H. Dennard, "Generalized scaling theory and is application to a ¼ micrometer MOSFET design," *IEEE Trans. Elect. Devices*, pp. 452-461 (1984).

[9] W. Haensch, et al., "Silicon CMOS devices beyond scaling," *IBM J. Res. & Dev.*, pp. 339-361 (2006).

[10] E.J. Nowak, "Maintaining the benefits of CMOS scaling when scaling bogs down," *IBM J. Res. & Dev*, pp. 169-180 (2002).

[11] B. Yu, et al., "Short-Channel Effect Improved by Lateral Channel-Engineering in Deep-Submicrometer MOSFET's," *IEEE Trans. Elect. Devices*, pp. 627-634 (1997).

[12] Z.K. Lee, M.B. McIlrath, D.A. Antoniadis, "Two-Dimensional Doping Profile Characterization of MOSFET's by Inverse Modeling Using I-V Characteristics in the Subthreshold Region," *IEEE Trans. Elect. Devices*, pp. 1640-1649 (1999).

[13] W. Zhao, Y. Cao, "New Generation of Predictive Technology Model for Sub-45nm Design Exploration," *Int. Symp. on Quality Elect. Design,* pp. 585-590, 2006.

[14] Z. Luo, et al., "High Performance and Low Power Transistors Integrated in 65nm Bulk CMOS Technology," *Int. Electron Devices Meeting*, pp. 661-664, 2004.

[15] *The Int. Technology Roadmap for Semiconductors*, 2005.

[16] B. Zhai, D. Blaauw, D. Sylvester, S. Hanson, "A sub-200mV 6T SRAM in 130nm CMOS," *Int. Solid-State Circuits Conf.,* 2007.

[17] B. Zhai, et al., "The Limit of Dynamic Voltage Scaling and Insomniac DVS," *IEEE Trans. VLSI Syst.*, 1239-1252 (2005).

[18] B. Calhoun, A. Chandrakasan, "Characterizing and Modeling Minimum Energy Operation for Subthreshold Circuits," *Int. Symp. on Low Power Electronics and Design*, pp. 90-95, 2004.

[19] Y. Taur, T.H. Ning, *Fundamentals of Modern VLSI Devices*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[20] E.P. Gusev, et al., "Advanced High-κ Dielectric Stacks with PolySi and Metal Gates: Recent Progress and Current Challenges," *IBM J. Res. & Dev.*, pp. 387-410 (2006).