Selective Light V_{th} Hopping (SLITH): Bridging the Gap Between Runtime Dynamic and Leakage Power Reduction

Hao Xu, Ranga Vemuri and Wen-Ben Jone
Department of Electrical and Computer Engineering, University of Cincinnati
Cincinnati, Ohio 45221-0030, USA
Email: {xuho, ranga, wjone}@ececs.uc.edu

***Abstract*—Ever since the invention of various leakage power reduction techniques, leakage and dynamic power reduction techniques are categorized into two separate sets. Most of them cannot be applied together during runtime. The gap between them is due to the large energy breakeven time (EBT) and wake-up time (WUT) of conventional leakage reduction techniques. This paper proposes a new leakage reduction technique (SLITH) based on V_{th} hopping. SLITH has very low EBT and WUT, yet keeps the effectiveness of leakage reduction. Thus, it is able to reduce the gap, and enables joint dynamic and leakage power reduction. SLITH can be applied together with clock gating, pre-computation and operand isolation etc., and significantly reduces both dynamic and active leakage power consumption.**

***Index Terms*—runtime leakage power reduction, energy breakeven time, wake-up time, V_{th} hopping**

I. Introduction

MOSFET scaling into deep sub-100nm has resulted in significant increase in leakage power consumption. Particularly, in 45nm technology generation and beyond, leakage power consumption will catch up with, and may even dominate, dynamic power consumption [1]. Subthreshold leakage, gate leakage and band-to-band tunneling leakage (BTBT) are the three main components contributing to the total leakage power consumption. Many leakage reduction techniques have been introduced and studied so far. They can be characterized into two classes: runtime techniques and design-time techniques. The runtime techniques, such as reverse body biasing, input vector control and power gating [1], tune the circuit into a low-leakage state during runtime when the circuit is idle.

On the other hand, existing runtime dynamic power reduction techniques, such as clock gating, pre-computation and operand isolation etc., prevent the circuit from switching when it is considered to be idle in the next clock cycle. Therefore in essence, although runtime dynamic and leakage power reduction are two separate sets of techniques, they have a common nature: they both look for circuit idleness to enforce the circuit into a low-power mode.

Despite of their common nature, dynamic and leakage power reduction techniques can hardly be applied together during runtime. Leakage reduction techniques usually incur high energy overhead and require long time for mode transition, whereas dynamic power reduction techniques do not have this problem. To explain this, we first introduce two important definitions: energy breakeven time (EBT) and wake-up time (WUT). EBT is defined as the minimal time for a circuit to stay in low-power mode, such that energy savings catch up with the energy overhead for mode transition. WUT is defined as the minimal time required for a circuit to fully recover from low-power mode to normal mode [2]. The EBT and WUT of most dynamic power reduction techniques are within one clock cycle, while they are multiple cycles for conventional leakage reduction techniques. For example, [2] have shown that in 130nm technology, the EBT and WUT of power gating are 100 clock cycles and 2 cycles, respectively.(Clock speed is 4GHZ.) The EBT and WUT of reverse body biasing are around 100 cycles and 4 cycles, respectively.

Researchers have been studying joint dynamic and leakage reduction. Abddollahi et al. [3] proposed the scheme of applying power gating together with pre-computation on a n-bit comparator, as shown in Figure 1. When the LSB of

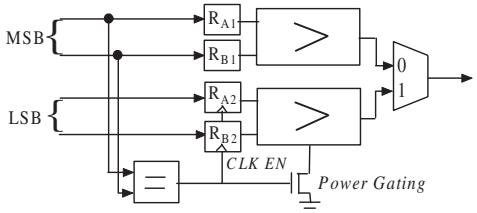

Fig. 1. Pre-computation and Power Gating

the comparator is detected to be useless in the next clock cycle, the flip-flips feeding LSB (R_{A2}, R_{B2}) will be disabled to reduce dynamic power. Meanwhile, power gating is applied to the combinational circuit of LSB to reduce leakage power. This scheme is reasonable in principle, but is not feasible in practices. The dynamic power reduction of the comparator works well. However for leakage power, assume that the EBT and WUT of power gating are E and W cycles, respectively. This scheme incurs two problems:

1) Energy overhead problem: To gain positive net energy saving, LSB should not be used in the next E+W cycles. This is because LSB requires W cycles to recover from sleep mode, and E cycles to guarantee that energy saving is larger than energy overhead.

978-3-9810801-5-5/DATE09 © 2009 EDAA

2) Wake-up delay problem: Since LSB requires W cycles to recover from sleep mode, it should be activated W cycles before it is used. This imposes an extra requirement for designers. They need to know the workload information of the comparator beforehand. Otherwise, the comparator needs to wait for W cycles to be functional again.

These two problems form the gap between dynamic and leakage power reduction. Both of they need to be addressed for a practical joint reduction scheme. In [3], Abddollahi et al. considered the energy overhead problem, and proposed a method to avoid power gating switching for short idleness (less than E cycles). However, their method requires series of D-fliplops to be inserted before the power gating controller, and thus incurs large area and power overhead. In additional, they did not address the wake-up delay problem. Min et al. [4] studied the joint application of clock gating and power gating. They proposed the scheme of "ZSCCMOS" to solve the wake-up delay problem. However, their method did not consider the energy overhead problem.

This paper emphasizes that the gap between dynamic and leakage power reduction is caused by the large EBT and WUT of conventional leakage reduction techniques. This gap can be diminished when EBT or WUT is reduced. If the sum of EBT and WUT can be reduced to one clock cycle, both energy overhead problem and wake-up delay problem will be completely solved, and the gap will be eliminated. We denote the sum of EBT and WUT as EAW. Based on this observation, we propose a new leakage reduction technique called "**S**elective **Li**ght V_{th} **H**opping" (SLITH), which reduces its EAW to as low as one clock cycle, while maintains the effectiveness of leakage reduction. In this way, SLITH addresses the above mentioned two problems and bridges the gap. As we will show in Section IV, SLITH can be applied together with most dynamic power reduction techniques. It is especially effective for active leakage reduction.

II. SELECTIVE LIGHT V_{th} HOPPING

A. Basic V_{th} Hopping Modeling

SLITH is based on one type of reverse body biasing technique, V_{th} hopping [1]. A basic V_{th} hopping diagram is illustrated in Figure 2, where V_P and $-V_N$ are the biasing voltage sources for PMOS and NMOS, respectively. They can either be provided by the off-chip voltage sources or be generated by the on-chip voltage converter. S_1 to S_4 are control transistors. When V_{th} hopping is applied, S_1 and S_3 will be turned on, while S_2 and S_4 will be turned off.

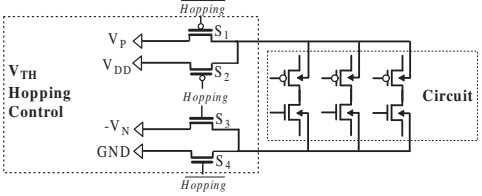

Fig. 2. V_{th} Hopping Implementation

Xu et al. [5] have derived the model for PMOS V_{th} hopping, as shown in Figure 3. In the following study, we will only show the modeling process for PMOS V_{th} hopping. NMOS V_{th} hopping can be modeled in a similar way. In Figure 3, V_b is the PMOS body voltage. R_1 and R_2 are the equivalent resistors of transistors S_1 and S_2 (in Figure 2), respectively. C_{bV} and C_{bG} are the equivalent capacitances of PMOS body-to-VDD and body-to-GND, respectively. I_{TV} and I_{TG} are the BTBT leakage. I_S is the subthreshold leakage. C_1 and C_2 (not shown in Figure 3) are the equivalent gate capacitances of transistor S_1 and S_2, respectively.

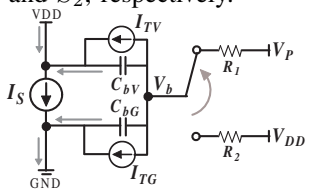

Fig. 3. Model of PMOS RBB Circuit

When V_{th} hopping is applied, V_b switches from V_{DD} to V_P. Then V_P starts to charge up the body capacitances C_{bV} and C_{bG}. The energy overhead for switching S_1 and S_2 is:

$$E_{swt} = V_P^2(C_1 + C_2) \tag{1}$$

The charging overhead to fully charge up the PMOS body is:

$$E_{charging} = (C_{bG} + C_{bV})\Delta V_P^2 + C_{bG}V_{DD}\Delta V_P \tag{2}$$

When the body is fully charged up to V_P, the total leakage (I) will be reduced to:

$$I = e^{B_s \Delta V_P}\widehat{I}_S + e^{-B_t \Delta V_P}\widehat{I}_T \tag{3}$$

where $\widehat{I}_S$ and $\widehat{I}_T$ are the original subthreshold leakage and BTBT leakage of the circuit without V_{th} hopping. B_s and B_t are the reduction exponents for subthreshold and BTBT leakage, respectively [5]. The energy saving per unit time is:

$$E_{saving/ut} = (\widehat{I}_S + \widehat{I}_T - I)V_{DD} \tag{4}$$

Thus, the EBT of V_{th} hopping can be calculated by:

$$T_{EBT} = \frac{E_{swt} + E_{charging}}{E_{saving/ut}} \tag{5}$$

According to [6], WUT can be calculated by the total charge stored in the body capacitance divided by the on-current (I_2) of the discharging transistor R_2:

$$T_{WUT} = \frac{(C_{bG} + C_{bV})\Delta V_P}{I_2} \tag{6}$$

Finally EAW can be calculated by:

$$T_{EAW} = T_{EBT} + T_{WUT} \tag{7}$$

B. Selective V_{th} Hopping

When V_{th} hopping is applied, the leakage energy saving and body charging overhead of each gate in the circuit can vary significantly. Due to this variation, some gates in the circuit, which have high saving and low overhead, are suitable for both active and standby leakage reduction. On the contrary, some gates with low saving and high overhead, are only suitable for standby leakage reduction. In order to demonstrate the variation of leakage saving versus body charging overhead, we apply V_{th} hopping on an ALU (C880 of ISCAS85). Then the EBT value of each gate in the ALU is calculated. Figure 4 shows the statistics of the EBT value of each gate. The X-axis is EBT value. The Y-axis is the count of the gates, whose EBT falls into the same range. In Figure 4, the EBT of each gate can vary up to 80 times, from 3 to 230 clock cycles. We can apply V_{th} hopping on a subset of the circuit, which

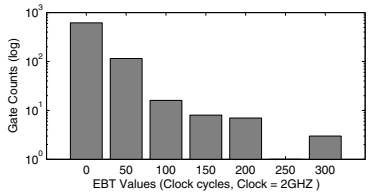

Fig. 4. Gate EBT Variation on C880 with 32nm Technology

only includes the gates with low EBT value. There are three advantages in doing so: 1). The EBT value of this subset will be much lower than the original circuit. 2) Low-EBT gates usually have small overhead, and thus small body capacitance (C_{bG}, C_{bV}). So according to Equation 6, the WUT value of this subset can be significantly reduced as well. 3) Low-EBT gates usually has high leakage. So by selecting them, we can still reduce a significant amount of leakage power. We call this method as "selective" V_{th} hopping.

Taking a clock-gated (clock=0) D-flipflop for example. We apply selective V_{th} hopping on a subset, instead of the whole DFF. The selected transistors are shown in gray color in Figure 5. Experimental results with 32nm predictive technology [7] show that in this scheme, leakage reduction ratio decreases from 97% to 60%. However, this scheme results in 2.2X reduction in EBT, and 3X reduction in WUT of the D-flipflop.

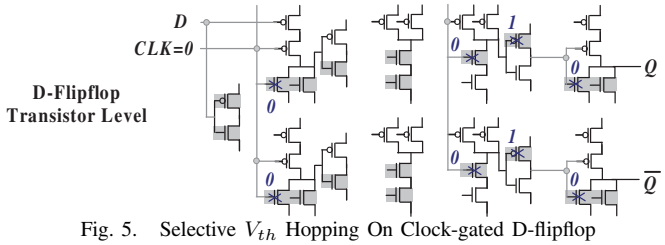

Fig. 5. Selective V_{th} Hopping On Clock-gated D-flipflop

C. Light V_{th} Hopping

Conventionally, the optimum biasing voltage occurs when the sum of subthreshold and BTBT leakage is minimized. However this optimum biasing voltage is usually too high for active leakage reduction, in which the body charging overhead is not negligible. For active leakage reduction, [5] has demonstrated that choosing small biasing voltage causes slight degradation on leakage reduction, but significantly reduces body charging overhead. We call V_{th} hopping with small biasing voltage as "light" V_{th} hopping. Since charging overhead has a quadratical dependency on biasing voltage, "light" V_{th} hopping is able to dramatically reduce EBT value. Additionally, WUT is reduced as well since small biasing voltage requires less time to be discharged to normal voltage.

Table I compares the experimental results of applying conventional versus light V_{th} hopping on an inverter chain. We can observe that the EBT of light V_{th} hopping is reduced by nearly 11 times. WUT is reduced by 2 times. The trade-off is that leakage reduction is degraded by 13%. However, 85% is still effective for active leakage reduction.

TABLE I
CONVENTIONAL VS. LIGHT V_{th} HOPPING ON INVERTER CHAIN

32nm Tech.	$V_P(V)$	$V_N(V)$	-Leak%	EBT (ns)	WUT (ns)
Conventional.	3.0	-2.0	98%	5.2	0.4
Light.	1.3	-0.4	85%	0.48	0.2

D. Selective Light V_{th} Hopping (SLITH)

Combining "selective" and "light" together, we are able to considerably reduce the EBT and WUT, while keep the effectiveness of leakage reduction. Figure 6 shows an example of SLITH. In Figure 6, several parameters need to determined:

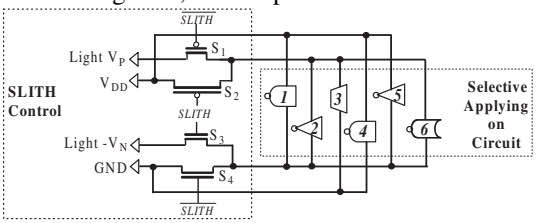

Fig. 6. Selective Light V_{th} Hopping (SLITH)

1) Biasing voltages V_P and $-V_N$. Biasing voltages control the effectiveness of leakage reduction. They are also the dominating factors affecting EBT and WUT. 2) Gate selection. We want to apply SLITH on more gates to reduce more leakage, meanwhile keep the EAW of the selected circuit within constraint. 3) Control transistors S_1 to S_4. Control transistors determine the charging and discharging speed of body capacitances, and thus affect EBT and WUT as well. As we have explained in Section I, if EAW can be reduced to one clock cycle, the energy overhead problem and wake-up time problem will be completely solved. And SLITH can enable seamless joint dynamic and leakage power reduction. So ideally, the constraint of EAW is one clock cycle. Finally, our problem definition is as following.

$$\begin{aligned} Determine: &\quad S_1, S_2, S_3, S_4, V_p, -V_N, \\ &\quad G_k^N = [0,1], G_k^P = [0,1] \ (k = 0..K) \\ Maximize: &\quad E_{saving/ut} \\ Constraint: &\quad T_{EAW} \leq M \cdot T_{clk} \quad (M = 1) \end{aligned} \tag{8}$$

where G_k^N and G_k^P represent the selection of the PDN and PUN of gate k, respectively. The PDN of gate k will be selected if G_k^N equals to 1. K is the total number of the gates in the circuit. $E_{saving/ut}$ and T_{EAW} are modeled in Equation 4 and 7, respectively. T_{clk} is the clock period. So $M \cdot T_{clk}$ $(M = 1)$ sets the constraint of EAW to be one clock cycles. Equation 8 can be solved by a heuristic solver to determine each design parameters of SLITH.

According to Section II.A, the value of EAW is heavily dependent on technology parameters. In older technologies, which in general have low leakage and large body capacitance, SLITH may not be able to archive one-cycle EAW. In this case, the M value in Equation 8 needs to be increased to find to a SLITH solution. As we will show in next section, one-cycle EAW is usually achievable for 32nm technology, while two-cycle EAW is achievable for 45nm technology.

III. EXPERIMENTAL RESULTS

HSPICE simulations have been conducted to verify the effectiveness of SLITH. The temperature is set to be 110 Celsius as runtime temperature. Clock frequency is 2GHZ.

Table II shows the results of applying SLITH on an inverter chain with 32nm and 45nm technologies. PDN% (PUN%) is the percentage of PDN (PUN) receiving SLITH. +A% is the extra area overhead of the charging and discharging transistors

S_1 to S_4. $-Leak\%$ is the reduction of leakage current with the biasing voltages $\triangle V_P$ and $\triangle V_N$. The first row shows that in 32nm, one-cycle EAW is achievable if SLITH is only applied to PDN. This is because PUN generally has larger area and causes larger charging overhead. So in this case, applying SLITH to PUN for just one clock cycle is not able to gain any energy saving. If we loose the constraint of EAW to two cycles, as shown in the second row, SLITH can be applied to both PUN and PDN. The leakage reduction exceeds 50% in this case. However for older technology of 45nm, as shown in the third row, PUN cannot receive SLITH even for two-cycle EAW. Figure 7 shows the HSPICE waveform of applying

TABLE II
SLITH ON AN INVERTER CHAIN WITH 32/45NM TECHNOLOGIES

Tech.	M	$\triangle V_P$ (V)	$\triangle V_N$ (V)	PDN%	PUN%	+A%	-Leak%
32nm	1	0	-0.26	0%	100%	1.7%	37%
32nm	2	0.21	-0.31	100%	100%	5.5%	56%
45nm	2	0	-0.33	0%	100%	0.9%	38%

SLITH with one-cycle EAW in 32nm. The left shows the waveform of N-well voltage. The right shows the waveform of subthreshold leakage current.

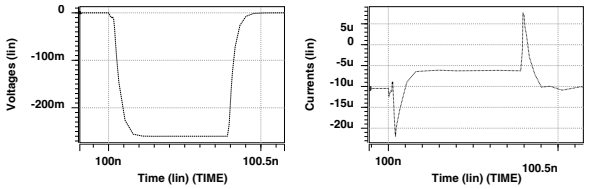

Fig. 7. HSPICE Waveforms of Applying SLITH On Inverter Chain in 32nm

The first row indicates that leakage current reduction is 37% with -0.26V biasing voltage for PDN. However, due to the charging overhead, the net energy saving is less. Figure 8 shows the net energy saving percentage after SLITH is applied for one-cycle EAW in 32nm. It is only 9% at the end of first cycle. However, if the invert chain stays in SLITH mode, the net energy saving increases and reaches 32% at the forth cycle. Therefore, if SLITH is optimized for M-cycle EAW, it yields significant net energy saving only after $M + 1$ cycles.

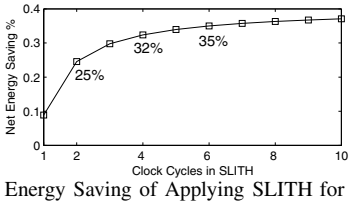

Fig. 8. Net Energy Saving of Applying SLITH for One-cycle EAW

Table III shows extensive experimental results in 32nm. For one-cycle EAW, SLITH is able to achieve 9% to 35% leakage reduction. For two-cycle EAW, the reduction is from 40% to 63%. More results are omitted due to page limitation.

TABLE III
SLITH WITH ONE-CYCLE AND TWO-CYCLE EAW IN 32NM TECH.

Circuit	M	$\triangle V_P$	$\triangle V_N$	PDN%	PUN%	+A%	-Leak%
D-flipflop	1	0	-0.20	0%	95%	1.2%	26%
P. Decoder(C432)	1	0	-0.18	0%	27%	0.3%	9%
ECAT(C499)	1	0	-0.29	0%	72%	0.6%	30%
ALU(C880)	1	0	-0.29	0%	73%	0.6%	27%
Multiplier(C6288)	1	0	-0.22	0%	83%	1%	35%
D-flipflop	2	0.26	-0.35	63%	95%	2.5%	50%
P. Decoder(C432)	2	0.19	-0.25	41%	93%	1.7%	40%
ECAT(C499)	2	0.22	-0.36	27%	97%	1.4%	51%
ALU(C880)	2	0.19	-0.30	86%	98%	3.7%	63%
Multiplier(C6288)	2	0.20	-0.36	46%	100%	1.9%	63%

SLITH with one-cycle EAW can seamlessly collaborate with any dynamic power reduction technique, such as clock gating, pre-computation and operand isolation. Even if the circuit is reactivated after one cycle, the net energy saving is guaranteed to be positive. And its body voltage will recover to the normal voltage. Figure 9 shows the diagram of joint power reduction by pre-computation and one-cycle SLITH. Note that a voltage level converter is needed to control SLITH.

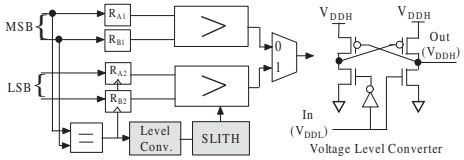

Fig. 9. Joint Reduction by Pre-comp.and One-cycle SLITH

SLITH with two-cycle EAW can be applied when the circuit has at least two cycles of idleness. Figure 10 shows its application in a pipeline. When the idle detector detects any two-cycle idleness in the workload, it generates SLITH and clock gating enable signals. Then the enable signals are passed to the remainder of the pipeline stage by stage.

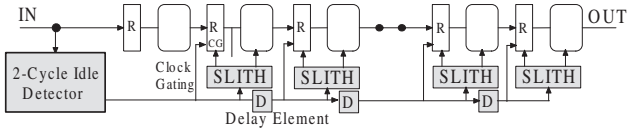

Fig. 10. Joint Reduction on DP. by CG. and Two-cycle SLITH

IV. CONCLUSION

In this paper, we emphasize that the gap between run-time dynamic and leakage power reduction techniques is caused by the large EBT and WUT of conventional leakage reduction techniques. This gap can be diminished, if EAW is reduced. Based on this observation, we propose a new leakage reduction technique called: SLITH. Our preliminary experiments show that SLITH is able to reduce its EAW, to as low as one cycle, while maintaining the effectiveness of leakage reduction. Therefore, it bridges the gap and enables joint dynamic and leakage power reduction. SLITH can be used together with most dynamic power reduction techniques and achieves significant active leakage power reduction.

REFERENCES

[1] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer cmos circuits," in *IEEE*, Feb. 2003, pp. 305–327.

[2] J. Tschanz, S. Narendra, Y. Ye, B. Bloechel, S. Borkar, and V. De, "Dynamic sleep transistor and body bias for active leakage power control of microprocessors," *IEEE J. Solid-State Circuits*, vol. 38, pp. 1838–1845, Nov. 2003.

[3] A. Abddollahi, M. Pedarm, F. Fallah, and I. Ghosh, "Precomputation-based guarding for dynamic and leakage power reduction," in *ICCD*, Oct. 2003, pp. 90–97.

[4] K.-S. Min, H.-D. Choi, H.-Y. Choi, H. Kawaguchi, and T. Sakurai, "Leakage-suppressed clock-gating circuit with zigzag super cut-off cmos (zsccmos) for leakage-dominant sub-70-nm and sub-1-v-v/sub dd/ lsis," *IEEE Trans. VLSI Syst.*, vol. 14, pp. 430–435, Apr. 2006.

[5] H. Xu, R. Vemuri, and W.-B. Jone, "Run-time active leakage reduction by power gating and reverse body biasing: An energy view," in *ICCD*, Oct. 2008.

[6] H. Singh, K. Agarwal, D. Sylvester, and K. Nowka, "Enhanced leakage reduction techniques using intermediate strength power gating," *IEEE Trans. VLSI Syst.*, vol. 15, pp. 1215–1224, Nov. 2007.

[7] A. S. University. Predictive technology model. [Online]. Available: http://www.eas.asu.edu/ ptm/