Sleep Mode Analysis and Optimization with Minimal-Sized Power Gating Switch for Ultra-low V_{dd} Operation

Mingoo Seok, Scott Hanson, David Blaauw, Dennis Sylvester

University of Michigan, Ann Arbor, MI

mgseok@umich.edu

Telephone: (734) 615-8930, Fax: (734) 763-9324

ABSTRACT

This paper investigates the optimization of sleep mode energy consumption for ultra-low V_{dd} CMOS circuits, which is motivated by our findings that minimization of sleep mode energy holds great potential for reducing total energy consumption. We propose a unique approach of using a power gating switch (PGS) in ultra-low V_{dd} regimes. Unlike the conventional manner of using PGSs, our optimization suggests using minimal-sized PGSs with a slightly higher V_{dd} to compensate for voltage drop across the PGS. In SPICE simulations, this reduces total energy consumption by ~125× compared to conventional approaches. The effectiveness of the proposed optimization is also confirmed by measurements taken from an ultra-low power microprocessor. Additionally, the feasibility of using minimal PGSs in ultra-low V_{dd} regimes is investigated using SPICE simulations and silicon measurements.

KEYWORDS

Ultra-low power, subthreshold operation, sleeps mode, standby mode, power gating switch, MTCMOS

Acknowledgement of Financial Support

The authors acknowledge support from the National Science Foundation.

Affiliation of Authors

- Mingoo Seok – Research Assistant at Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, US
- Scott Hanson – Post-doctoral researcher at Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, US
- David Blaauw – Professor at Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, US
- Dennis Sylvester –Professor at Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, US

Major Changes from Our Conference Publication [28]

- We extend the analysis on the total energy derivation of the circuit with a power gating switch in ultra-low Vdd regimes, in particular, by comparing the effect of Vdd change and PGS width in Chapter 3.1. New SPICE simulation results are added.
- Section 4.5 includes measurement results and discussion of an ultra-low power microprocessor employing the proposed power gate switches.
- In Sections 4.1 and 4.2 we explicitly discuss the difference of the strategy using PGS between nominal supply voltage and ultra-low supply voltage. New SPICE simulation results are shown to support the idea.
- We discuss the impact on the virtual ground rail when using minimal-sized power gating switches (PGS) in Section 5. To support the idea, we add new SPICE simulation results from inverter chains as well as new measurement results from a low power micro-system.
- Seven new figures are added: Figures 8,9,10,13,14,15, 16

1. Introduction

Voltage scaling is well known as an effective method to reduce energy-per-operation due to the quadratic relationship between switch energy (E_{switch}) and supply voltage. Therefore, dynamic voltage scaling (DVS) has been used in microprocessors to scale down the supply voltage to the point where a task is completed just before the deadline, thereby saving a significant amount of energy [1][2].

However voltage scaling has limitations in providing energy savings [3]. Metal Oxide Semiconductor Field Effect Transistors (MOSFET) become exponentially slow once the supply voltage scales below the threshold voltage (V_{th}) of devices due to the small subthreshold current, as captured by the well-known subthreshold current equation (EQ1). This performance degradation causes a rapid increase of leakage energy (E_{leak}), which eventually offsets the savings of E_{switch}. Therefore, the total energy consumption starts to increase once the supply voltage scales down below a certain point, which we refer to as V_{min}. The optimal energy consumption, which occurs at V_{min}, is defined as E_{min}. This relationship is illustrated in Figure 1 and analytically modeled in EQ2. [3]

$$I_{sub} = \mu \cdot Cox \cdot W/L \cdot (m-1) \cdot V_T^{\,2} \cdot \exp(V_{gs} - V_{th} / mV_T) \cdot (1 - \exp(-V_{ds} / V_T) \quad \text{[EQ1]}$$

$$E = E_{switch} + E_{leak} = \frac{1}{2} nCV_{dd}^{\,2} \left[\alpha + \eta \cdot n \cdot e^{(-\frac{V_{dd}}{mV_T})} \right] \quad \text{[EQ2]}$$

where μ: mobility, Cox: oxide capacitance, W: width, L: length,
m: subthreshold slope factor, V_T: thermal voltage, V_{th}: threshold voltage,
V_{gs}: gate-source voltage, V_{ds}: drain-source voltage, n: length of inverter chain,
η: fitting coefficient

Operating CMOS circuits at V_{min} usually leads to large performance degradation. For example, recent publications show that microprocessors operate at clock frequencies of hundreds of kHz at 300~400mV [4][5][6][7]. However many energy-constrained applications, such as biomedical and environmental sensor systems, have relaxed performance requirements [8]. Therefore, ultra-low V_{dd} operation represents a viable option for them.

Studies of ultra-low V_{dd} operation have been conducted at the technology, circuit, and architecture levels. At the technology level, the existing scaling strategies of modern CMOS that emphasize high performance can be sub-optimal for minimizing energy consumption in ultra-low V_{dd} regimes. Therefore, there have been proposals on new device designs [9][10] as well as technology selection [11][12] for ultra-low voltage operations. At the circuit level, design methodologies for better energy [13], variability [14], and performance [15][16] have been investigated. In particular, static random access memory (SRAM) has been intensively studied, since the reduced on-current to off-current ratio degrades the robustness of back-to-back inverters in ultra-low V_{dd} regimes. To improve robustness, different bitcell topologies including 6T [17][18], 8T [19][20], 10T [21], and 12T [22] SRAM have been proposed to replace the conventional 6T. Research at the architecture level has focused on simple and energy-efficient architectures for ultra-low power microprocessors [36].

Ultra-low V_{dd} computational cores [4][23] and general microprocessors [7][25][26] have been designed and tested showing that ultra-low V_{dd} designs achieve the active energy consumption of several pJ per cycle. However, these designs have often overlooked the importance of sleep energy consumption. Sleep energy, which has become important in modern CMOS processes due to the increasing contributions of subthreshold and gate leakage current, becomes more significant in ultra-low V_{dd} operations for two reasons. First, the reduced switching energy consumption from scaled supply voltages renders the sleep energy a more significant portion of total energy consumption. Second, ultra-low power applications often have low duty cycles. Although they run slowly at V_{min}, there is a considerable amount of sleep time between the moment of completing a task (T_{min}) and the start of a new task ($T_{deadline}$), as defined in Figure 2. Since there is a considerable amount of sleep energy consumption during the period, an optimization method that considers sleep energy consumption is vital to an energy-optimal design.

This paper extends one of the earliest works regarding sleep energy analysis and optimization in ultra-low V_{dd} regimes [28]. We start by proving the importance of sleep energy for reducing total energy consumption. Then, we discuss the effects of power gating switches (PGSs) [29], a well-known sleep energy reduction scheme, on energy consumption in ultra-low Vdd regimes. Our proposed optimization,

which modulates PGS size and supply voltage simultaneously, suggests using very small PGSs with a supply voltage higher than V_{min}, unlike conventional practices in which a large PGS is often used (typically ~10% of total NFET width). In SPICE simulations of generic circuits, the optimization method achieves 125× reduction in total energy consumption and 50× savings in PGS area. The effectiveness of this proposed optimization is also confirmed by measurement results from a fabricated microprocessor. We also discuss the functional feasibility of using minimal PGSs with SPICE simulations and silicon measurements. Finally, other approaches to perform power gating are quantitatively compared for energy optimal designs.

2. Impact of Sleep Energy on Total Energy Consumption

We first investigate the case in which circuits experience non-zero sleep time. In other words, T_{min}, the time when circuits complete a task at the traditional $V_{dd}=V_{min}$ comes earlier than $T_{deadline}$, the moment when the circuit begin a new task. In this case, the total energy is the sum of sleep energy (E_{sleep}) and active energy ($E_{switch} + E_{leak}$). We define duty cycle K_{duty} as $T_{deadline} / T_{min}$, which represents the ratio of maximum allotted-time to actual used-time (i.e., circuit delay at V_{min}). If $K_{duty} > 1$, then circuits experience sleep time and consume additional energy.

For this scenario, we run SPICE simulations using inverter chains to estimate the contribution of sleep energy consumption to total energy consumption. In this paper, SPICE simulations are performed using a commercial 0.13µm CMOS technology. Unless mentioned explicitly, a 99-stage inverter chain is used. Inverters use regular V_{th} devices while PGSs uses high V_{th} device. The $V_{th,high\text{-}Vth}$ is ~560mV and $V_{th,regular}$ is 350mV at nominal conditions. At $V_{dd}=V_{min}$ (220mV), E_{min} of the inverter chain is simulated as 15.4fJ/cycle at a delay of 5.66µs (176 kHz). NFETs and PFETs in inverters are sized at 0.32µm. Wiring parasitics are not included in simulations. The logic depth of the inverter chain is equivalent to 25 Fan-Out-of-4 (FO4) delays, which is shorter than most ultra-low V_{dd} designs. For a single inverter chain, the circuit activity rate is 1. The chosen logic depth and switching activity approximate the worst-case voltage drop scenario across power gating switches, and provide conservatism in the results.

We initially assume that there is no cutoff technique applied in sleep mode. The total energy consumption for inverter chains can be expressed as EQ3, which is derived from EQ2. EQ3 shows that nearly the same amount of leakage current exists for both sleep and active time. Therefore, a significant increase in total energy consumption is expected. Figure 4 shows that sleep energy contributes a large amount of energy consumption at lower duty cycles or higher K_{duty} (i.e., circuits spend more time in sleep mode). Since ultra-low power applications often have low duty cycles, it is paramount to consider sleep energy in total energy optimization frameworks.

$$\begin{aligned} E_{Total} &= E_{switch} + E_{leak} + E_{sleep} \\ &= E_{switch} + t_{delay}P_{leak} + (T_{deadline} - t_{delay}) \cdot P_{leak} \\ &= E_{switch} + T_{deadline}P_{leak} \end{aligned} \qquad \text{[EQ3]}$$

Another interesting observation is that both E_{switch} and $T_{deadline} \cdot P_{leak}$ in EQ3 are proportional to V_{dd}, resulting in lower energy-optimal supply voltage than conventional V_{min}, as shown in Figure 4. The optimal supply voltage can be scaled down until CMOS gates fail to function, while it is often bounded by the contribution of leakage energy in the conventional analysis. The minimal functional voltage for CMOS gates is assumed to be ~100mV, although this assumption has little impact on the results of this work.

3. The Effects of Cutoff Structures on Total Energy Consumption

Given the significant contribution of sleep energy to total energy consumption, PGSs are attractive for improving overall energy efficiency. While several other methods can be used in sleep mode, such as reverse body-biasing, PGSs are considered the most effective measure to reduce leakage energy consumption [37][38]. However, PGS design in ultra-low Vdd regimes differs from conventional practices. Therefore, in this section, we first study the effects of PGSs on energy consumption of circuits operating in ultra-low voltage regimes. Section 4 then lays out a strategy for using PGSs to minimize total energy consumption based on our findings in this section.

The purpose of employing PGSs in circuits is to reduce sleep power by strongly shutting off leakage paths during sleep modes. However, the benefit of reducing sleep energy consumption comes

with performance degradation due to the voltage drop across PGSs [29]. In ultra-low voltage regimes, the performance degradation can induce extra active energy consumption since circuits consume extra leakage energy for longer periods of active time. Therefore, designers should be aware of the effects of PGSs on sleep and active energy consumption in ultra-low Vdd regimes.

To capture the effects of PGSs on energy consumption, we propose two parameters in EQ4. The first parameter, denoted by K_{leak}, sleep energy reduction factor, is the ratio of sleep power with PGSs to sleep power without such structures. The second parameter, the delay degradation factor, denoted by $1/K_{delay}$, is the ratio of circuit delay with PGSs to delay without them.

$$\frac{1}{K_{delay}} = \frac{t_{delay\_w/\_PGS}}{t_{delay\_w/o\_PGS}} \qquad K_{leak} = \frac{I_{leak,w\_PGS}}{I_{leak,w/o\_PGS}} \qquad \text{[EQ4]}$$
$$(0 < K_{delay},\ K_{leak} < 1)$$

3.1 Theoretical Power Gating Switch

This section investigates Emin assuming that circuits use a theoretical PGS having independent controls on K_{leak} and $1/K_{delay}$. For example, EQ5 shows the total energy consumption of circuits with the PGS of K_{leak} and $1/K_{delay}$, where T_{min} denotes the delay of main circuits at V_{min} without the PGS; P_{leak} denotes the leakage power without the PGS; and t_{delay} expresses the delay of main circuits at a specific V_{dd} with the PGS. In EQ5, E_{switch} is technically affected by the PGS due to the change of the voltage swing. However, this can be ignored without sacrificing much accuracy. However, we include the changes of E_{switch} after this section for a more complete analysis.

$$\begin{aligned} E_{Total} &= E_{switch} + E_{leak} + E_{sleep} \\ &= E_{switch} + \frac{1}{K_{delay}} t_{delay} P_{leak} + (K_{duty} T_{\min} - \frac{1}{K_{delay}} t_{delay}) \cdot K_{leak} P_{leak} \end{aligned} \qquad \text{[EQ5]}$$

where t_{delay} : delay of circuits without PGSs, $T_{\min}$: delay of circuits at V_{min} without PGSs

We investigate the changes of V_{min} and E_{min} while sweeping either K_{leak}, as shown in Figures 5(a) and 5(b), or $1/K_{delay}$ as shown in Figures 5(c) and 5(d). Figures 5(a) and 5(b) show that small values of K_{leak} can reduce E_{sleep} and push V_{min} to a conventional V_{min}. On the other hand, Figures 5(c) and 5(d) show that

large values of $1/K_{delay}$ increases E_{leak} due to the longer delay. Since higher V_{dd} can alleviate the performance degradation, higher V_{min} is preferred to offset the increase of E_{leak} for this case.

3.2 Practical Power Gating Switch

While we assume PGSs with independent controls on K_{leak} and $1/K_{delay}$ in the previous section, they are actually co-related in practical PGS designs. For a simple PGS (Figure 3), we can derive $1/K_{delay}$ and K_{leak} in ultra-low Vdd regimes, as shown in EQ6 and EQ7. In the derivation, it is assumed that the voltage across the PGS in sleep mode is ~V_{dd}, due to the very high resistance of the PGS when it is off. V_{swing}, which is a highly non-linear function of PGS width and technology parameters, reduces for wider PGSs and lower threshold voltages.

$$\frac{1}{K_{delay}} = \frac{t_{delay\_w/\_PGS}}{t_{delay\_w/o\_PGS}} = \frac{CV_{swing}}{CV_{dd}} \cdot \frac{\mu_{main} C_{ox} \frac{W}{L} V_T^{\,2} (m-1) e^{\frac{-V_{swing}}{mV_T}} (1 - e^{\frac{-V_{swing}}{mV_T}})}{\mu_{main} C_{ox} \frac{W}{L} V_T^{\,2} (m-1) e^{\frac{-V_{dd}}{mV_T}} (1 - e^{\frac{-V_{dd}}{mV_T}})} = \frac{V_{swing}}{V_{dd}} e^{\frac{V_{dd} - V_{swing}}{mV_T}} \quad \text{[EQ6]}$$

$$K_{leak} = \frac{I_{leak\_w/\_PGS}}{I_{leak\_w/o\_PGS}} = \frac{\mu_{PGS} C_{ox} \frac{W_{PGS}}{L_{PGS}} \exp(\frac{-V_{th\_PGS}}{mV_T})(1 - \exp(-\frac{V_{dd}}{V_T})}{\mu_{main} C_{ox} \frac{W_{main}}{L_{main}} \exp(\frac{-V_{th\_main}}{mV_T})(1 - \exp(-\frac{V_{dd}}{V_T})} = k \cdot W_{PGS} \quad \text{[EQ7]}$$

where V_{swing} : voltage swing without voltage drop in PGSs, μ : mobility,
Cox : oxide capacitance, W : width, L : length, m : subthreshold slope factor,
V_T : thermal voltage, V_{th} : threshold voltage, V_{dd} : supply voltage

Both $1/K_{delay}$ and K_{leak} are functions of PGS width and supply voltage, as shown in EQ6 and EQ7. $1/K_{delay}$ can quickly approach 1 by increasing the width of PGSs at high supply voltages, while it slowly increases at low supply voltages. On the other hand, K_{leak} is a linear function of the width of PGSs at a wide range of supply voltages. Figures 6(a) and 6(b) compare the derived equations against SPICE simulations, demonstrating acceptable accuracy. Figure 6(c) shows the inter-relationship between K_{leak} and $1/K_{delay}$ as the width of PGSs is swept. The ideal cutoff structure point is at the point where K_{leak} =0 and

$1/K_{delay}$ =1. This figure also provides a means to quantitatively compare the efficacy of different PGSs for ultra-low V_{dd} regimes, as discussed further in Section 6.

As shown in EQ8, total energy consumption can be derived from EQ5, 6, and 7. The change of E_{switch} from PGS is included here for higher accuracy. EQ8 shows that the total energy is a function of V_{dd}, K_{leak}, K_{delay} and technology parameters. T_{min} is the circuit delay without PGSs evaluated at its own V_{min}, and is thus constant.

$$
\begin{aligned}
E_{Total} &= E_{switch} + E_{leak} + E_{sleep} \\
&= \frac{1}{2} \cdot n \cdot C V_{dd} V_{swing} \cdot \alpha \\
&\quad + \frac{1}{2} n C {V_{dd}}^2 \eta \cdot n e^{\frac{-V_{dd}}{m V_T}} \cdot \left(\frac{V_{swing}}{V_{dd}} e^{\frac{V_{dd} - V_{swing}}{m V_T}} \right) \qquad \text{[EQ8]} \\
&\quad + (K_{duty} T_{\min} - e^{\frac{V_{dd} - V_{swing}}{m V_T}} \cdot t_{delay}) \cdot k \cdot W_{PGS} \cdot I_{leak\_w/o\_PGS} V_{dd} \\
E_{Total} &\approx K_1 V_{dd} V_{swing} + K_2 {V_{dd}}^2 e^{\frac{-V_{dd}}{m V_T}} \left(\frac{V_{swing}}{V_{dd}} e^{\frac{V_{dd} - V_{swing}}{m V_T}} \right) + K_3 K_{duty} W_{PGS} V_{dd}
\end{aligned}
$$

where n: length of inverter chain, C: capacitance per inverter, α: activity factor,
V_{swing}: voltage swing without voltage drop in PGSs, η: fitting coefficienct, V_{dd}: supply voltage
T_{min}: delay of inverter chain without PGSs at original V_{min}, m: subthreshold slope factor,
t_{delay}: delay of inverter chain without PGSs, k: fitting coefficient from EQ7, V_T: thermal voltage

Since K_{leak} and K_{delay} are functions of supply voltage and PGS width, we investigate energy consumption by sweeping both of these parameters. Sleep energy consumption is roughly proportional to both supply voltage and PGS width. Here, the effect of t_{delay} on sleep energy consumption is ignored since for large K_{duty} the t_{delay} term in E_{sleep} is much smaller than $K_{duty} \times T_{min}$ while for small K_{duty} the sleep energy consumption itself becomes small and less important in E_{total}. Additionally, subthreshold leakage current, the dominant source of sleep energy consumption, is nearly constant with supply voltage in the ultra-low Vdd regime while it often increases in super-threshold regimes due to short-channel effects. Therefore, we use a lumped coefficient, K_3, for simplicity in EQ8.

On the other hand, active energy consumption has a complex relationship with supply voltage and PGS width. First, PGS width affects the performance of circuits. For example, small PGSs (i.e., larger

$1/K_{delay}$) induce longer delay, resulting in higher E_{leak} consumption in circuits. In near-threshold regimes (V_{dd}>450mV for this technology), the increase in E_{leak} is relatively small, while it can significantly increase total energy consumption in sub-threshold regimes due to the importance of E_{leak}, as shown in Figure 7.

The effect of supply voltage on active energy consumption is similar to the traditional analysis [3]. Lowing V_{dd} causes performance degradation and thus leads to higher E_{leak} consumption (i.e. higher $1/K_{delay}$), while it quadratically reduces E_{switch}.

One interesting observation is that large values of $1/K_{delay}$ or E_{leak} can be alleviated by either using larger PGSs or raising supply voltages. However there is a difference between these approaches. Using larger PGSs reduces the voltage drop across PGSs, leading to lower active energy consumption compared to raising supply voltage. However, raising supply voltage is more effective in improving performance with a smaller increase in sleep energy consumption. To confirm these trends, we perform SPICE simulations where circuits initially have excessive E_{leak} consumption that must be alleviated using either of the discussed methods. Figure 8 shows that both raising V_{dd} and widening PGS can reduce active energy consumption but with differing impacts on sleep energy consumption. The larger PGS increases sleep energy consumption by 30× while raising supply voltage incurs only a 25% penalty. Given the advantage of widening PGS is improved active energy consumption compared to raising Vdd, this approach should be used in cases of small K_{duty}, where active energy is more important than E_{sleep}, which will be confirmed in Section 4.1.

4. Strategy of Using Power Gating Switches in Ultra-Low V_{dd} Regimes

4.1 PGS Design Strategies in Ultra-Low V_{dd} Regimes

This section presents a strategy for using PGSs in ultra-low V_{dd} regimes based on the findings in Section 3. We first review the conventional methods of designing PGSs. Then, we propose our PGS design method employing co-optimization in ultra-low V_{dd} regimes. In this method, supply voltage and PGS width are simultaneously chosen to achieve full energy savings at a given duty cycle.

For the designs targeted at nominal V_{dd} operations the performance degradation is often constrained by less than 5-10%. Therefore, the width of PGSs needs to be large enough to supply proper

current and minimize virtual ground bounces. Often, the constraints lead to large PGS width, often ~10% of total NFET width of main circuits [29][30].

Also, high V_{th} devices have been a popular choice for PGSs since they have similar on-current but much smaller off-current than regular V_{th} devices. Figure 9 shows that in this technology, high V_{th} devices have 600× smaller off-current, while they have only 1.7× smaller on-current at V_{dd} =1.2V. Therefore, high V_{th} PGSs can provide ~352× reduction in off-current at the same on-current.

In ultra-low V_{dd} regimes, PGS design can be different. High V_{th} devices become less attractive since they have the similar on-current to off-current ratio as regular V_{th} devices in ultra-low V_{dd} regimes. Here on-current is defined as saturation current since the V_{ds} required for device saturation is only 3-4V_T in ultra-low V_{dd} regime. Using high V_{th} PGSs is beneficial only for the case where circuits draw a current smaller than what a minimum-sized regular V_{th} PGS can deliver. Figure 10 shows that circuits with current of less than ~30nA can exploit high V_{th} PGSs for the targeted technology. For the higher current draw, regular V_{th} devices are preferred due to an unnecessary use of area by high V_{th} PGSs. The crossover point between regular V_{th} and high V_{th} PGS is technology-dependent, thus requiring careful evaluations for each technology.

In this sense, devices with a large on-current to off-current ratio are preferred for PGSs in ultra-low V_{dd} regimes. One way of improving the ratio is to use longer channel devices [16], as shown in Figure 10. Note that in this particular technology, high V_{th} devices exhibit a slightly better on-current to off-current ratio than regular V_{th} devices. However, since the ratio is technology-dependent, a careful evaluation is needed for each technology.

Another important factor to consider is that the conventional practices of sizing PGSs for maintaining performance is no longer valid since minimizing total energy consumption is a more important goal for ultra-low power applications. Therefore, PGSs should be optimized for minimizing total energy consumption. Since both PGS width and supply voltage affect total energy consumption, as we discuss in Section 3, we propose an optimization method, called co-optimization, for designing PGSs. In this

proposed method, PGS width and supply voltage are simultaneously selected for minimizing total energy consumption.

We investigate total energy consumption at different duty cycles by sweeping all combinations of PGS widths and supply voltages in the SPICE simulations using inverter chains. If K_{duty} is equal to one, then the optimal energy consumption can be achieved by supplying the conventional V_{min} without PGSs. This is because PGSs induce extra delay and more E_{leak} consumption. Since there is no sleep time, i.e. K_{duty} =1, the sleep leakage reduction is of no use in this case. The results are shown at the left end of Figure 11.

When K_{duty} falls roughly between 1 and 100, the optimal V_{dd} is similar to the conventional V_{min} and the optimal PGS width becomes large. These relatively small values of K_{duty} imply that E_{sleep} is small. Therefore, the increase in E_{sleep} caused by the use of larger PGSs is a negligible part of total energy consumption. This is well matched to the idea expressed in Section 3, that increasing PGS width is more energy-efficient than raising V_{dd} when sleep time is small. This is well supported by SPICE simulations using inverter chains, as shown in Figure 11. If the large PGS causes too much area overhead, it can be omitted with a relatively small sleep energy penalty.

When K_{duty} > 100, small PGSs and V_{dd}>V_{min} are preferred for minimizing total energy consumption since raising V_{dd} imposes a lower penalty on E_{sleep}, as discussed in Section 3. This is confirmed by SPICE simulations using inverter chains, as shown in Figure 11. The small PGSs force the effective voltage between virtual rails to approach conventional V_{min}.

Typical sensor-type applications have K_{duty} of ~10^4 [8]. Therefore, to achieve optimal energy consumption, the regular V_{th} PGS can be downsized to 0.01% of total NFET width of main circuits, as shown in Figure 11. However, since 0.01% of total NFET width is smaller than the minimum width of device in this technology, a high V_{th} PGS is instead used. For the same on-current, the high V_{th} PGS should be sized at 1% of total NFET width of the main circuits.

As stated earlier the logic depth and switching activity of the test circuits incur worst-case voltage drop across PGSs. Since higher logic depth or less activity reduces the current delivery requirement, optimized PGSs can be made even smaller in many practical settings. As a reference data point, the co-

optimization for an inverter chain with 2× logic depth and ½ the activity factor relative to the baseline system of this work suggests the use of a 50% smaller PGS at 25mV higher V_{dd} for Kduty=10^4 to achieve optimal energy consumption. The smaller voltage glitch on the virtual ground rail allows further scaling of PGS size while the lower circuit switching activity increases V_{min} due to a larger leakage to dynamic energy ratio. Baseline PGS size and V_{dd} settings incur a 6.7% energy penalty in this longer and lower activity circuit rather than its optimal values.

4.2 Comparisons of the Optimization Methods

We run SPICE simulations using inverter chains to compare our proposed co-optimization with two baseline approaches for designing PGSs. The first baseline approach is to use no cutoff structure and optimize supply voltage only. The second baseline approach, referred to as fixed-V_{min}-optimization, uses PGSs at a conventional fixed V_{min}. Figure 12 (a) shows the change of V_{min} for each strategy. It illustrates that the co-optimization calls for a higher V_{dd} than the conventional V_{min} for large values of K_{duty}. However, the V_{min} is scaled down to the functional limit of supply voltage that allows the task to be completed in a given time (K_{duty}) for the no-cutoff approach.

On the other hand, Figure 12(b) illustrates the optimal PGS width for each optimization approach. The co-optimization suggests the use of extremely small PGSs for energy optimization. However, the fixed -V_{min}-optimization cannot suggest such small PGSs since they degrade performance and thus consume extra active energy at the fixed V_{min}.

Finally, the total energy consumption of these strategies is compared in Figure 12 (c). Even at relatively small values of K_{duty}, the no-cutoff strategy consumes a significantly large amount of energy. The fixed-V_{min} optimization and co-optimization exhibit comparable energy consumption for small values of K_{duty}. However, co-optimization saves a considerable amount of total energy consumption when K_{duty}>1000. Note that sensor applications often have K_{duty} larger than 1000. For these high K_{duty} applications, the co-optimization can save up to ~99% of total energy consumption, compared to the other approaches.

4.3 Case Study Using a Fabricated Microprocessor

We apply the proposed design method to a microprocessor designed for ultra-low power applications [26]. It is fabricated in 0.18µm CMOS and consists of ~4000 gates. The total NFET width is ~6000µm. The microprocessor has tunable PGSs with widths ranging from 0.66µm to 28µm for mitigating the effects of process variations on PGSs. Using the smallest PGS, E_{active} is measured as 2.35pJ/cycle with I_{sleep}=2pA. The processor operates at 60 kHz with V_{dd}=0.475V. For the smallest and the largest PGSs, we measure active energy consumption and sleep energy consumption. We estimate that 1000 instructions are executed during active mode. We then calculate the total energy consumption at several values of K_{duty}.

Figure 13 shows that as sleep time become small (i.e., larger K_{duty}) the ideal strategy transitions from using the widest (28µm) PGS to employing the 0.66µm PGS. The large PGS is slightly more energy efficient at high duty cycles due to less performance degradation and smaller voltage drop across the PGS. However, the small PGS becomes energy-optimal at low duty cycles since sleep energy consumption represents a large portion of total energy consumption. These strategies cross over when $T_{deadline}$ is 4 seconds. Since $T_{deadline}$ for most ultra-low power systems is larger than 4 second [8], small PGSs are energy-optimal for these applications. If 1000 seconds (16 minutes) sleep time is assumed, the small PGS provides 4.6× lower total microprocessor energy consumption compared to the large PGS. We cannot measure T_{min} of the microcontroller (i.e., the microprocessor delay at V_{min} without PGSs), therefore we approximate it as the delay at V_{min} with the large PGS. With the estimated T_{min}, the K_{duty} for 10 seconds is ~10^6.

5. Feasibility of Minimal-Sized PGSs in Ultra-Low Vdd Regimes

Even if performance degradation is ignored, designers are unlikely to view extremely small PGSs as viable options since the voltage drop across PGSs may cause functional robustness problems. In super-threshold regimes, it is true that the small PGSs cause functional failures. Figure 14 shows that the microprocessor discussed in Section 4 is not functional with the small PGSs at V_{dd} > 0.8V. However, in ultra-low V_{dd} regimes, the microprocessor with the small PGSs is functional. Therefore, it is important to understand the different feasibilities of small PGSs in ultra-low V_{dd} regimes.

One reason that the small PGS functions well in ultra-low V_{dd} regimes can be found in the relationship of V_{ds} and subthreshold current. As shown in EQ1, subthreshold current becomes insensitive to V_{ds} once V_{ds} is larger than 3~4 V_T. In other words, even if the microprocessor attempts to draw a large current, for example, because of many simultaneous internal node switches, the V_{ds} or voltage drop across the PGS changes only by a small amount. Instead, the current draw is limited and the microprocessor is slowed. However, linear and saturated current of devices in super-threshold regimes have a linear relationship with V_{ds}. Therefore, the V_{ds} of the PGS quickly rises to the point at which the PGS can supply a large current. This V_{ds} increase appears as a large virtual ground bounce, making the minimal PGS less robust in super-threshold regimes.

To confirm these concepts, we perform SPICE simulations with two different sets of inverter chains. The first set has one inverter chain that is switching and four chains that are not switching. The second set has five inverter chains that are switching. Each inverter chain is identical, thus the second set draws ~5× higher current draw. We investigate voltage drops across PGSs for these circuits PGSs are sized at 0.05% of total NFET width for each set.

Figure 15 illustrates that relative virtual ground levels are smaller for ultra-low V_{dd} regimes for both low and high work load cases, which is expected, given the different relationships of V_{ds} with drain current in two different V_{dd} regimes. Additionally, in ultra-low V_{dd} regimes, the relative increase of the virtual ground level from low to high work load is smaller. The final observation is that the relative virtual ground level goes up at $V_{dd} < 0.4V$. This is because the V_{ds} of the PGS gets close to $3\sim4V_T$ and then decreases only slightly.

The 0.13 and 0.18μm technologies considered in this paper exhibit less process variations than leading-edge scaled technologies. In such cases robustness can be improved by using a wider PGS at the cost of sleep energy consumption [37]. To further mitigate process variations, trimmable PGSs such as those in [26] can be used for selecting appropriate width PGSs post-silicon to minimize sleep energy. Since robust operation is of critical importance, statistical simulations across PVT (process, voltage, and temperature) variations should be considered.

6. Beyond Basic PGSs

So far, we have discussed only the basic PGS topology. However there are many variations for PGSs to improve the fundamental tradeoff between performance degradation and sleep energy reduction. In this section, we quantitatively compare different flavors of PGSs and provide guidelines for choosing energy-optimal PGSs in ultra-low V_{dd} regimes.

Figure 16 shows three well-known PGS topologies: basic PGS, DTCMOS PGS, and stack-forcing PGS. In DTCMOS PGSs, the gate and the body of the PGSs are tied to increase on-current. Therefore, DTCMOS PGSs are expected to have a smaller $1/K_{delay}$, compared to the basic PGS. The stack-forcing PGS uses two FETs in series to reduce off-current using the stack effect [34]. These series-connected FETs induce negative V_{gs} at the upper FET, which exponentially decreases off-current. Therefore, it exhibits smaller K_{leak} than the basic PGS. However, $1/K_{delay}$ can be worse. At V_{dd}=0.5V, the K_{leak}-K_{delay} curves of these structures are shown in Figure 17. For the same K_{leak}, the DTCMOS structure provides the smallest $1/K_{delay}$, and thus the smallest E_{leak}, followed by stack-forcing PGS.

Super-cutoff PGS [32] is not considered in comparisons since the penalty of generating bias voltages is difficult to quantify. However, it can be a promising design choice due to the exponential relationship between subthreshold current and supply voltage in ultra-low V_{dd} regimes. In [35], a detailed analysis on the tradeoff between generating bias voltages and sleep energy reduction is presented for ultra-low V_{dd} operations.

7. Conclusions

This paper investigates the interaction of optimal energy, supply voltage and PGS for ultra-low V_{dd} designs. The results show that ignoring sleep leakage energy in ultra-low V_{dd} regimes can significantly degrade energy efficiency. Therefore, we propose several approaches for designing PGSs including co-optimization, which seeks to achieve optimal energy by simultaneously adjusting both PGS size and V_{dd}. Unlike typical practices in higher Vdd regimes, in which large PGSs and nominal supply voltage are often chosen, our proposed optimization suggests using minimal PGS and higher V_{dd} for those applications with long sleep time. This reduces energy by 125× in SPICE simulations. The effectiveness of the proposed

method is confirmed by the silicon measurements from an ultra-low power microprocessor. Finally, the feasibility of using minimal-sized PGSs in ultra-low Vdd regimes is studied with the focus of functional robustness using SPICE simulations and silicon measurements.

References

[1] Intel XScale. http://www.intel.com/design/intelxscale

[2] IBM Power PC. http://www.chips.ibm.com/products/powerpc

[3] B. Zhai, D. Blaauw, D. Sylvester, K. Flautner, "Theoretical and Practical Limits of Dynamic Voltage Scaling," Design Automation Conference, pp. 868-873, June, 2004

[4] A. Wang, A. Chandrakasan, "A 180-mV subthreshold FFT processor using a minimum energy design methodology," IEEE Journal of Solid-State Circuits, vol.40, no.1, pp.310-319, Jan 2005

[5] G. Chen, et al, "Millimeter-Scale Nearly Perpetual Sensor System with Stacked Battery and Solar Cells," ISSCC., pp.288-289, 2010

[6] J. Kwong, et al, "A 65nm Sub-Vt Microcontroller with Integrated SRAM and Switched-Capacitor DC-DC Converter," ISSCC., 2008

[7] B. Zhai, L. Nazhadili, J. Olson, A Reeves, M Minuth, R Helfand, S. Pant, D. Blaauw, T. Austin, "A 2.60pJ/Inst Subthreshold Sensor Processor for Optimal Energy Efficiency," Symposium on VLSI Circuits, pp.154-155,2006

[8] L. Nazhadili, M. Minuth, T. Austin, "SenseBench: toward an accurate evaluation of sensor network processors," Workload Characterization Symposium, pp.197-203, Oct, 2005

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

[10] S. Hanson, M. Seok, D. Blaauw, D. Sylvester, "Nanometer Device Scaling in Subthreshold Circuits," on Design Automation Conference (DAC), June 2007

[11] M. Seok, D. Sylvester, D. Blaauw, "Optimal Technology Selection for Minimizing Energy and Variability in Low Voltage Applications," on International Symposium on Low Power Electronics and Design, August, 2008

[12] D. Bol, et, al, “Technology Flavor Selection and Adaptive Techniques for Timing-Constrained 45nm Subthreshold Circuits,” International Symposium on Low Power Electronics and Design, pp.21-26, Aug, 2009

[13] B. Calhoun, A. Wang, A. Chandrakasan, “Device sizing for minimum energy operation in subthreshold circuits”, Custom Integrated Circuits Conference, pp. 95~98, Sep, 2004

[14] J. Kwong, A. Chandrakasan, “Variation-Driven device sizing for minimum energy sub-threshold circuits”, International Symposium on Low Power Electronics and Design, pp8~13, 2006

[15] J. Keane, H. Eom, T.-H. Kim, S. Sapatnekar, C. Kim, “Subthreshold logical effort: A systematic framework for optimal subthreshold device sizing”, DAC, pp425~428, 2006

[16] T. Kim, H. Eom, J. Keane, and C. Kim, “Utilizing Reverse Short Channel Effect for Optimal Subthreshold Circuit Design,” International Symposium on Low Power Electronics and Design, Oct, 2006

[17] B. Zhai, D. Blaauw, D. Sylvester, S. Hanson, “A sub-200mV 6T SRAM in 130nm CMOS”, International Solid-State Circuits Conference, 2007

[18] S. R. Sridhara, et al., “Microwatt Embedded Processor Platform for Medical System-on-Chip Applications,” Symposium on VLSI circuits, 2010

[19] N. Verma, A. Chandrakasan. “A 65nm 8t sub-Vt SRAM employing sense-amplifier redundancy,” International Solid-State Circuits Conference, pages 328–606, Feb. 2007.

[20] L. Chang, et al., “A 5.3GHz 8T-SRAM with operation down to 0.41V in 65nm CMOS,” IEEE Symposium on VLSI Circuits, 2007

[21] I. Chang, J.-J. Kim, S.P. Park, K. Roy. “A 32kb 10t subthreshold SRAM array with bit-interleaving and differential read scheme in 90nm CMOS,” International Solid-State Circuits Conference, pages 388–622, Feb. 2008.

[22] J.P. Kulkarni, et, al, “A 160mV Robust Schmitt Trigger Based Subthreshold SRAM,” Solid-State Circuits, IEEE Journal of, vol 42, issue 10, pp 2303-2313, Oct, 2007

[23] H. Kaul, et al, "A 320mV 56uW 411GOPS/Watt Ultra-Low-Voltage Motion-Estimation Accelerator in 65nm CMOS," International Solid-State Circuits Conference, 2008

[24] Y. Pu, et al., "An Ultra-Low-Energy/Frame Multi-Standard JPEG Co-Processor in 65nm CMOS with Sub/Near-Threshold Power Supply," IEEE International Solid-State Circuits Conference, pp.146-147, 2009

[25] S. Hanson, B. Zhai, M. Seok, B. Cline, K. Zhou, M. Singhal, M. Minuth, J. Olson, L. Nazhandali, T. Austin, D. Sylvester, D. Blaauw, "Performance and variability optimization strategies in a sub-200mV, 3.5pJ/Inst, 11nW Subthreshold Processor," Symposium on VLSI circuits, pp 152-153, 2007

[26] M. Seok, S. Hanson, Y.-S. Lin, Z. Foo, D. Kim, Y. Lee, N. Liu, D. Sylvester, D. Blaauw, "The Phoenix Processor: A 30pW Platform for Sensor Applications," on Symposium on VLSI Circuits, June 2008

[27] ARM Cortex M0. http://www.arm.com/products/CPUs/ARM-Cortex-M0.html

[28] M. Seok, S. Hanson, D. Sylvester, D. Blaauw, "Analysis and Optimization of Sleep modes in Subthreshold Circuit Design," Design Automation Conference, June 2007

[29] S. Mutoh, T. Douseki, Y. Matsuya, T. Aoki, S. Shigematsu, J. Yamada, "1V power supply high-speed digital circuit technology with multithreshold voltage CMOS," IEEE Journal of Solid-State Circuits, vol.30, NO.8, August, 1995

[30] J. Kao, A. Chandrakasan, "Dual-Threshold voltage techniques for low power digital circuits," IEEE Journal of Solid-State Circuits, vol.35, NO.7, July, 2000

[31] K. Kumagai, et, al, "A Novel Powering-down Scheme for Low Vt CMOS Circuits," Symposium on VLSI Circuits Digest of Technical Papers, pp 44-45, 1998

[32] H. Kawaguchi, et, al, "A CMOS Scheme for 0.5V Supply Voltage with Pico-Ampere Standby Current," Solid-State Circuits Conference, 1998. ISSCC 1998. Digest of Technical Papers. IEEE International, pages 192–193, Feb. 1998.

[33] B. Razavi, “Design of Analog CMOS Integrated Circuits,” McGraw-Hill, 2001

[34] S. Narendra, et, al, “Scaling of stack effect and its application for leakage reduction,” International Symposium on Low Power Electronics and Design, pp 195-200, 2001

[35] Y. Lee, M. Seok, S. Hanson, D. Blaauw, D. Sylvester ”Standby Power Reduction Techniques for Ultra-Low Power Processors,” European Solid-State Circuits Conference, September, 2008

[36] L. Nazhandali, et al., “Energy Optimization of Subthreshold-Voltage Sensor Network Processors,” International Symposium on Computer Architecture, June, 2005

[37] D. Bol, et al., “Robustness-Aware Sleep Transistor Engineering for Power-Gated Nanometer Subthreshold Circuits,” International Symposium on Circuits and Systems, pp.1484-1487, 2010

[38] D. Bol, et al., “Analysis and Minimization of Practical Energy in 45nm Subthreshold Logic Circuits,” International Conference on Computer Design, pp.294-300, Oct, 2008

FIGURES

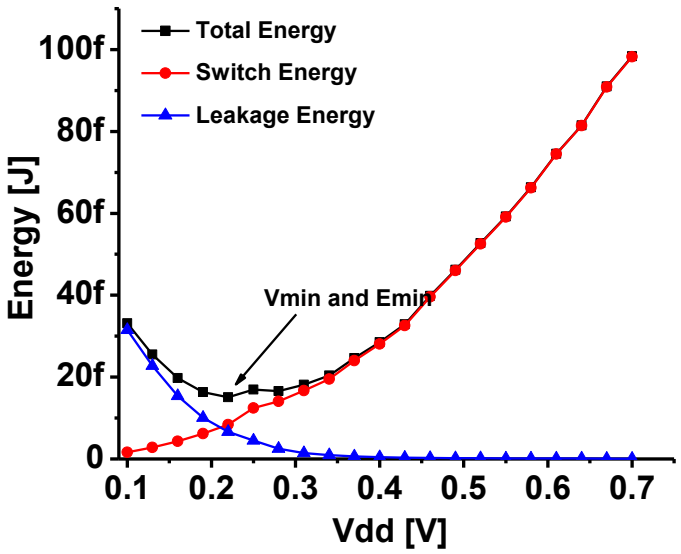
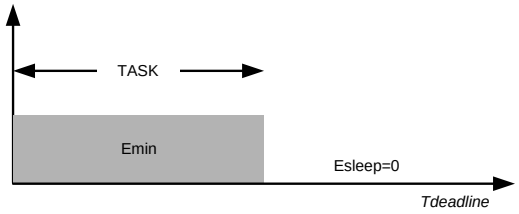
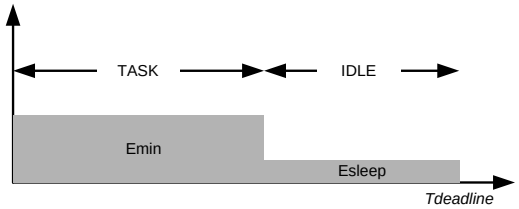


Figure 1 V_{min}/E_{min} curve with no consideration on sleep energy



(a) Task is completed before Tdeadline at Vmin, assuming only Emin consumed



(b) Task is completed before Tdeadline at Vmin, consuming Emin + Esleep

Figure 2 Illustration of task scheduling at different deadlines

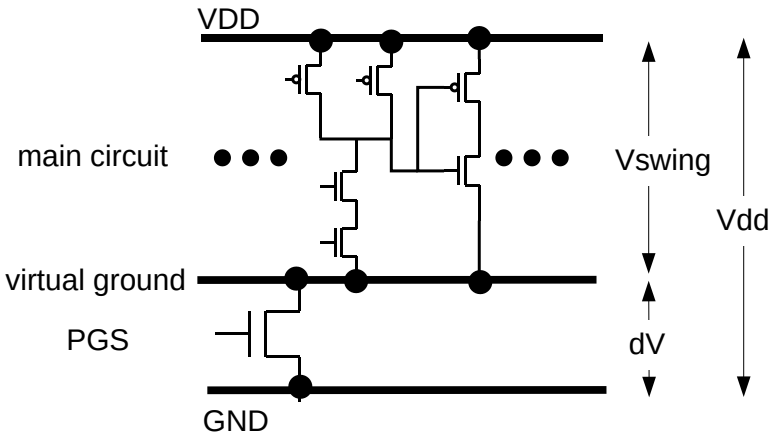


Figure 3 Basic PGS configuration

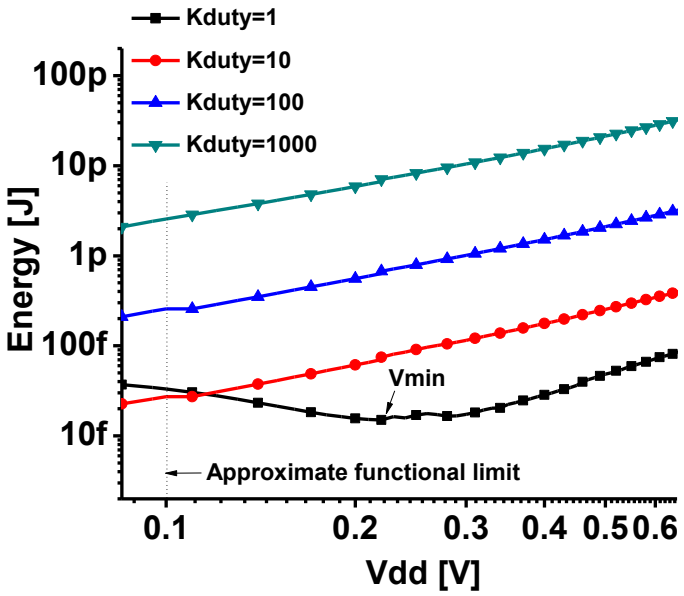
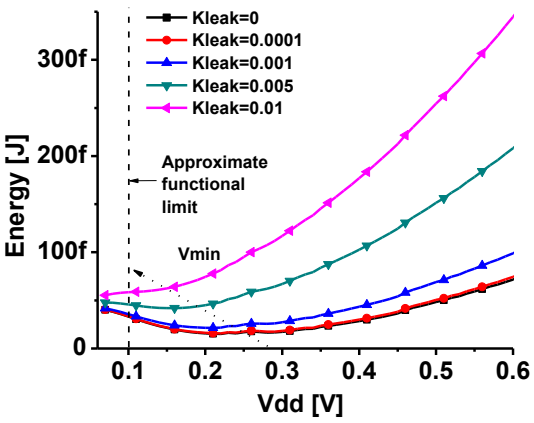
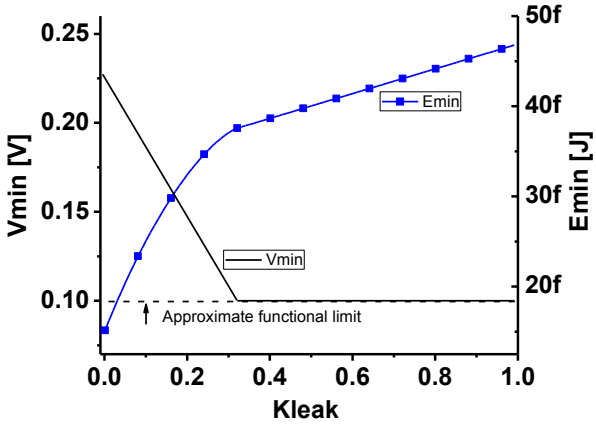


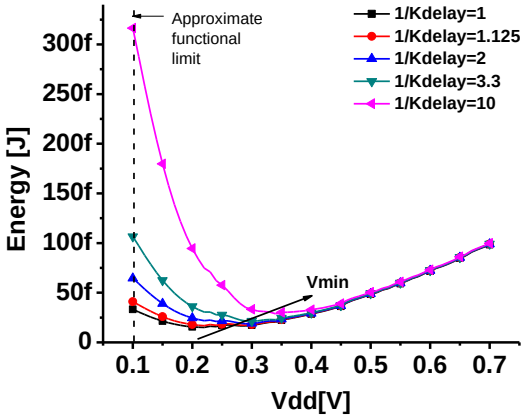
Figure 4 V_{min} /E_{min} curves with different K_{duty} considering sleep energy



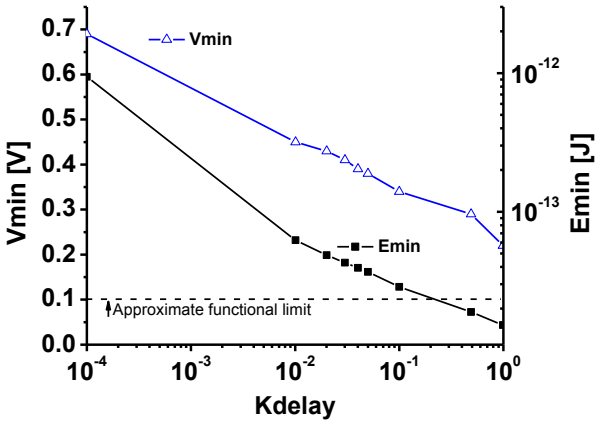
(a) V_{min} /E_{min} curves



(b) K_{leak} - V_{min} /E_{min}

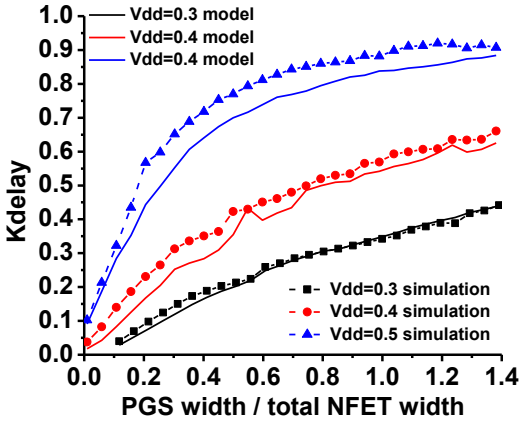


(c) V_{min} /E_{min} curves

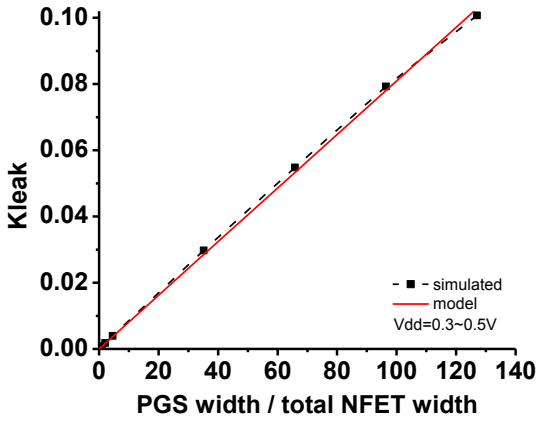


(d) K_{delay} - V_{min} /E_{min}

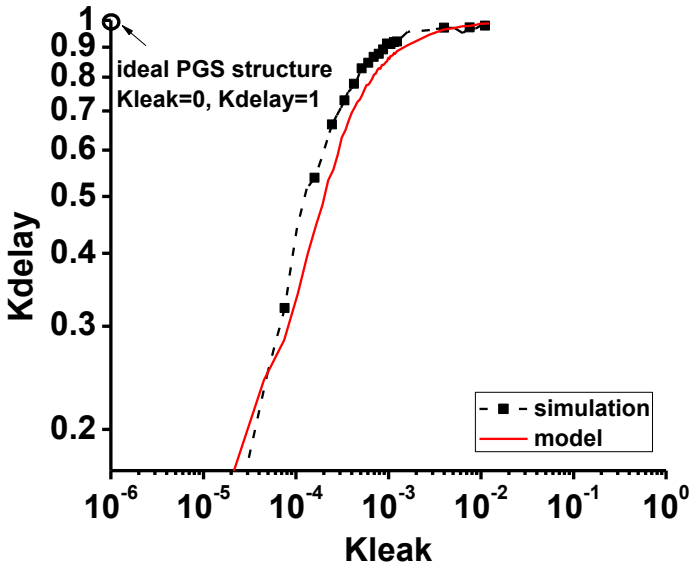
Figure 5 V_{min} /E_{min} change with K_{leak} and K_{delay}



(a) width - K_{delay}



(b) width - K_{leak}



(c) K_{leak} – K_{delay}

Figure 6 K_{leak} and K_{delay} change with PGS width and V_{dd}

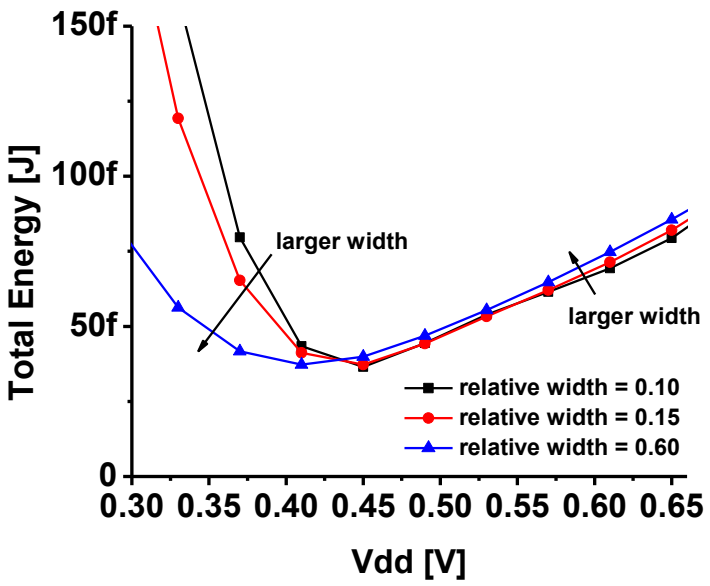


Figure 7 V_{min} /E_{min} with different PGS sizes, K_{duty}=100

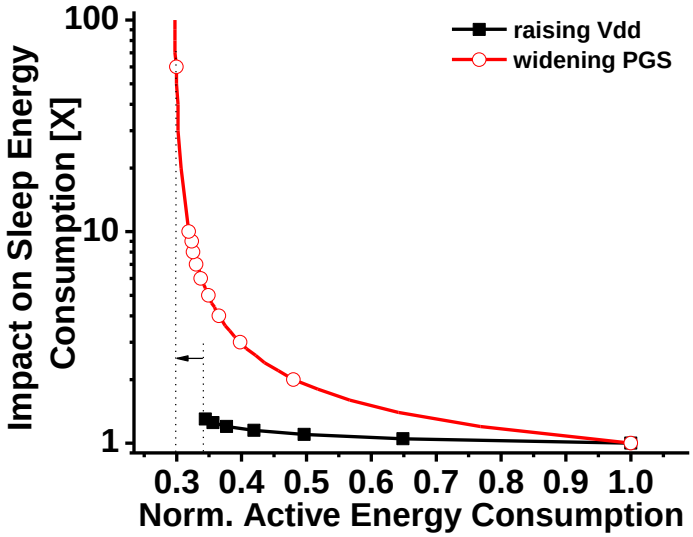


Figure 8 Comparison between raising V_{dd} and upsizing PGS in energy optimization

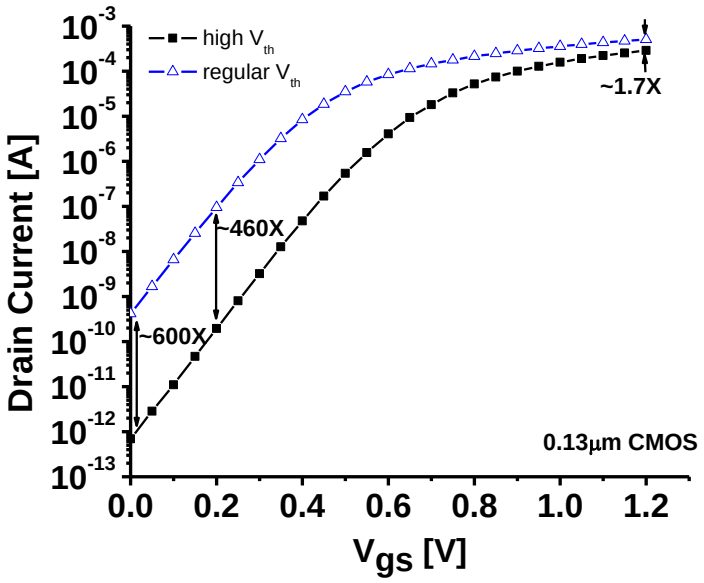


Figure 9 On/off-current of high V_{th} and regular V_{th} devices.

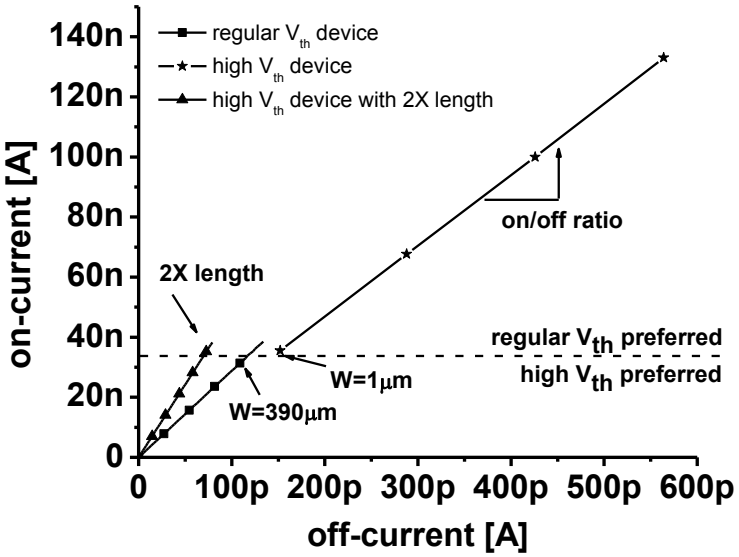


Figure 10 Off-current vs. on-current as sweeping PGS width.

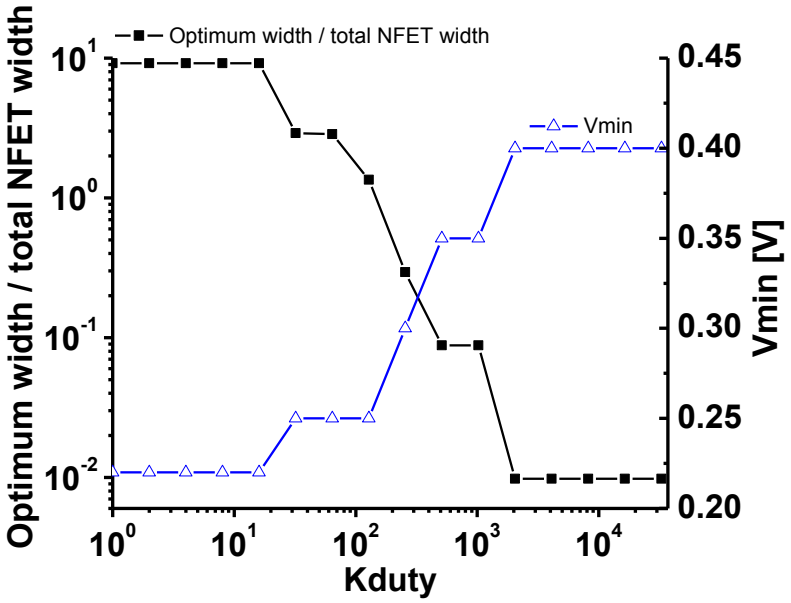
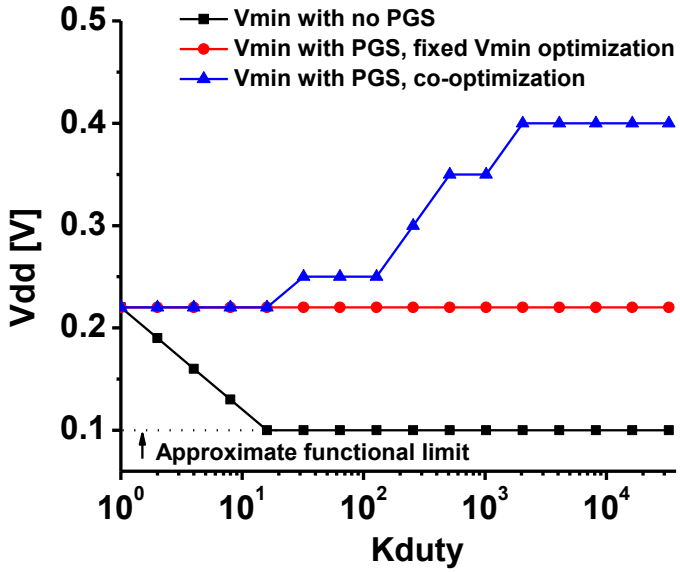
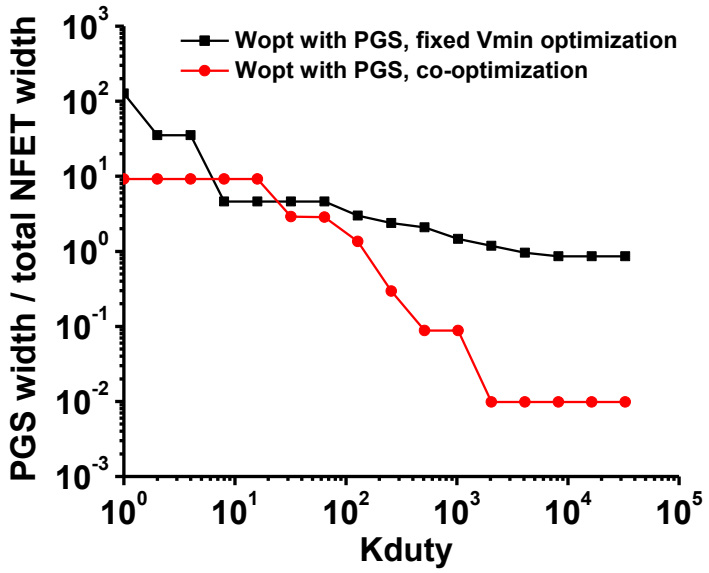


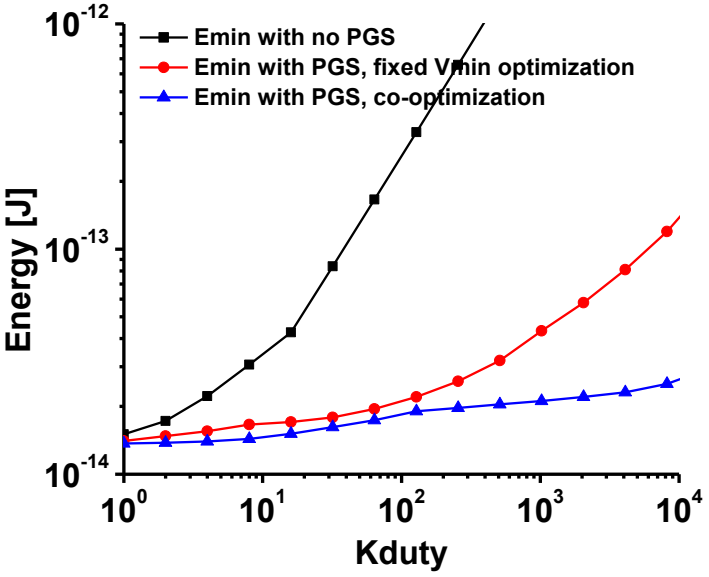
Figure 11 New V_{min} and optimal PGS size at different K_{duty}.



(a) K_{duty} – V_{min} over three strategies



(b) K_{duty} – optimal PGS width over three strategies



(c) K_{duty} – E_{min} over three strategies

Figure 12 Comparison of three optimization strategies.

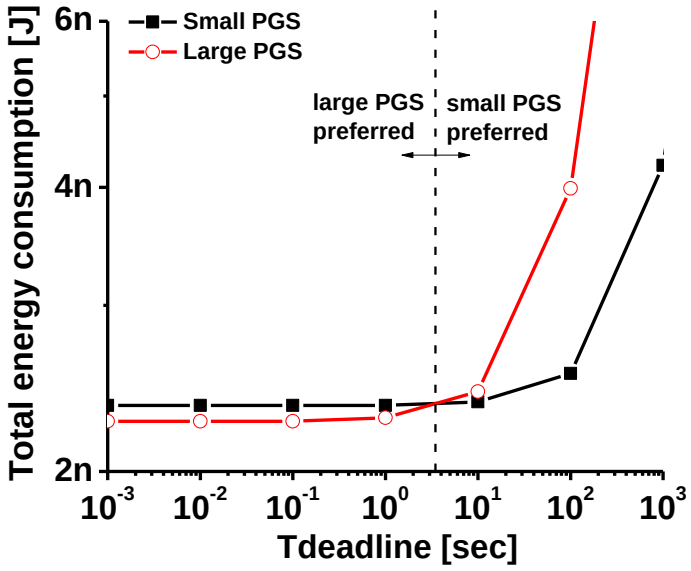


Figure 13 Measured total energy consumption with two different PGS sizes from a test microprocessor.

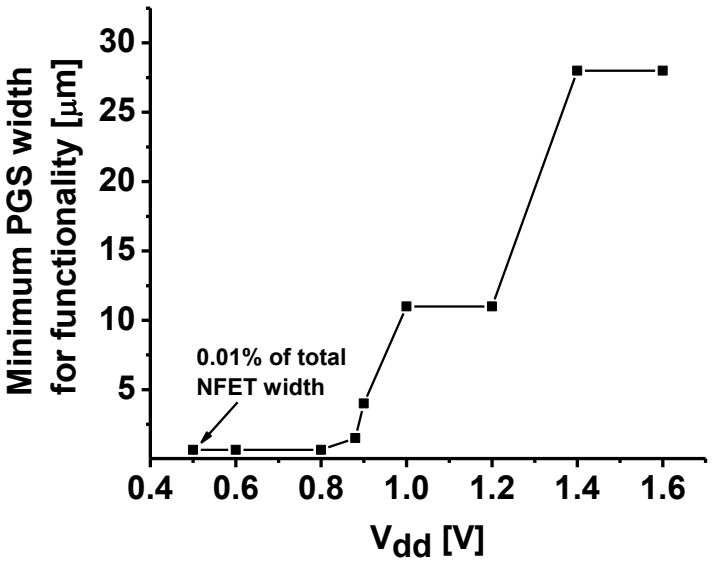


Figure 14 Measured minimal PGS size for functionality.

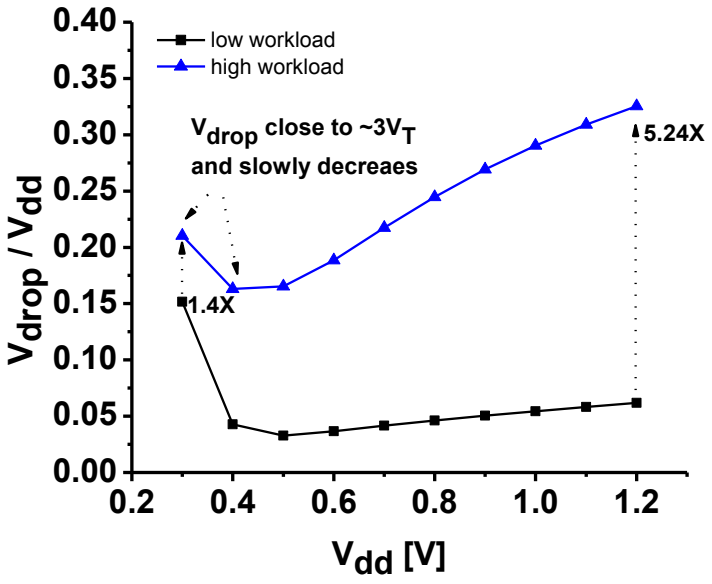


Figure 15 Simulated virtual ground level over different workload and supply voltage.

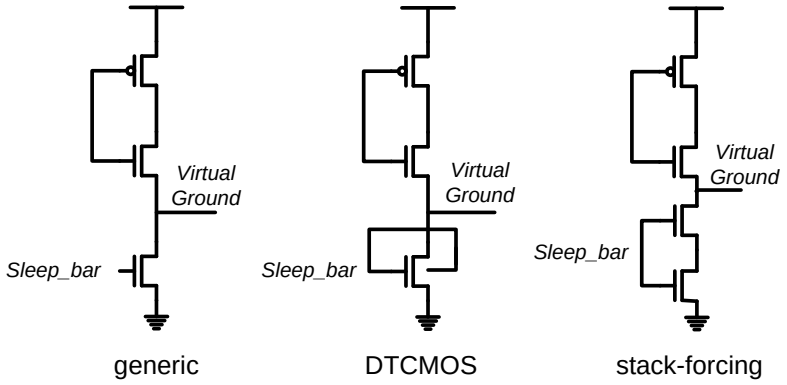


Figure 16 Generic, DTCMOS, and stack-forcing PGS.

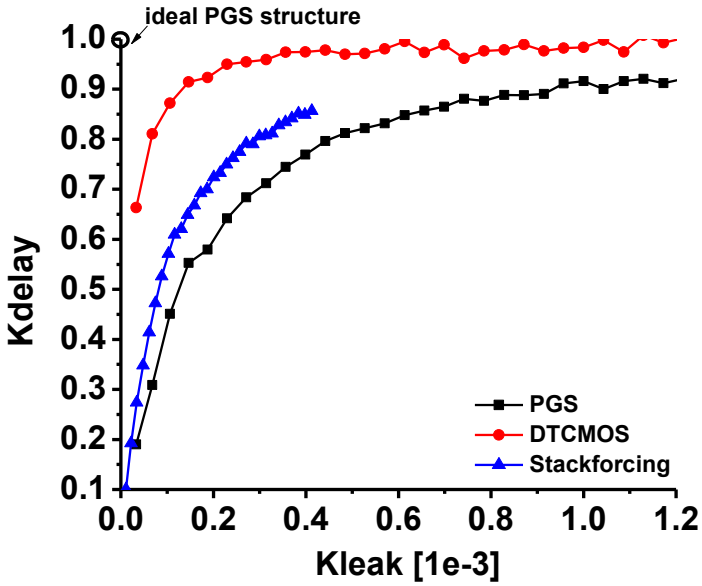


Figure 17 K_{leak} - K_{delay} curves with different PGSs.

RESPONSE TO REVIEWER COMMENTS

We would first like to thank the reviewers for taking the time to make constructive comments on the manuscript. We have revised the paper to consider all reviewer comments. Please see below for specific responses to individual comments.

Thank you,

Mingoo Seok, Scott Hanson, David Blaauw, and Dennis Sylvester

Reviewer's Comments

Reviewer: 1

SPECIFIC FEEDBACK TO AUTHORS

The authors did a good job in addressing the comments.

-Thank you.

Reviewer: 2

SPECIFIC FEEDBACK TO AUTHORS

The authors have adequately addressed my comments in this revision

-Thank you.

Reviewer: 3

SPECIFIC FEEDBACK TO AUTHORS

I thank the authors for the clarifications and response to my earlier comments. I believe it answers most of the pertinent questions. A few minor comments are:

1. The authors mention that the use of a RO represents the worst-case load condition. I do agree with that. And hence my previous comment, that this design is probably suboptimal. For most datapath circuits the activity and hence load is much smaller. The capacitance on the virtual Vss is also different depending on the logic depth and so on. Hence I wonder if a better design can be obtained by sizing the PGS for a particular logic block. Or is RO-representation of the logic block close enough to the optimal solution. A discussion to this effect would improve the quality of the paper.

- Thank you for your comment. We agree that the activity and hence current load is smaller for most datapath circuits and hence it would be instructive to know the sensitivity of optimal Power Gating Switch (PGS) design to activity. We extend our experiments of the co-optimization with lower activity circuits, i.e., 2× larger logic depth and ½ activity inverter chain and find the optimal solutions. The optimization suggests using a narrower PGS and higher Vdd, more specifically 0.005% of total nfet width for PGS width and 425mV for supply voltage. (For the baseline circuits, the optimal combination is the 0.01% of total nfet width and 400mV as shown in Figure 11). The reduced voltage drop arising from lower activity allows the use of smaller PGSs while lower activity raises the Vmin (energy optimal supply voltage), as suggested in [3]. The energy penalty incurred by using the baseline PGS size and V_{dd} value in this longer and lower activity inverter chain is 6.7%. We have added this discussion at the end of Section 4.1.

2. Please mention the conditions for the SPICE simulations in the paper. It is important to know if the routing parasitics were considered.

- We have not included routing parasitics in our SPICE simulations. We have added information on the simulation conditions at the beginning of Section 2 for clarification.