Accurate Energy Breakeven Time Estimation for Run-time Power Gating

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

Abstract—Run-time Power Gating (RTPG) is a recent technique, which aims at aggressively reducing leakage power consumption. Energy breakeven time (EBT), or equivalent sleep time has been proposed as a critical figure of merit of RTPG. Our research introduces the definition of average EBT in a runtime environment. We develop a method to estimate the average EBT for any given circuit block, considering the impact of circuit states. HSPICE simulation results on ISCAS85 benchmark circuits show that the average EBT model has on the average 1.8% error. The CAD tool implemented based on the model can perform fast estimations with a speedup of 3000× over HSPICE.

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]. This makes leakage power reduction an indispensable component in nano-era low power design. Subthreshold leakage, gate leakage and band-to-band tunneling leakage are the three main components contributing to the total leakage power.

Many leakage reduction techniques have been introduced and studied so far, such as forced stacking [2], input vector control [3] and power gating [4]. Among these, power gating, or sleep transistor technique, has been proven to be the most effective one. Currently, most of the power gating applications disable the circuit power when they are in standby mode. However as the technology scales down, more aggressive leakage reduction techniques are required. As a promising technique, runtime power gating (RTPG) has drawn more attention recently [5-10]. As shown in Figure 1, RTPG turns off the power gate once it detects sufficient idleness in circuit workload, even when the circuit is in the active mode. In this way, it is able to exploit more circuit slackness, and thus reduce more leakage power. Furthermore, because the leakage in the active mode is significantly larger due to the higher die temperature in active mode [5], the study of RTPG is even more important.

Since RTPG switches the power gate more frequently, it incurs more energy penalty. Hence, to guarantee RTPG’s effectiveness, designers need to make sure the leakage energy saving by applying RTPG is larger than the switching penalty. This has been identified as the key design problem of RTPG in [6-8]. The concept of energy breakeven time (EBT) [6], or equivalent sleep time [7], or minimum idle time [8] has been introduced to quantify the trade-off between energy saving and penalty. Precisely, EBT \( T_B \) is defined as the sleep time of a circuit, at which leakage energy saving \( E_s(t) \) catches up with the energy penalty \( E_p \) for entering the sleep mode:

\[
E_s(T_B) = E_p \tag{1}
\]

Thus, if the circuit stays in the sleep mode for a period longer than EBT, energy saving can be archived. In order to calculate EBT, the accurate leakage energy saving model \( E_s(t) \) as a function of time is required. This further necessitates the modeling of leakage reduction process after the power is gated. For example, consider ground power gating in Figure 1. After the circuit ground is gated, its internal nodes will be gradually charged up by the leakage current. In return, the leakage current decreases. Finally, the virtual ground voltage reaches a final value. The leakage current is then minimized to a steady value \( I_{min} \) at time \( T_{steady} \), in Figure 2. The leakage current varies significantly over time before \( T_{steady} \).

\[
T_B = \frac{E_p}{P_{active} - P_{sleep}} \tag{2}
\]

where \( P_{active} \) and \( P_{sleep} \) are the leakage power consumption in active mode and sleep mode, respectively. This expression assumes that the leakage current is a constant \( I_{min} \) after the power gating is applied. This assumption introduces error (shade area in Figure 2) into the energy saving and EBT.
estimations. For standby mode power gating, since the circuit sleep time is long ($T_{\text{standby}}$ in Figure 2), this error is negligible. However, for RTPG applications, the sleep time ($T_{\text{RTPG}}$) is generally short. If $T_{\text{RTPG}}$ is designed to be less than $T_{\text{steady}}$, this error is quite significant. This is quite possible in the future technology generations. For example, our HSPICE simulation results show that in 32nm technology, the leakage current approaches its steady value at around 80ns on the average, while the average EBT occurs around 2ns. In this case, $T_{\text{RTPG}}$ can be selected as any value between 2ns and 80ns. The modeling of the leakage current variation before $T_{\text{steady}}$ is then essential for the accurate EBT estimation.

Some recent researches have considered the leakage current variation before $T_{\text{steady}}$. Yu et al. [5] multiply the total leakage energy saving with an empirical value (0.73) to cover the variation. In [10] and [6], Hu et al. and Usami et al. derive EBT models factoring in the variation. However, their models do not consider the impact of circuit topologies and circuit states. Since leakage has a strong dependency on the circuit states [1], the models in [10] and [6] are rather high level models and cannot be used when accurate estimates are necessary.

In this paper, we develop a fast and accurate method to estimate the EBT. Our method has three novel features: (1) It is applicable to any static CMOS technology; (2) It is applicable to any circuit topology; (3) It includes the impact of circuit states. This paper is organized as follows. Section II shows the modeling of leakage current reduction process for power gating. Section III gives the definition of average EBT in a run-time environment and derives the average EBT model. Section IV shows the experimental results. Finally Section V concludes the paper.

II. LEAKAGE CURRENT REDUCTION MODELING FOR GROUND GATING

In order to derive the leakage energy saving and EBT model, we first model the leakage current reduction (LCR) process after the circuit power is gated. To this end, we derive the LCR model from the transistor level to the circuit level.

In this paper, we only study the ground power gating, since it is widely adopted in most design practices. Supply power gating can be studied with a similar method. Also, we consider the sub-threshold leakage current as the major leakage source, since the gate leakage is projected to have a reduction of 100× with the use of high-k dielectrics [11]. Additionally, due to the exponential dependency of the sub-threshold leakage on temperature, the sub-threshold leakage is still the dominant leakage component during run-time. So we focus on the sub-threshold leakage and refer to it as leakage.

A. Leakage Current Reduction Modeling At Transistor Level

We start from the full-fledged subthreshold leakage current model for a single transistor [1]:

$$I = A \cdot e^{\frac{1}{m} \cdot m \cdot r \cdot V_G - V_S - V_{th0} - \gamma' \cdot V_S + \eta \cdot V_{DS}} \cdot (1 - e^{-V_{DS}/v_T})$$

with $A = \mu_0 C_{ox} \frac{W}{L_{w}} (v_T)^2 \cdot e^{V_{th0}/v_T}$

(3)

where $V_{th0}$ is the zero bias threshold voltage, $v_T$ is the thermal voltage, $\gamma'$ is the linearized body effect coefficient, and $\eta$ is the DIBL coefficient. When $V_{DS} \gg v_T$, the term $(1 - e^{-V_{DS}/v_T})$ in the above equation can be neglected. Assuming that the temperature and the body bias voltage are fixed, we can simplify Equation 3 into:

$$I = \tilde{I} \cdot e^{K_1 V_G + K_2 V_D - K_3 V_S}$$

(4)

where $K_1$, $K_2$ and $K_3$ are technology-dependant exponents, and $\tilde{I}$ is the leakage current under the normal condition terminal voltages. Note that the simplification of the term $(1 - e^{-V_{DS}/v_T})$ in Equation 3 causes inaccuracy when $V_{DS}$ is comparable with $v_T$. (When $V_{DS}$ equals to $4v_T$, the error is 1.8%). Hence our model is accurate when $V_{DS} > 4v_T$. When $V_{DS} < 4v_T$, the leakage current is very small such that it can be approximated as zero in EBT calculation.

Now apply ground gating to a single transistor. With ground gating, the terminal voltages of the off-state NMOS or PMOS have the patterns as shown in Figure 3. Note that $V_{VG}$ stands for virtual ground voltage.

The leakage current of them can be expressed as:

$$\begin{cases}
  I = \tilde{I} \cdot e^{K_1 V_G + K_2 V_D - K_3 V_S} \\
  V_G = V_{VG}(NMOS) \\
  V_G = V_{DD}(PMOS)
\end{cases}$$

(5)

Solving the above equations yields:

$$\begin{cases}
  I_N = \tilde{I}_N \cdot e^{K_1 V_{VG} + K_2 V_{DD} - K_3 V_{VG}} \\
  I_P = \tilde{I}_P \cdot e^{K_1 V_{DD} + K_2 V_{VG} - K_3 V_{DD}}
\end{cases}$$

(6)

where $\tilde{I}_P$ ($\tilde{I}_N$) is the leakage current of PMOS (NMOS) with zero $V_{VG}$, and $K_P$ ($K_N$) is the leakage reduction exponent of PMOS (NMOS). So a ground gated off-state transistor can be modeled as a $V_{VG}$-controlled current source. The footer has a similar model since essentially it is a NMOS transistor with high threshold voltage.

Next, we study the LCR of a ground gated NMOS. As soon as the footer is turned off, the leakage current starts to charge the internal nodes. As a result, the $V_{VG}$ level gradually rises. Consequently, the leakage current is reduced according to Equation 6. As shown in Figure 4, we model this as a charging process of the virtual ground capacitor by $V_{VG}$-controlled current sources. The resistance of on-state transistors and interconnects does not affect the charging process, since the charging is due to an equivalent current source. The resistance is then neglected in our model. We have designed an experiment in Section IV to verify this.
By considering $V_{VG}$ and leakage current as functions of time, the charging process can be characterized as:

\[
\begin{align*}
V_{VG}(t) &= \frac{1}{C} \int_0^t I_{\text{charge}}(t) \, dt \\
I_{\text{charge}}(t) &= I_{NMOS}(t) = I_{\text{foot}}(t) \\
I_{NMOS}(t) &= \frac{1}{2} \cdot e^{-K_N V_{VG}(t)} \\
I_{\text{foot}}(t) &= \frac{1}{2} e^{-K_F V_{VG}(t)}
\end{align*}
\]

(7)

where $t$ is the time after the ground is gated, and $C$ is the equivalent capacitor attached to the virtual ground. In this case, $C$ includes the junction capacitances of the NMOS and the footer, and gate capacitance of the NMOS. $I_F$ is the footer leakage current when $V_{VG}$ equals to $V_{DD}$. $I_F$ is the leakage reduction exponent of the footer. In the above equations when the $V_{VG}$ level is low, the leakage current of the footer is very small and can be ignored. (It will be considered in the circuit level model.) Equations 7 can then be solved as:

\[
\begin{align*}
V_{VG}(t) &= \frac{1}{K_N} \ln \left( \frac{K_N t}{C} \right) \\
I_{NMOS}(t) &= \frac{1}{K_N (t+C)}
\end{align*}
\]

(8)

The above equations give the transistor level LCR model.

B. Leakage Current Reduction Modeling At Gate Level

In the following, we derive the LCR model at the gate level.

1) Example of Two-input NAND Gate

Consider a two-input NAND gate. Since the input vectors have significant impacts on the leakage current, we model the gate by each input vector. As shown in Figure 5, the two-input NAND gate is modeled into four $V_{VG}$-controlled leakage current sources ($I_i$), with equivalent capacitances ($C_i$) attached to the virtual ground of the gate. The next question is to find out the relationship between the leakage current ($I_i$) of the gate and the $V_{VG}$. Since the off-state transistors control the amount of leakage current, our answer starts from studying two basic gate structures: off-state transistors in series and parallel.

2) Off-state Transistors In Series

Case Four in Figure 5 is the simplest series structure with two off-state NMOS transistors in stack. Our first goal is to find out the impact of stacking effect on the leakage current. Consider a general case of transistors in series as shown in Figure 6a. Assume that the upper terminal voltage of the four-transistor stack is $V_{VG}$, and the virtual ground voltage is $V_{VG}$. By using Equation 4, the leakage current for each transistor is given by:

\[
\begin{align*}
I_1 &= \tilde{I}_s e^{-K_1 V_{VG}} + K_2 V_a - K_3 V_t = I_{\text{series}} \\
I_2 &= \tilde{I}_s e^{-K_1 V_{VG}} + K_2 V_a - K_3 V_t = I_{\text{series}} \\
I_3 &= \tilde{I}_s e^{-K_1 V_{VG}} + K_2 V_a - K_3 V_t = I_{\text{series}} \\
I_4 &= \tilde{I}_s e^{-K_1 V_{VG}} + K_2 V_a - K_3 V_t = I_{\text{series}}
\end{align*}
\]

(9)

where $I_{\text{series}}$ is the leakage current of the stack. By substituting $\tilde{I}_s (i=1,2,3,4)$ with $e^{t_i}$, Equations 9 turn into:

\[
\begin{align*}
K_2 V_a - K_3 V_t + I_1 &= \ln(I_{\text{series}}/e^{K_1 V_{VG}}) \\
K_2 V_a - K_3 V_t + I_2 &= \ln(I_{\text{series}}/e^{K_1 V_{VG}}) \\
K_2 V_a - K_3 V_t + I_3 &= \ln(I_{\text{series}}/e^{K_1 V_{VG}}) \\
K_2 V_a - K_3 V_t + I_4 &= \ln(I_{\text{series}}/e^{K_1 V_{VG}})
\end{align*}
\]

(10)

Solving the above equations yields:

\[
I_{\text{series}} = e^{K_1 V_{VG}} \cdot e^{K_2 V_a - K_3 V_t}
\]

(11)

With the stacking effect, the leakage current still has an exponential dependency on $V_{VG}$. The difference is that the leakage reduction exponent turns into an equivalent exponent $K_s$, and the zero-$V_{VG}$ leakage current of the stack turns into $\tilde{I}_{\text{series}}$. Now consider a special case where one of the transistors in the stack is on as shown in Figure 6b. Assume that the voltage difference between $V_1$ and $V_2$ is negligible. This case is then equivalent to a three-transistor stack. Hence if some transistors in the stack are in the on-state, the total leakage current of the stack still satisfies Equation 11.

3) Off-state Transistors In Parallel

Case one in Figure 5 is the simplest example of off-state transistors in parallel. Since two off-state PMOS have the same leakage reduction exponent $K_F$, the total leakage remains an exponential function, despite of the transistors size. Similarly, for a parallel structure with only one transistor in each branch (Figure 7a), the total leakage is an exponential function of $V_{VG}$. The parallel structure can be complex when a branch has more than one transistor. Due to the stacking effect, it can have different $K$ values depending on the number of off-state transistors in that branch. For example in Figure 7b, by using

\[
\begin{align*}
I_{\text{parallel}} &= \tilde{I}_s e^{-K_s V_{VG}} + \tilde{I}_s e^{-K_s V_{VG}} + \tilde{I}_s e^{-K_s V_{VG}} + \ldots
\end{align*}
\]

(12)

where $K_s (i=a,b,c,...)$ is the equivalent leakage reduction exponent of each branch. Since $K_s$ can be different, the total leakage current is no longer a simple exponential function. However, we still approximate $I_{\text{parallel}}$ into a lumped exponential function of $V_{VG}$ because of two reasons:

a) Most gates have the simple parallel structure shown in Figure 7a, for example, NAND gates, NOR gates and Buffers.

b) For complex parallel structures, the $K$ values can be different because the number of off-state transistors in each branch is different. In this case, due to the stacking effect,
the branches with the least number of off-state transistors (branches b and c in Figure 7b) have much larger leakage current than other branches (at least one order of magnitude [1]) and become the dominant ones. Then \( I_{parallel} \) is close to the sum of dominant branches current. Since the dominant branches current have the same \( K \), \( I_{parallel} \) can be approximated as an exponential function:

\[
I_{parallel} = I_0 e^{-K_0 V_{CG}} + I_0 e^{-K_1 V_{CG}} + I_0 e^{-K_2 V_{CG}} + ... \\
\approx I_0 e^{-K_0 V_{CG}} + I_0 e^{-K_1 V_{CG}} \\
= (I_0 + I_0) e^{-K_0 V_{CG}} \quad (K_0 = K_1)
\] (13)

We have shown that the leakage current of any number of off-state transistors in series is an exponential function of \( V_{CG} \). And in parallel, it can be approximated into an exponential function. So for any type of gate structure, we construct a lumped exponential function of \( V_{CG} \) to model the total leakage current of the gate. Conclusively, we define the full-fledged gate model as follows. For each input vector \( i \), the gate is modeled as an equivalent virtual ground capacitor \( C_i \) and a \( V_{CG} \)-controlled current source \( I_i \), satisfying:

\[
I_i = I_0 e^{-K_0 V_{CG}}
\] (14)

where \( I_0 \) is the zero-\( V_{CG} \) leakage current, and \( K_0 \) is the equivalent reduction exponent of the gate. Use this model to substitute the NMOS model in Equations 7 and solve them in a similar way. The gate level LCR model can be obtained.

C. Leakage Current Reduction Modeling at Circuit Level

At the circuit level after the ground is gated, each gate in the circuit can be modeled as a \( V_{CG} \)-controlled current source with a capacitor attached to the virtual ground of the circuit, as shown in Figure 8. Taking the footer leakage current into account, we derive the circuit level model by 3 steps.

![Fig. 8. LCR Modeling At Circuit Level](image)

a) For those gates (\( i \)), which have the same gate type (\( j \)) and receive the same input vector (\( i \)), they can be linearly combined into a single current source (\( I_{ij} \)), with a total equivalent capacitor (\( C_{ij} \)). Hence in this step, we build a downsized circuit model with maximally \( 2^2 \) current sources.

\[
I_{ij} = \sum_{j} (I_{ij}) e^{-K_{ij} V_{CG}}
\] (15)

\[
C_{ij} = \sum_{j} C_{ij}
\]

b) Now the total charging current to the virtual ground of the circuit turns into the summation of all \( I_{ij} \) of each gate in the circuit, subtracted by the footer leakage current:

\[
\begin{align*}
I_{charge} &= I_{circuit} - I_{footer} \\
I_{circuit} &= \sum_{ij} I_{ij} e^{-K_{ij} V_{CG}} \\
I_{footer} &= I_{PE} e^{-K_{PE} V_{CG}}
\end{align*}
\] (16)

The total virtual ground capacitor of the circuit is a linear summation of all the equivalent capacitors \( C_{ij} \). By considering the \( V_{CG} \) and the current as functions of time, the LCR process of the circuit can be characterized as:

\[
\begin{align*}
V_{CG}(t) &= \frac{I_{charge}(t) \text{dt}}{C_{total}} \\
I_{charge}(t) &= \sum_{ij} I_{ij} e^{-K_{ij} V_{CG}(t)} - I_{PE} e^{-K_{PE} V_{CG}(t)} \\
C_{total} &= \sum_{ij} C_{ij}
\end{align*}
\] (17)

The above equations do not have a closed-form solution. To address this issue, we construct a piecewise linear model for the leakage current of each gate, as well as the footer. In detail, we first divide the full \( V_{DD} \) into \( R \) consecutive voltage regions (< \( V_{R}^*, V_{F}^* \) >), where \( V_{R}^* \) (\( V_{F}^* \)) represents the low (high) bound voltage of region \( r \). In each region \( r \) \((r = 1..R)\), we perform linear regression on the exponential leakage current models in Equations 16. We then have:

\[
\begin{align*}
I_{ij}^r &= -S_{ij}^r V_{CG} + Z_{ij}^r \\
I_{footer}^r &= -S_{PE}^r V_{CG} + Z_{PE}^r
\end{align*}
\] (18)

where \( S_{ij}^r \) and \( Z_{ij}^r \) are the linearized coefficients for gate level leakage current models in each region \( r \), and \( S_{PE}^r \) and \( Z_{PE}^r \) are the linearized coefficients for the footer leakage model. Linear regression incurs error \( D_{ij}^r(D_{PE}) \) to the leakage current models. It can be alleviated by increasing the \( R \) value. However, the computation complexity also increases when \( R \) is larger. So the selection of \( R \) is a trade-off between accuracy and speed. In Algorithm I, we set an error threshold \( (D_{th}) \) to determine \( R \).

c) Once we obtain the linearized leakage current models, we use them to substitute all exponential leakage current models in Equations 17. Then the total charging current also transforms into a piecewise linear model:

\[
\begin{align*}
I_{charge} &= \sum_{ij} (-S_{ij}^r V_{CG} + Z_{ij}^r) - (-S_{PE}^r V_{CG} + Z_{PE}^r) \\
&= (-S_{ij}^r V_{CG} + Z_{ij}^r) - (-S_{PE}^r V_{CG} + Z_{PE}^r) \\
&= -S^r V_{CG} + Z^r
\end{align*}
\] (19)

where \( S^r \) and \( Z^r \) are the linearized model coefficients for the total leakage current of the circuit in each voltage region \( r \), and \( S^r \) and \( Z^r \) are the linearized model coefficients for the total charging current to the virtual ground. Figure 9 illustrates the transform in Equation 19. Substitute the total charging current in Equations 17 with the above linearized model. Make \( V_{CG} \) and the charging current as functions of time. Solving it yields:

\[
\begin{align*}
V_{CG}(t) &= \frac{\tau_{r} \text{dt}}{S_{total}} \quad + \quad \frac{S_{r} V_{r}^* + Z^r e^{-\frac{Z^r}{S_{total}} (t-\tau)}}{S_{total}} \\
I_{charge}(t) &= I_0 e^{-K_0 V_{CG}(t)}
\end{align*}
\] (20)

where \( V_{r}^* \) and \( \tau \) are the initial charging current, initial voltage and initial time of the region \( r \), respectively. The initial voltage and time for region 1 are both zero. The initial charging current of region 1 is the original leakage current.

![Fig. 9. Derivation Of Piecewise Total Charging Current](image)
without ground gating. In region $r$ ($r=2..R$), these three initial conditions can be obtained by solving the following equations in region $r-1$ recursively:

$$\begin{align*}
V_{G}^{r-1}(t) &= V^{r-1} \\
I_0^r &= I_{\text{charge}}^{r-1}(t) \\
V_i^r &= V_h^r
\end{align*}$$

(21)

Finally, the total leakage current of the circuit is given by:

$$I_{ckt}^r(t) = -S_r^t V_{G}^r(t) + Z_c^r$$

(22)

It gives the piecewise LCR model of the whole circuit.

D. Secondary Physical Phenomena

Some other secondary physical phenomena will occur during the leakage reduction process. For example in the previous study, the equivalent capacitance attached to the virtual ground is considered to be a constant. However, it is shown that the gate capacitance of a transistor has a dependency on its $V_{GS}$[12], which is a changing value during the $V_{G}$ charging process. So our model needs to take this phenomenon into consideration. To address this problem, we obtain an equivalent capacitor $C_{\text{total}}^r$ in each voltage region $r$, and use it to replace the $C_{\text{total}}$ in Equation 20.

Due to the page limitation, other secondary phenomena will not be discussed in this paper.

III. EBT MODELING

In this section, we first derive the EBT model for the circuit, which has only one specific state when entering the sleep mode. Secondly, we give the definition of average EBT for the circuit during run-time, with multiple possible states. Lastly, we develop a method to estimate the average EBT.

A. EBT Modeling for Circuit In A Specific State

Based on the piecewise LCR model in Equation 22, we have the energy saving model $(E_{s}(t)^{r})$ in each region $r$:

$$E_{\text{piece}}^{r}(t) = \int_{0}^{t} V_{DD}(I_{ckt0} - I_{ckt}^{r}(t))dt$$

(23)

where $I_{ckt0}$ is the original leakage current of the circuit without power gating. It can be obtained by applying Equation 22 in voltage region 1 at time zero:

$$I_{ckt0} = -S_1^t V_{G}^1(0) + Z_c^1$$

(24)

$E_{\text{piece}}^{r}(t)$ represents the energy saving only in region $r$, until time $t$. The total energy saving $(E^{r})$ of all previous regions until time $t$ is:

$$E^{r}(t) = \sum_{m=1-(r-1)}^{m} E_{\text{piece}}^{m} + E_{\text{piece}}^{r}(t)$$

(25)

where,

$$E_{\text{piece}}^{m} = \int_{0}^{t} V_{DD}(I_{ckt0} - I_{ckt}^{m}(t))dt$$

(26)

where successive region $m$ spans from $\tau_{m}$ to $\tau_{m+1}$. $E_{\text{piece}}^{m}$ then represents the energy saving in region $m$ from the beginning to the end of the region, as shown in Figure 10. So the summation of all $E_{\text{piece}}^{m}$ from 1 to $r-1$ represents the total energy saving until region $r-1$.

Equation 25 gives the piecewise total energy saving model. Based on it, we are able to solve the EBT $(T_B)$ by matching the energy saving with energy penalty $(E_P)$:

$$E^{r}(T_B) = E_P$$

(27)

The energy penalty is the dynamic power consumption for switching the gate capacitance $(C_{\text{foot}})$ of the footer:

$$E_P = C_{\text{foot}} V_{DD}^2$$

(28)

EBT can be solved by putting Equations 20, 22, 24, 25 and 27 together:

$$\begin{align*}
I_{ckt}^r(Z_1 - I_{ckt}^r(t))dt &= E_P - \frac{1-(r-1)}{\sum_{m=1}^{r-1} E_{\text{piece}}^{m}} E_{DD}^m \\
I_{ckt}^r(t) &= -S_r^t V_{G}^r(t) + Z_c^r \\
V_{G}^r(t) &= \frac{2r^2}{a^2} + \frac{S_r^t V_{G}^r(t)}{e^{-\frac{t}{\tau}}} - \frac{2r^2}{a^2} (t-r^+)
\end{align*}$$

(29)

Simplifying the above equations yields:

$$a_1 e^{\theta (t-r^+)} + a_2 t + a_4 = 0$$

(30)

where $a_1$ to $a_4$ are the constants given by $Z^r$, $S^r$, $Z_c^r$, $V_{G}^r$, $C_{\text{total}}^r$, $Z_c^r$ and $E_P$. The above equation leads to no closed-form solution. Once again, we use linear regression to approximate it.

In detail, assuming that EBT happens in voltage region $\theta$ (Figure 10), linear regression will be performed on the leakage current model $I_{ckt}^\theta(t)$ and yields:

$$I_{ckt}^\theta(t) = -Ft + G$$

(31)

Linear regression on $I_{ckt}^\theta(t)$ is only needed in region $\theta$, instead of every region. This evolves the determination of $\theta$. It will be explained in detail in Section III.C.

Substitute the $I_{ckt}^\theta(t)$ in Equation 29 with 31 and set $r$ as $\theta$. Simplifying them yields:

$$A(T_B - \tau^\theta)^2 + B(T_B - \tau^\theta) - C = 0$$

where

$$\begin{align*}
A &= \frac{1}{\tau} F \\
B &= Z_c^\theta - G \\
C &= E_P - \frac{1-(r-1)}{\sum_{m=1}^{r-1} E_{DD}^m} E_{DD}^m
\end{align*}$$

(32)

Solving it yields:

$$T_B = \tau^\theta + \sqrt{\frac{B^2}{4A^2} + \frac{C}{A} - \frac{B}{2A}}$$

(33)

The above equation gives the EBT model for the circuit under a specific state.

B. Average EBT Definition For Multiple States Circuit

So far, we have modeled the EBT with the assumption that the circuit enters the sleep mode with one specific state. However during run-time, the circuit may be in different states when the footer is turned off. Since in different states, the leakage power consumption can vary significantly [1], our EBT estimated for one state may not be accurate for another state, as shown in Figure 11.
So our next question is: is there an average EBT value \( (T_U) \) for the circuit, which has \( U \) possible states when entering the sleep mode?

Assuming that \( O_u \) is the number of occurrences for the circuit sleeping under state \( u \), our average EBT definition is:

\[
E_U^u(t) = \sum_{u} O_u E_U^u(T_U) = E_P \sum_{u} O_u
\]

\( E_U^u(t) \) is our piecewise energy saving model for the circuit sleeping under state \( u \). \( E_U^u(T_U) \) is the energy saving at time \( T_U \). By multiplying \( E_U^u(T_U) \) with its corresponding \( O_u \) and summing them up, the left side of the above equation is the overall energy saving of all sleep occurrences, at time \( T_U \). The right side is the total energy penalty. In some states, the energy saving \( E_U^u(T_U) \) will be more than its penalty \( E_P \), and in other states less. However, the definition of the average EBT guarantees that the overall saving compensates the penalty. Equation 34 can also be expressed by the probability \( (P_u) \) of the occurrences of each state:

\[
\sum_{u} P_u E_U^u(T_U) = E_P
\]

We use the above probability expression in the following study.

C. Average EBT Estimation

We develop the average EBT estimation method in 4 steps:

a) Build piecewise LCR model \( I^u_r(t) \) for each state \( u \) in each region \( r \), using Equation 22. Build piecewise energy saving model \( E_U^u(t) \) using Equation 25. Also, calculate the initial time \( (\tau^u_r) \) using equation 21.

b) Denote overall energy saving of every sleep occurrence at time \( t \) as \( E^u_r(t) \). By the average EBT definition, we have:

\[
E^u_r(t) = \sum_{u} P_u E_U^u(t)
\]

So we need to obtain the overall energy saving as a function of time. However, the piecewise energy model \( E_U^u(t) \) cannot be simply summed up by the voltage region \( r \). This is because each \( E_U^u(t) \) may have different time range \( (\tau^u_r, \tau^u_{r+1}) \) for the same \( r \). For example in Figure 12, the individual energy saving of the three states \( (E_1^1(t), E_2^1(t), E_3^1(t)) \) in voltage region 1 can not be summed up, because they have different ending time.

We create time regions to address this issue. In detail, we sort all the initial time \( \tau^u_r \) into a low-to-high list. Between every two elements in the list, create a time region \( w \) \((w = 1..W, W = RU)\), as shown in Figure 12.

Thus, we can sum up \( E^u_r(t) \) by the time region \( w \) and build a piecewise overall energy saving mode \( E^w_r(t) \):

\[
E^w_r(t) = \sum_u P_u E^u_r(t)
\]

c) By starting from the first region, we traverse through each region \( w \). In each \( w \), calculate the overall energy saving \( (E^w_r) \) until the end time of \( w \). This can be achieved simply by calculating \( E^w_r(t) \) at time \( \tau^w_{r+1} \):

\[
E^w_r = E^w_r(\tau^w_{r+1})
\]

Assume that in time region \( w \), the overall saving catches up with the penalty:

\[
E^\theta_{w-1} < E_P \leq E^\theta_w
\]

It means that the average EBT occurs in \( \theta \). Similar to the single state EBT, we perform linear regression on the \( I^\theta_{u}(t) \) for each state \( u \). It yields:

\[
I^\theta_u(t) = -F_u t + G_u
\]

d) Now, we can solve the average EBT in time region \( \theta \). By the average EBT definition, and using the linearized circuit leakage current model in Equation 40, we have:

\[
\left\{ \begin{array}{l}
\theta^w_r(T_U) = E_P \\
E^w_r(t) = \frac{E^w_r - E^w_r(0)}{E^w_r(\theta^w_r(T_U)) - E^w_r(0)} \sum_u P_u \int_{t=0}^{\theta^w_r(T_U)} \left( Z^1_u \right) dt \\
I^\theta_u(t) = -F_u t + G_u
\end{array} \right.
\]

Solving the above equations yields:

\[
T_U = \tau^\theta + \sqrt{\frac{B^2}{4A^2} + \frac{C}{A}} - \frac{B}{2A}
\]

where

\[
\begin{align*}
A &= \sum_u \frac{1}{2} P_u F_u \\
B &= \sum_u P_u (Z^1_u - G_u) \\
C &= \frac{E_P - E^\theta_{w-1}}{\frac{B}{2A}}
\end{align*}
\]

Equation 42 gives the final solution of the average EBT. The whole solving process of the average EBT can be described in Algorithm 1.

**Algorithm 1: Average EBT Estimation**
IV. EXPERIMENTAL RESULTS

We conduct experiments to compare our model estimates with HSPICE simulation results. The ISCAS85 benchmark circuits in 32nm, 45nm and 65nm technologies [13] are used in the experiments. The gate level implementations are from [14]. We insert footers into the benchmark circuits to implement the ground gating. Since footer sizing does not affect the correctness of our model, it is designed to be equal to the total width of all the parallel NMOS in the circuit for simplicity. The simulation temperature is set to be 110°C to emulate the runtime temperature. The gate leakage is set to be zero.

The most direct way to verify the LCR model is to perform comparisons on the time-varying leakage current. However, since our final target is EBT, the accuracy of leakage energy estimation is more important towards this goal. Therefore in the following experiments, all comparisons are performed on the time-varying leakage energy consumption. In the simulations, right after the footer is turned off, we collect the data \(E_{\text{exp}}(n)\) on the time-varying leakage energy consumption of the whole circuit. This is done by integrating the leakage current of the circuit \(V_D\) over time. Note that each \(n\) in \(E_{\text{exp}}(n)\) represents a sample point of time. In our experiments the step of \(n\) is 0.1ns.

### A. Resistance Impact On The Model

To verify the resistance impact on the model, we replaced all the original parasitic resistances in each circuit by 5 (30) times of their own values. Then the simulated \(E_{\text{exp}}(n)\) of the new circuit, with larger resistances is compared with the simulated \(E_{\text{exp}}(n)\) of the original circuit. As shown in Table I, the larger resistance causes a maximal variation of 1.48% to the \(E_{\text{exp}}(n)\) in all three technologies. This experiment proves that resistance can be neglected in the model.

<table>
<thead>
<tr>
<th>BM.</th>
<th>Gate</th>
<th>Cnts.</th>
<th>32nm</th>
<th>45nm</th>
<th>65nm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>5x 30x</td>
<td>5x 30x</td>
<td>5x 30x</td>
</tr>
<tr>
<td>C432</td>
<td>160</td>
<td>0.20%</td>
<td>0.14%</td>
<td>0.03%</td>
<td>0.53%</td>
</tr>
<tr>
<td>C1355</td>
<td>546</td>
<td>0.12%</td>
<td>0.30%</td>
<td>0.10%</td>
<td>0.23%</td>
</tr>
<tr>
<td>C2670</td>
<td>1193</td>
<td>0.01%</td>
<td>0.13%</td>
<td>0.06%</td>
<td>0.36%</td>
</tr>
<tr>
<td>C3540</td>
<td>1669</td>
<td>0.02%</td>
<td>0.08%</td>
<td>0.08%</td>
<td>0.45%</td>
</tr>
<tr>
<td>C7552</td>
<td>3521</td>
<td>0.01%</td>
<td>0.07%</td>
<td>0.02%</td>
<td>0.28%</td>
</tr>
</tbody>
</table>

**B. Lumped Gate Level Leakage Current Model Accuracy**

In Section II.B, we model the leakage current of a gate as a lumped exponential function of \(V_{\text{VG}}\). To verify this, we conduct experiments on an 8-input AND gate in 32nm technology. This complex gate consists of a 2-input NAND gate, two 3-input NAND gates and a 3-input NOR gate. Given a certain input vector as shown in Figure 13, this gate includes all the serial and parallel structures discussed in Section II.B.

According to our method, the leakage current of a gate is modeled as a lumped \(V_{\text{VG}}\)-controlled current source. So in this experiment, we set the \(V_{\text{VG}}\) of the gate to 7 different values and measure the leakage current of the gate. The predicted leakage current values and HSPICE simulations are shown in Table II and Figure 14. The error is negligible when the \(V_{\text{VG}}\) level is low (\(V_{\text{VG}} < 500\,\text{mV}\)). The error percentage is higher when the \(V_{\text{VG}}\) level is high (\(V_{\text{VG}} > 500\,\text{mV}\)). However when \(V_{\text{VG}}\) is high, since the absolute value of the leakage current is very small, the error has minor impact on the EBT model.

**TABLE II**

<table>
<thead>
<tr>
<th>(V_{\text{VG}}) (mV)</th>
<th>0</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
<th>500</th>
<th>600</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model (I_{\text{leak}}) (nA)</td>
<td>56.9</td>
<td>24.7</td>
<td>10.7</td>
<td>4.6</td>
<td>2.0</td>
<td>0.9</td>
<td>0.4</td>
</tr>
<tr>
<td>Sim. (I_{\text{leak}}) (nA)</td>
<td>56.9</td>
<td>23.9</td>
<td>10.2</td>
<td>4.5</td>
<td>2.2</td>
<td>1.2</td>
<td>0.7</td>
</tr>
</tbody>
</table>

**Fig. 14. Correlation of Lumped Gate Leakage Current Model on AND8**

**C. Circuit Level Leakage Reduction Process Model Accuracy**

To verify the circuit level LCR model, we compare the model estimates on the time-varying leakage energy consumption with simulation data \(E_{\text{exp}}(n)\) for each benchmark circuit in each technology. The average and worst case errors of our model estimates are shown in Table III. Figure 15 shows the correlation over time on benchmark circuit C7552 in 32nm technology. The time (X axis) is normalized by EBT and spans from 0 to 12 EBT. The accurate modeling of LCR in this time range captures the variation of the leakage current. Especially, the error at the time point of 1 EBT and vicinity determine the accuracy of EBT estimation.

**TABLE III**

<table>
<thead>
<tr>
<th>BM.</th>
<th>Gate</th>
<th>Cnts.</th>
<th>32nm</th>
<th>45nm</th>
<th>65nm</th>
</tr>
</thead>
<tbody>
<tr>
<td>C432</td>
<td>160</td>
<td>2.1%</td>
<td>1.6%</td>
<td>1.2%</td>
<td>0.9%</td>
</tr>
<tr>
<td>C499</td>
<td>202</td>
<td>4.2%</td>
<td>1.3%</td>
<td>1.3%</td>
<td>0.7%</td>
</tr>
<tr>
<td>C880</td>
<td>383</td>
<td>3.4%</td>
<td>2.1%</td>
<td>3.5%</td>
<td>2.0%</td>
</tr>
<tr>
<td>C1355</td>
<td>546</td>
<td>2.2%</td>
<td>1.9%</td>
<td>2.4%</td>
<td>1.9%</td>
</tr>
<tr>
<td>C1908</td>
<td>880</td>
<td>3.3%</td>
<td>1.4%</td>
<td>3.4%</td>
<td>2.1%</td>
</tr>
<tr>
<td>C2670</td>
<td>1193</td>
<td>4.6%</td>
<td>2.5%</td>
<td>4.3%</td>
<td>1.6%</td>
</tr>
<tr>
<td>C3540</td>
<td>1669</td>
<td>4.1%</td>
<td>2.4%</td>
<td>3.5%</td>
<td>2.2%</td>
</tr>
<tr>
<td>C3515</td>
<td>2307</td>
<td>3.9%</td>
<td>1.9%</td>
<td>2.9%</td>
<td>1.9%</td>
</tr>
<tr>
<td>C6288</td>
<td>2406</td>
<td>4.1%</td>
<td>2.2%</td>
<td>3.6%</td>
<td>1.6%</td>
</tr>
<tr>
<td>C7552</td>
<td>3521</td>
<td>5.7%</td>
<td>2.7%</td>
<td>3.3%</td>
<td>2.5%</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td>6.7%</td>
<td>2.0%</td>
<td>3.4%</td>
<td>1.7%</td>
</tr>
</tbody>
</table>

**Fig. 15. Error of Leakage Energy Consumption Estimates on C7552**
D. Average EBT Estimation Accuracy

To verify the average EBT estimation, we generate 64 random input vectors for each benchmark circuit. The circuit is then put into sleep 64 times with a different state each time. The probabilities of the occurrences of each state are assumed to be the same. ($P_u = 1/64$). In each sleep occurrence, we simulate the leakage current ($I_{exp}$) before ground gating, and simulate the time-varying leakage energy consumption ($E_{exp}(N)^n$) after ground gating. Also, we simulate the switching energy penalty of the footer. We multiply this penalty by a factor ($\delta$) in order to emulate the extra energy penalty of the control circuit. Since the control circuit penalty is very likely to be smaller than the footer switching penalty, we set $\delta$ as 2 to be conservative. Finally with all the data, we find out a sample point ($N$) satisfying:

$$E_{penalty} = N \sum_{1 \leq i \leq 64} \frac{1}{64} V_{DD} I_{exp}^n - \sum_{1 \leq i \leq 64} \frac{1}{64} E_{exp}(N)^n$$

(43)

The time corresponding to ($N$) is the simulated average EBT. It is compared with the model estimates in Table IV. The results show that our model estimates have on the average 1.8%, maximally 3.0% error when compared with HSPICE.

<table>
<thead>
<tr>
<th>BM</th>
<th>32nm</th>
<th>Sim. Mod. Er.</th>
<th>46nm</th>
<th>Sim. Mod. Er.</th>
<th>65nm</th>
<th>Sim. Mod. Er.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C443</td>
<td>1.59</td>
<td>2.00</td>
<td>2.3%</td>
<td>1.82</td>
<td>2.30</td>
<td>2.7%</td>
</tr>
<tr>
<td>C449</td>
<td>1.58</td>
<td>1.61</td>
<td>1.9%</td>
<td>1.59</td>
<td>1.65</td>
<td>1.7%</td>
</tr>
<tr>
<td>C880</td>
<td>1.50</td>
<td>1.54</td>
<td>2.7%</td>
<td>1.53</td>
<td>1.57</td>
<td>2.3%</td>
</tr>
<tr>
<td>C3515</td>
<td>1.46</td>
<td>1.49</td>
<td>2.0%</td>
<td>1.50</td>
<td>1.54</td>
<td>2.4%</td>
</tr>
<tr>
<td>C1098</td>
<td>1.28</td>
<td>1.32</td>
<td>2.8%</td>
<td>1.29</td>
<td>1.33</td>
<td>2.7%</td>
</tr>
<tr>
<td>C6270</td>
<td>1.36</td>
<td>1.39</td>
<td>2.3%</td>
<td>1.37</td>
<td>1.40</td>
<td>2.5%</td>
</tr>
<tr>
<td>C3540</td>
<td>1.41</td>
<td>1.45</td>
<td>2.4%</td>
<td>1.44</td>
<td>1.47</td>
<td>2.5%</td>
</tr>
<tr>
<td>C5315</td>
<td>1.44</td>
<td>1.48</td>
<td>2.7%</td>
<td>1.47</td>
<td>1.50</td>
<td>2.3%</td>
</tr>
<tr>
<td>C6288</td>
<td>2.19</td>
<td>2.26</td>
<td>3.0%</td>
<td>2.30</td>
<td>2.36</td>
<td>2.9%</td>
</tr>
<tr>
<td>C7552</td>
<td>1.39</td>
<td>1.43</td>
<td>2.5%</td>
<td>1.43</td>
<td>1.47</td>
<td>2.5%</td>
</tr>
</tbody>
</table>

Overall 2.5% 1.9% 1.1%

In the above table for 32nm technology, the variation of the average EBT for different circuit topologies can be 71% (C1908 VS. C6288). However, by using the EBT model in [5,6,10], the average EBT for different circuits will be the same. Hence their models can have at least 71% inaccuracy when circuit topology is considered.

Circuit states also have significant impact. Using the single state circuit EST model in Section III.A, we have an individual model for each technology. Since the average EBT is determined by the distribution of $T_u$, this observation indicates that for a large circuit, its average EST is likely to converge quickly by considering a limited number of circuit states.

HSPICE takes 47 hours to obtain the average EBT of C7552, while our model estimation takes 1 minute (3000× speedup).

V. CONCLUSION

In this paper, we emphasize the energy breakeven time as a critical design parameter for RTPG. We first derive a model to characterize the leakage current reduction process of ground gated circuits. Next, we give the definition of average EBT in a run-time environment. Finally, we develop a method to estimate the average EBT for any given circuit topology, counting in the impact of circuit states. HSPICE Simulation results of ISCAS85 benchmark circuits show that the average EST model has on the average 1.8%, maximally 3.0% error. The CAD tool implemented based on the model can perform fast estimations with a speedup of 3000× over HSPICE.

The simulation results provides insights to the EBT problem. In Table IV, the EBT value reduces when the technology scales down. This is because the ratio of dynamic to leakage power consumption decreases with smaller transistor feature size. In 32nm technology, the average EBT is around 1ns to 2ns. Assume that the clock speed is 5GHz. Then within 5 to 10 clocks, the energy saving obtained by ground gating is able to compensate the penalty. Therefore, we consider that RTPG has substantial potentials in the 32nm technology.

REFERENCES


