2011 24th Annual Conference on VLSI Design

A Robust and Reconfigurable Multi-Mode Power Gating Architecture*

Z. Zhang[1], X. Kavousianos[1,2], K. Chakrabarty[1] and Y. Tsiatouhas[2]
[1]Dept. of Electrical & Computer Engineering, Duke University, 27708 Durham, NC, USA
[2]Dept. of Computer Science, University of Ioannina, 45110 Ioannina, Greece
e-mail: zz18@ee.duke.edu, kabousia@cs.uoi.gr, krish@ee.duke.edu, tsiatouhas@cs.uoi.gr

***Abstract*- Multi-threshold CMOS is a very effective technique for reducing standby leakage power during long periods of inactivity. Recently, a power-gating scheme was presented to support multiple power-off modes and reduce the leakage power during short periods of inactivity. However, this scheme suffers from high sensitivity to process variations, which impedes manufacturability and also limits its applicability to at most two intermediate power-off modes. We propose a new power-gating technique that is tolerant to process variations and scalable to more than two intermediate power-off modes. In addition, the proposed design requires minimum design effort and offers greater power reduction and smaller area cost than the previous method. Analysis and extensive simulation results demonstrate the effectiveness of the proposed design.**

I. INTRODUCTION

As chip density increases relentless along Moore law, power consumption is emerging as a major burden for contemporary systems, especially for standalone devices [6]. Dynamic power consumption has been effectively tackled by the reduction of the supply voltage level, which is accompanied by a reduction of the transistor threshold voltage for maintaining system performance. The reduction of the threshold voltage has in turn adversely affected the sub-threshold leakage current, which has increased exponentially in recent times. Moreover, as devices keep shrinking, the channel length shortens and the gate oxide thickness reduces, which leads to a reduction of threshold voltage due to drain induced barrier lowering effect. Therefore, the gate oxide tunneling current and the junction leakage are considerably increased [17]. For technologies below 90 nm, leakage power is so high that it is comparable in magnitude to dynamic power consumption.

Many techniques have been presented in the literature for reducing static power. One common approach is to exploit the delay slack of parts of the circuit by implementing non-critical domains using high-V_t cells [4]. High-V_t cells reduce the leakage current at the expense of reduced performance; thus their use on non-critical circuit domains reduces the leakage power considerably without affecting circuit performance. Another very efficient technique involves the partitioning of the system into islands, where each island is a logic region with separate supply rail and unique power characteristics [11], [14], [15]. Separate power management policies can be applied in each region, depending on the performance requirements of the system, thereby further reducing both dynamic and static power.

Various power management techniques can be applied in both active and standby operation modes of the circuit, to exploit idle periods during system use. For reducing power consumption in active mode, dynamic voltage scaling (DVS) is widely used [1], [5], [19]. DVS targets the reduction of dynamic energy consumption, which is proportional to the square of the processor's supply voltage. Thus, using a lower supply voltage level yields a quadratic reduction in the energy consumption at the expense of increased execution time. For managing the power consumption during standby mode, Multi-threshold CMOS (MTCMOS) technology is used [4], [10]. In this approach, high-V_t power switches are inserted between the circuit and the power supply or the ground rail, which are turned off during idle mode, thereby suppressing the leakage current. Power switches are carefully sized as they affect circuit performance due to the reduced gate drive as well as due to the increased threshold of the circuit transistors caused by the body effect [7].

A limitation of MTCMOS is that the time required for recovering from the idle mode, referred to as the *wake-up time*, is long relative to circuit clock rates; therefore, the wake-up time prohibits the use of power switches during short periods of inactivity. In [2], [9], [13], [18] it was shown that further leakage power savings can be achieved by exploiting the short periods of inactivity as well. The authors of [9] proposed a structure with one intermediate power-off mode, which reduces the wake-up time at the expense of reduced leakage current suppression. Similar structures were proposed in [2], [13]. The authors of [18] extended this trade-off between wake-up overhead and leakage power savings into multiple power-off modes. Using these techniques, instead of consuming power by remaining in the active mode during the short periods of inactivity, the circuit is put into an appropriate power-off mode (i.e., low-power state), which is determined by both the wake-up time and the length of the idle period. The longer the period of inactivity, the higher are the power savings achieved by using the most aggressive power-off mode that can be tolerated.

Even though the architecture proposed in [18] is efficient for reducing leakage power, it has several drawbacks that seriously limit its applicability. First, it cannot be easily extended to support more than two intermediate power-off modes and thus it cannot fully exploit the power reduction potential of the power-gating structure, especially for high performance circuits. Second, the architecture in [18] consumes a significant amount of power, and this reduces the benefits offered by the power switches. Third, this structure is very sensitive to process variations, which can adversely affect its manufacturability and predictability. Finally, it is not-easily testable as it consists of analog components.

In this work, we present an effective and robust multi-mode power-gating architecture that has none of the above drawbacks of the architecture proposed in [18]. The proposed structure requires minimal design effort since it is very simple, and with no analog components. It is considerably

*This research was supported in part by the National Science Foundation under grant no. CCF-0903392, and by the Semiconductor Research Corporation under contract no. 1992.

1063-9667/11 $26.00 © 2011 IEEE
DOI 10.1109/VLSID.2011.29

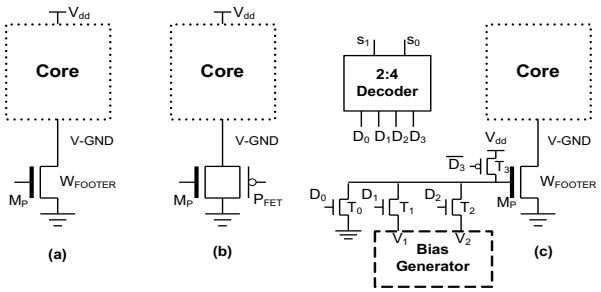


Fig. 1. Various Power-Gating Architectures

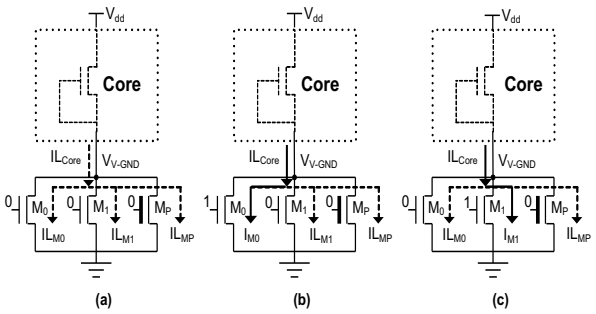


Fig. 2. Proposed scheme: (a) Snore Mode, (b) Dream Mode, (c) Sleep Mode

smaller than the architecture proposed in [18] and offers greater power savings for similar wake-up times. The proposed architecture is also more tolerant to process variations than [18], thus its operation is more predictable. Finally, a reconfigurable version of the proposed architecture is also proposed, which can tolerate even greater process variations, enabling thus the utilization of the proposed architecture for newer technologies.

II. Background

The classical power-switch architecture is shown in Fig. 1(a). It consists of a footer transistor M_p connected between the core and the ground rail. When the footer is "on", the core operates in the normal functional mode. When it is "off" (i.e., during idle mode), the virtual ground rail (*V-GND*) charges to a voltage level close to the power supply and it suppresses the leakage power of the transistors of the circuit due to the body effect. In order to minimize the impact on circuit performance during normal operation, the footer transistor has to be large enough to constitute a strong driver. In practice, instead of using a large footer transistor, many small transistors connected in parallel are used.

In order to restore the virtual ground rail to its nominal value when the circuit transitions from the power-off mode to the active mode, the parasitic capacitance at the *V-GND* node has to be completely discharged through the power switches. However, power switches are relatively small high-V_t transistors and thus the wake-up time is usually long relative to circuit clock rate. This limits the applicability of this technique to idle periods that are longer than the wake-up time of the circuit. To overcome this limitation, [9] proposed the use of an intermediate power-off mode, where the virtual ground node is left charged to an intermediate voltage level. This is achieved through the use of a pMOS device connected in parallel with the nMOS footer M_p, as shown in Fig 1(b). The pMOS is turned-on in the intermediate power-off mode and the virtual ground potential is adjusted to the threshold voltage of the pFET. Then the virtual ground node requires less time to completely discharge, although at the expense of less leakage reduction compared to the complete power-off mode. Similar architectures were proposed in [2], [13]. The authors of [18] proposed a power-switch structure with two intermediate power-off modes and they showed that for various applications on a 64-bit Alpha processor, the use of two intermediate power gating modes offers further reduction in leakage of about 17% compared to single-mode gating.

The architecture proposed in [18] is presented in Fig. 1(c). It consists of the power switch M_p, a decoder, the bias generator, which is an analog circuit, and the transistors T_0-T_3. Using this structure, the gate voltage of the power switch M_p is regulated to four different voltage levels 0, V_1, V_2 and V_{dd}. Transistor T_0 adjusts the gate voltage of M_p at the ground level, and thus it completely turns off the power switch. This is the "Snore" mode where the leakage power is minimized and the wake-up time is very high (M_p has to completely discharge the virtual ground rail when it is turned on). The next two modes, namely "Dream" and "Sleep", are determined by the two sub-threshold gate voltages V_1, V_2, ($V_1<V_2<V_{TH\text{-}SW}$ where $V_{TH\text{-}SW}$ is the threshold voltage of the power switch transistor M_P) generated by the bias generator and applied to the gate of the power switch through transistors T_1, T_2 respectively. In both cases, the virtual ground is charged to a potential that is lower than V_{dd} and thus the wake-up time drops, at the expense however of increased leakage power consumed. Finally, by turning on transistor T_3 the gate voltage level is set to V_{dd} and the core is put into "Active" mode. The authors of [18] have reported comprehensive studies to evaluate the leakage-saving advantages of this architecture compared to the baseline power-gating architecture. In addition, they showed that by using the intermediate modes, the ground bounce in neighboring circuits which is an inherent problem of all MTCMOS designs [3], [8] can be also reduced.

A major drawback of the structure proposed in [18] is the sensitivity of the bias generator in Fig. 1(c) to process variations. The correct operation of the structure depends on the precise generation of the two sub-threshold voltages V_1 and V_2 that are very close one to the other. However, generation of such fine-tuned voltage levels requires the fabrication of a very accurate bias-generator circuit, which is very difficult to achieve under process variations. Moreover, the generation of more than two sub-threshold voltages requires an even more accurate bias generator. Therefore, this architecture cannot be easily scaled to support more than two intermediate power-off modes. Moreover, the bias generator is an analog circuit and consumes static power, which reduces the overall efficiency of the structure and introduces complexity for testing and fault diagnosis.

In this work, we propose a new multi-mode power switch architecture with the following major advantages:

1. It is very simple, all-digital, and minimally sized.
2. It provides more than two intermediate power-off modes.
3. It consumes low static power.
4. It has high tolerance to manufacturing process variations.

In addition, by inserting a small amount of redundancy, the proposed scheme can be easily modified to a reconfigurable structure that is robust to high process variations.

III. Multi-Mode Power Gating Architecture

In this section, we first present the proposed design for the same power-off modes with [18]. Later in the section, we explain the extension to more power-off modes.

A. Proposed Architecture

Fig. 2 presents the proposed design. It consists of the main power switch transistor M_p and two small transistors M_0 and M_1, each corresponding to an intermediate power-off mode (M_0 corresponds to the dream mode and M_1 corresponds to the sleep mode). Transistor M_p is a high-V_t transistor and it remains on only during the active mode. Transistors M_0 and M_1 are small low-V_t transistors that are turned on only during the corresponding power-off mode. The various modes of operation are as follows:

Active mode: Transistor M_P is on. Transistors M_0 , M_1 are off.

Snore mode: Transistors M_P, M_0 and M_1 are off as shown in Fig. 2(a). In this case, the leakage current of the core, IL_{core}, is equal to the aggregate leakage current flowing through transistors M_0, M_1, M_P ($IL_{core} = IL_{M0}+IL_{M1}+IL_{MP}$), which is very small (note that M_0, M_1 are small transistors and M_P is a high-V_t transistor). Thus the voltage level at virtual ground rail is close to V_{dd} (i.e. $V_{V\text{-}GND} \approx V_{dd}$) and the circuit consumes a negligible amount of energy, but the wake-up time is high.

Dream mode: Transistor M_0 is on and transistors M_P and M_1 are off as shown in Fig. 2(b). In this case, the current flowing through transistor M_0 (and thus the aggregate current flowing through M_0, M_1 and M_P) increases because M_0 is on ($I_{M0} > IL_{M0}$). The exact value of I_{M0} depends on the size of transistor M_0, and it sets the virtual ground node at a voltage level which is lower than V_{dd} (i.e., $V_{V\text{-}GND} < V_{dd}$). Thus the static power consumed by the core is higher compared to the snore mode, but the wake-up time is less.

Sleep mode: Transistor M_1 is on, and M_P, M_0 are off as shown in Fig. 2(c). Provided that transistor M_1 has larger aspect ratio than M_0 ($W_{M1}/L_{M1} > W_{M0}/L_{M0}$), the aggregate current flowing through M_0, M_1, and M_P increases even more when M_1 is on (note that $I_{M1}>I_{M0}$). Consequently, the voltage level at the virtual ground node is further reduced compared to the dream mode and thus the wake-up time decreases at the expense of increased static power consumption.

B. Design Method

The correct operation of the proposed design depends on the correct sizing of transistors M_0, and M_1. In the sequel, we will provide an analytical calculation for the aspect ratio of each of these transistors. For simplicity, as in [18], we model the core with a single equivalent NMOS transistor, and we consider only the sub-threshold leakage current.

Let us consider the dream mode shown in Fig. 2(b). In this mode transistor M_0 is on. Assuming that $V_{V\text{-}GND} < V_{dd} - V_{THC}$ (V_{THC} is the threshold voltage of the low-V_t transistors M_0, M_1) we can deduce that M_0 operates in the linear region when it is on. Therefore, the current flowing through M_0 is given by the following equation:

$$I_{M_0} = \mu_n C_{ox}(W_{M0}/L_{M0})((V_{dd} - V_{THC})V_{V-GND} - V_{V-GND}^2/2) \quad (1)$$

where W_{M0}, L_{M0} are the width, length, respectively, of transistor M_0. The sub-threshold leakage current of the core IL_{core} is calculated using (2):

$$IL_{core} = I_0^{core}[1 - e^{(V_{dd}-V_{V-GND})/v_t}] \cdot e^{(V_{V-GND}-V_{dd}-V_{THC}-V_{off}')/nv_t} \quad (2)$$

where I_0^{core} is a constant, which depends on the width and length of the equivalent transistor corresponding to the core and on process parameters. Likewise, the leakage current of the power switch M_P is expressed in (3)

$$IL_{M_P} = I_0^{M_P}[1 - e^{-V_{V-GND}/v_t}] \cdot e^{-V_{TH-SW}-V_{off}'/(nv_t)} \quad (3)$$

The leakage current of transistor M_1 is calculated in the same way. Based on Kirchhoff's current law, we can obtain the following equation:

$$IL_{core} = I_{M0} + IL_{M1} + IL_{MP}$$

Note that $W_{MP}/L_{MP} >> W_{M1}/L_{M1}$ thus $IL_{M1} << IL_{MP}$. Therefore, the equation above is simplified to (4):

$$IL_{core} = I_{M0} + IL_{MP} \quad (4)$$

Substituting (1), (2), (3) into (4), we get

$$\frac{W_{M_0}}{L_{M_0}} = \frac{2(IL_{core} - IL_{M_P})}{\mu_n C_{ox}(2(V_{dd} - V_{THC})V_{V-GND} - V_{V-GND}^2)} \quad (5)$$

By using Equation (5) we can adjust the voltage level *V-GND* to any value in the range (0, $V_{dd}-V_{THC}$) and we can calculate the aspect ratio of transistor M_0. The wake-up time is calculated as follows:

$$T_{wake\text{-}up} = C_{total} \cdot R_{eq}$$

where C_{total} is the parasitic capacitance of the virtual ground and R_{eq} is the equivalent resistance of transistor M_0 when it discharges the virtual ground node (R_{eq} is the average resistance of M_0 for the conducting time duration [16]). Thus, the wake-up time is provided by the following equation:

$$T_{wake-up} = C_{total} \cdot R_{eq} = C_{total} \cdot (1/(t_2 - t_1)) \int_{t1}^{t2} (V_{V-GND}(t)/I_D(t))dt$$

or equivalently

$$T_{wake-up} = C_{total} \times \frac{1}{-V_{V-GND}} \int_{V-GND}^{0} \frac{V}{I_D(V)} dV \quad (6)$$

Since M_0 is in the linear region during the wake-up operation ($V_{GS}=V_{dd}$), equation (6) is written as follows:

$$T_{wake-up} = C_{total} \times \frac{-2L}{\mu_n C_{ox} W V_{V-GND}} \int_{V-GND}^{0} \frac{1}{2(V_{dd} - V_{THC}) - V} dV$$

and thus

$$T_{wake-up} = \frac{2C_{total}L(\ln(2(V_{dd} - V_{THC})) - \ln(|V_{V-GND} - 2(V_{dd} - V_{THC})|))}{\mu_n C_{ox} W V_{V-GND}} \quad (7).$$

The same analysis can be used for calculating the size and the wake-up time of transistor M_1; the "Sleep" mode case is presented in Fig. 2(c). Equations (5) and (7) can be used for calculating the transistor size required to set the virtual ground rail at any particular voltage level in the range (0, $V_{dd}-V_{THC}$). Thus the extension of the design to more power-off modes is straightforward. Note that in the above analysis, we considered only the sub-threshold leakage current for every device that is turned-off. For a more accurate estimation, however, the total leakage current of the core and the power switch M_P must be used in Equation (5). Finally, note that as in the previous architectures [9], [18], the wake-up time also depends on the internal state of the core since leakage current is input-pattern dependent. Average-case analysis as well as worst-case analysis of the core can be used to calculate the leakage current during idle mode. Worst-case analysis assumes that each cell receives the most leaky logic combination at its inputs. Even though this is a pessimistic scenario, it guarantees the correct operation of the power-gating structure independent of the core state.

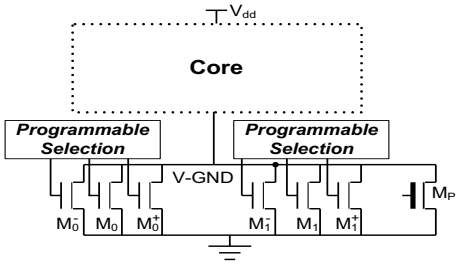


Fig. 3. Reconfigurable architecture

C. Reconfigurable Architecture

As shown in the next section, the proposed architecture exhibits considerable tolerance to process variations. However, for cases where even higher tolerance to process variations is required, we propose the reconfigurable structure shown in Fig. 3. Each of the M_0, M_1 transistors is replaced by a triplet of transistors $\left(M_0^-, M_0, M_0^+\right)$, $\left(M_1^-, M_1, M_1^+\right)$ respectively. The aspect ratios of M_0 and M_1 are calculated analytically as in the previous subsection. The aspect ratios of $\left(M_0^-, M_0^+\right)$, $\left(M_1^-, M_1^+\right)$ are selected to be close to the aspect ratio of M_0, M_1, respectively. Specifically,

$$W_{M_0^-}/L_{M_0^-} = (1+\alpha/100)W_{M_0}/L_{M_0}, W_{M_0^-}/L_{M_0^-} = (1-\alpha/100)W_{M_0}/L_{M_0}$$

where $a \in (0\%, 100\%)$. The parameter a is selected in such a way as to reflect the process variations of the particular technology used. Specifically, it induces an artificial variation in the aspect ratio of these transistors in order to counterbalance some of the process variations. Note that process variations will shift the desired aspect ratio of transistors M_0, M_1 a little above or below the nominal value calculated by Equation (5). The length of this shift depends on the magnitude of the process variations. The use of a pair of transistors in each triplet with their aspect ratios already shifted by a% above and below the nominal value increases the probability that one of the transistors of each triplet provides the required voltage at the virtual-ground node in the presence of process variations. For new technologies, which tend to suffer from high process variations, a large value of a must be used whereas for older or mature technologies, a smaller value of a will suffice. The selection of the proper transistor of each triplet can be done after the manufacturing process using a programmable structure, e.g., fuses commonly used for built-in memory self-repair. Except for the selected transistor, the other transistors in each triplet will be permanently off.

The reconfigurable architecture offers the advantage of low cost due to its simplicity and the small size of transistors M_0, M_1. Moreover, even for higher tolerance to process variations, the reconfigurable structure can be easily extended to accommodate groups of n pairs of transistors per mode with their aspect ratios shifted by a_1%, a_2%, …, a_n% above and below the nominal value ($a_1 < a_2 < \ldots < a_n$).

IV. Evaluation & Comparisons

For evaluating the proposed architecture, we considered a logic core consisting of 9 Million transistors. Even though this logic core is not a real circuit, it is representative of a realistic industrial circuit in terms of static power consumption during DC operation in power-off mode. The size of the logic core is not crucial for simulation, since the desired power saving and wake-up time can always be achieved by adjusting the sizes of power switches. We used the 45 nm predictive technology [20] for 1.1 volts power supply. The leakage power consumption of the core in idle mode with no power gating is equal to 10.001mW.

Table I. Transistor sizes in two structures

Proposed structure (W/L)		Structure in [18] (W/L)	
M_0	250 nm / 45 nm	T_0	120 nm / 45 nm
		T_1	120 nm / 45 nm
M_1	480 nm / 45 nm	T_2	120 nm / 45 nm
		Bias Generator	7920 nm / 45 nm
M_p	43.2×10^6nm/45nm		

We implemented both the architecture proposed in [18] (see Fig. 1(c)) and the proposed architecture (see Fig. 2) for the aforementioned logic core. As it was suggested in [18], the width of the main power switch (transistor denoted as M_P) was set equal to 12% of the total width of the nmos transistors in the logic core. We have to note that transistor T_3 in the architecture of [18] shown in Fig. 1(c) has to be a strong pull-up driver for quickly charging the large gate capacitance of the power switch M_P during activation of the core after any power-off mode. A similar strong driver is used in the classical power-gating architecture, and this is the case for the proposed design as well. Finally, in order to provide fair comparison between the proposed architecture and that of [18], the transistor sizes in both architectures were selected in such a way as (a) to be of minimum size required, and (b) to provide similar wake up times, in both architectures. Moreover, in the proposed scheme, the sizes of transistors M_0 and M_1 have been selected in such a way as to provide the same voltage level at the virtual ground node with the scheme proposed in [18] at each power-off mode. Thus, the logic core dissipates the same amount of static power in both architectures at each power-off mode.

At first, we compare both architectures in terms of area overhead measured as aggregate transistor sizes. The sizes of the main transistors in two structures are listed in Table I (the width of the bias generator is reported as the summation of the width of all its transistors). For the comparison we excluded the main switch transistor M_P, the decoder, the large transistor T_3 in the case of [18] and the large buffer driving M_p transistor in the proposed architecture which are similar in both architectures. The rest of the circuitry (taking also into account power-off signal drivers) occupies in the proposed architecture almost one fifth (1/4.8) of the area of the architecture in [18]. Even though this is an estimate based on transistor sizes, it is apparent from Fig. 1(c) and Fig. 2 that the proposed architecture is also much simpler than [18] and thus it requires less routing overhead.

Fig. 4 presents the leakage power consumed during the various sleep modes by the core and the power-gating logic in the two architectures. The x-axis presents the three power-off modes, sleep, dream and snore, which require 3, 5 and 8

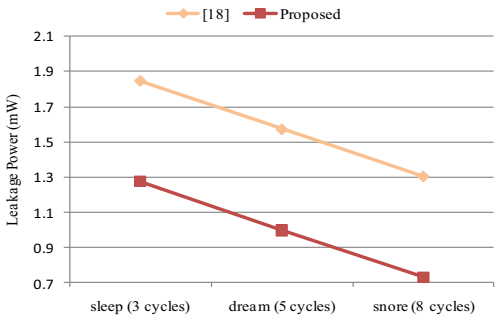
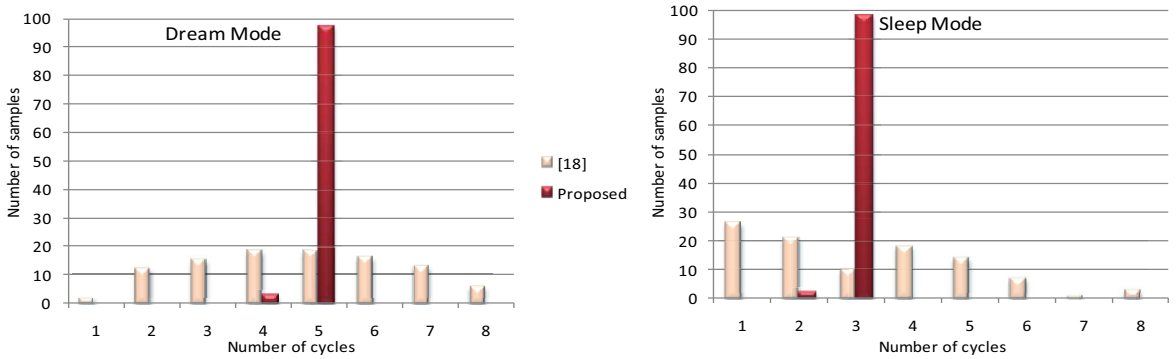


Fig. 4. Leakage power comparison between [18] and proposed architecture



g. 5. Distribution of cycles needed for wake-up from dream and sleep mode

wake-up cycles respectively (the clock frequency is taken to be 1GHz). The y-axis shows the leakage power consumed in each case. It is obvious that both architectures provide a trade-off between the wake-up time and the static power reduction. However, the proposed scheme is more effective than [18] in reducing the total static power for the same number of wake-up cycles. As mentioned earlier, the logic core consumes the same static power at each power-off mode in both schemes because the voltage level at the virtual ground node is the same for both architectures.

In the next experiment, we study the effect of process variations on the two architectures using Monte-Carlo simulations with sample size 100. Note that even though 30 Monte Carlo runs are adequate according to [12], we ran 100 simulations in order to ensure higher level of confidence in the simulation results. In Monte-Carlo simulations for both schemes, variations are set to 3.5% for transistor width, 7% for transistor length and T_{ox}, and 12% for threshold voltage, based on data available from ITRS [6].

Fig. 5 presents the distribution of the number of wake-up cycles for each scheme at the dream and sleep power-off modes for the 100 samples. In both charts the x-axis presents the number of cycles and the y-axis presents the corresponding percentage of samples. Note that the number of cycles needed for wake-up from the dream and sleep mode with no process variations (i.e. the number of cycles calculated during the design phase) is 5 for the dream mode and 3 for the sleep mode respectively. It is obvious that the design proposed in [18] is affected considerably by the process variations, in both intermediate power-off modes (in the "snore" mode both schemes exhibit similar tolerance). Specifically, less that 20% of the samples operate as designed in both dream and sleep modes. In contrast, the proposed scheme is very tolerant as more than 95% of the samples are not affected by the process variations.

Fig. 6 shows the static power consumption of the 100 samples for the dream and the sleep mode. As in the case of wake-up cycles, the variation for the architecture of [18] is very high for both intermediate power-off modes, while the variation for the proposed scheme is negligible. The high variation for the design from [18] can be attributed to the bias generator which fails to generate accurate bias voltages in the presence of process variations. In contrast, the proposed scheme is much more tolerant to process variations.

In the next experiment we considered our design for four intermediate power-off modes, namely dream, sleep, slumber and nap (snore mode is the complete power-off mode). In this case, we considered a 2 GHz clock frequency. Fig. 7 presents the tradeoff between wake-up time and power consumption for the proposed design. The left y-axis in Fig. 7 presents the number of wake-up cycles, while the right y-axis presents the power consumption for each power-off mode. We see that the tradeoff between wake-up time and power reduction can be effectivelly extented to more power-off modes by using the proposed scheme. This is particularly useful in cases where the wake-up time from the complete power-off mode is large enough to allow for finer segmentation into power-off modes and thus better exploitation of the short periods of inactivity.

Finally, we highlight the effectiveness of the reconfigurable architecture for higher levels of process variations. We ran 100 Monte Carlo simulations for both the reconfigurable and non-reconfigurable architecture for five power-off modes, assuming 3.5% for transistor width, 10% for transistor length, 3% for T_{ox}, and 30% for threshold voltage (these values are obtained from a current VDSM technology in industry). In

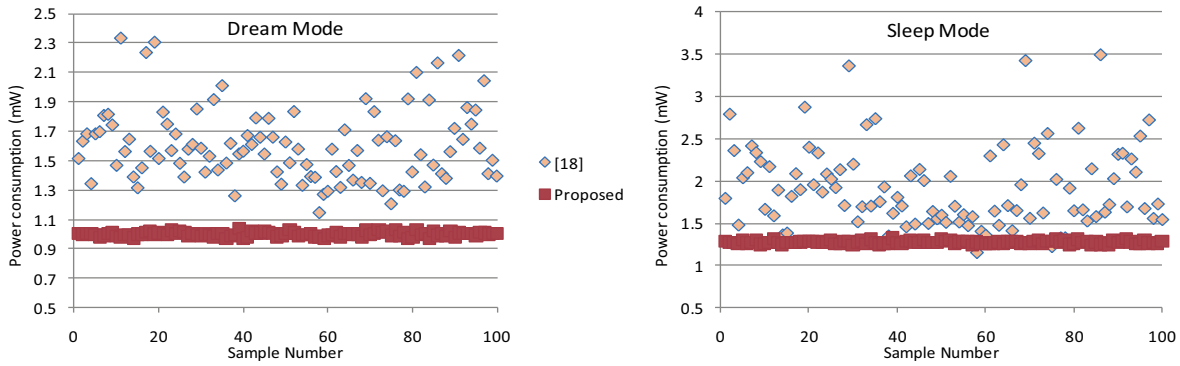


Fig. 6. Distribution of power consumption for dream and sleep mode

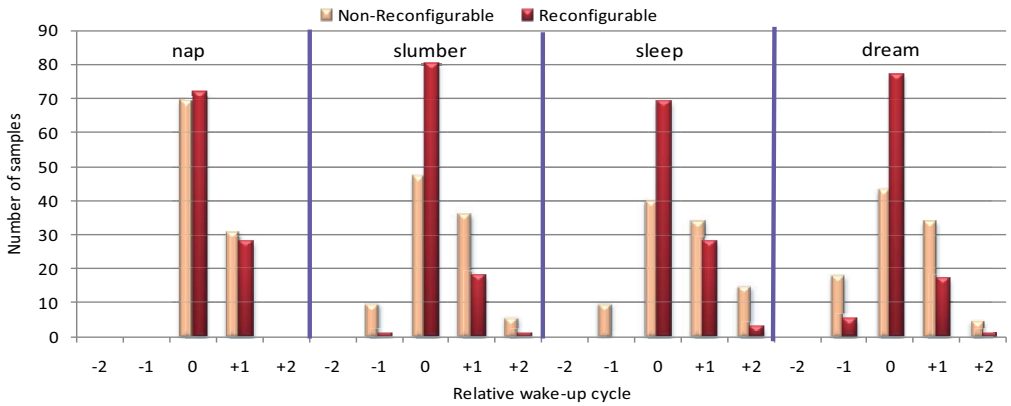


Fig. 8. Distribution of the wake-up cycles for the proposed design, assuming four intermediate modes.

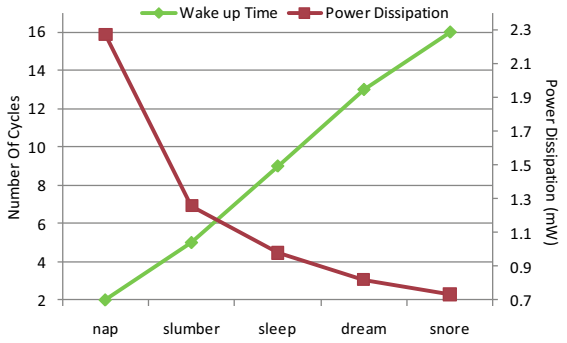


Fig. 7. Tradeoff between wake-up time and power consumption with 5 modes

the reconfigurable architecture we used triplets of transistors as in Fig. 3, with a = 5% for dream mode and 10% for the rest modes. From each triplet, we selected the transistor that best matched the nominal case in terms of the number of wake-up cycles. Fig. 8 shows the distribution of the samples in respect with the wake-up cycles for the four intermediate power-off modes, for both reconfigurable and non-reconfigurable architecture. For each mode we present the number of samples which need the same number of wake-up cycles with the case of no-process variations (denoted as relative wake-up cycle 0 in the chart). Additionally, we present the number of samples that need one or two more cycles (denoted as relative wake-up cycle '+1', '+2') as well as the number of samples that need one or two less cycles (denoted as relative wake-up cycle '-1', '-2'). It is obvious that in all cases the reconfigurable architecture offers higher percentage of samples operating as designed, especially in the slumber, sleep and dream modes. Note that even higher tolerance can be achieved by using larger groups of transistors per mode. Therefore, we can cocnclude that the reconfigurable architecture is much less affected by high levels of process variation than the non-reconfigurable one.

V. Conclusions

We have described a new power-gating scheme that provides multiple power-off modes. The proposed design offers the advantage of simplicity and it requires minimum design effort. Extensive simulation results show that, in contrast to a recent power-gating method for multiple power-off modes, the proposed design is robust to process variations, and it is scalable to more than two power-off modes. Moreover, it requires significantly less area and consumes much less power than the previous design. Finally, a reconfigurable version of this method can be used to increase the manufacturability and robustness of the proposed design in technologies with larger process variations.

References

[1] ARM 1176JZ(F)-S documentation, http://www.arm.com/products/CPUs/ARM1176.html.

[2] M. Chowdhury, J. Gjanci and P. Khaled, "Innovative Power Gating for Leakage Reduction", in *Proc. ISCAS*, 2008, pp. 1568-1571.

[3] S. Henzler et al., "Sleep Transistor Circuits for Fine-grained Power Switch-off with Short Power-down Times", in *Proc. International Solid-State Circuits Conference*, 2005, pp. 302–303.

[4] S. Ingunji, "Case Study of Low Power MTCMOS based ARM926 SoC: Design, Analysis and Test Challenges", in *Proc. International Test Conference*, 2007.

[5] Intel Corp, "Intel XScale Core Developer's Manual", 2003, http://developer.intel.com/design/intelxscale/.

[6] Semiconductor Industry Association, Int. Tech. of Semiconductors, 2007 (http://www.itrs.net/Links/2007ITRS/Home2007.htm).

[7] J. Kao, S. Narendra and A. Chandrakasan, "MTCMOS Hierarchical Sizing Based on Mutual Exclusive Discharge Patterns", in *Proc. Design Automation Conference*, 1998, pp. 495-500.

[8] S. Kim, S. Kosonocky and D. Knebel, "Understanding and Minimizing Ground Bounce during Mode Transition of Power Gating Structures", in *Proc. IEEE ISPLED*, 2003, pp. 22-25.

[9] S. Kim, S. Kosonocky, D. Knebel and K. Stawiatsz, "Experimental measurement of a novel power gating structure with intermediate power saving mode", in *Proc. SLPED*, 2004, pp. 20-25.

[10] S. Kosonocky, et al., "Enhanced Multithreshold (MTCMOS) Circuits with Variable Well Bias", in *Proc. IEEE ISPLED*, 2001, pp. 165-169.

[11] D. Lackey, et al., "Managing power and performance for system-on-chip designs using voltage islands", in *Proc. IEEE International conference on ICCAD*, 2002, pp. 195–202.

[12] R.C. Lopez, et al., "Reuse-based methodologies and tools in the design of analog and mixed-signal integrated circuits", Springer, 2006.

[13] E. Pakbaznia and M. Pedram, "Design and Application of Multimodal Power Gating Structures", in *Proc. ISQED*, 2009, pp. 120-126.

[14] R. Puri, et al., "Pushing ASIC performance in a power envelope", in *Proc. IEEE ACM Design Automation Conference*, 2003, pp. 788–793.

[15] R. Puri, D. Kung and L. Stok, "Minimizing power with flexible voltage islands" in *Proc. IEEE ISCAS*, 2005, pp. 21–24.

[16] J. M. Rabaey, A. Chandrakasan, B. Nicolic, "Digital Integrated Circuits A Design Perspective", Prentice Hall, 2003.

[17] K. Roy, S. Mukhopadhyay and H. Mahmoodi-Meimand, "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicron CMOS Circuits", *Proc. IEEE*, 2003, vol. 91, pp. 305-327.

[18] H. Singh, K. Agarwal, D. Sylvester and K. Nowka, "Enhanced Leakage Reduction Techniques Using Intermediate Strength Power Gating", *IEEE Trans. on VLSI*, 2007, vol. 15, no. 11, pp. 1215-1224.

[19] Transmeta Corporation, "Crusoe processor documentation", 2002 http://www.transmeta.com.

[20] W. Zhao and Y. Cao, "New Generation of Predictive Technology Model for sub-45nm Early Design Exploration," *IEEE Trans. on Electron Devices*, 2006, vol. 53, no. 11, pp. 2816-2823.