A Multi-Mode Power Gating Structure for Low-Voltage Deep-Submicron CMOS ICs

Suhwan Kim, *Member, IEEE*, Stephen V. Kosonocky, *Member, IEEE*, Daniel R. Knebel, Kevin Stawiasz, and Marios C. Papaefthymiou, *Senior Member, IEEE*

***Abstract*—Most existing power gating structures provide only one power-saving mode. We propose a novel power gating structure that supports both a cutoff mode and an intermediate power-saving and data-retaining mode. Experiments with test structures fabricated in 0.13-μm CMOS bulk technology show that our power gating structure yields an expanded design space with more power-performance tradeoff alternatives.**

***Index Terms*—Deep-submicrometer CMOS, ground bounce noise, low voltage, multi-threshold CMOS (MTCMOS).**

I. Introduction

THE trend toward high-performance portable system-on-a-chip (SoC) designs for communication computing equipment has made power dissipation a critical constraint. Supply voltage scaling is known as the most effective way to reduce power dissipation, especially in CMOS digital circuits; but a reduced supply voltage increases circuit delay, making it necessary to decrease threshold voltages in order to maintain performance. Unfortunately, this leads to a dramatic increase in leakage current, due to the exponential nature of leakage current in the subthreshold regime of the transistor [1].

The use of a multi-threshold CMOS (MTCMOS) circuit, called a power gating structure, is a well-known technique for reducing leakage power in standby mode, while still permitting high-speed operation in active mode [2]. It utilizes low-leakage, high-threshold devices as sleep transistors, which disconnect idle blocks from the power supply, from the ground, or from both, to reduce the leakage in standby mode. This is achieved by using a pMOS and an nMOS transistor in series with the transistors of each logic circuit to create a virtual power supply and a virtual ground.

We propose a novel power gating structure in which only nMOS transistors are used in series with the transistors of each logic circuit, so as to reduce the on-resistance in active mode. To support an intermediate power-saving and state-retaining mode at a low supply voltage, a single pMOS is added in parallel with the nMOS. This intermediate mode minimizes the ground bounce noise induced by power mode transitions of the power gating structures, yielding an expanded space for power-performance tradeoffs that supports three different power modes. We call these modes RUN/IDLE, PARK, and COLD. We have evaluated our new power gating structure by designing and fabricating three differently configured macros on a test chip in 0.13-μm CMOS bulk technology, using single-threshold devices for both logic and sleep transistors. Measured results from the macros show the potential benefits of our new approach.

II. Background

Recently, many vendor products in the low power embedded space provide power-gating support in the form of "sleep" modes, typically software control. One of multiple processor cores, in such as system, runs at the maximum operating frequency and the other processor cores can be power-gated off when the operating system detects a long idle loop [3]. The aggressive power-saving strategy above, however, has the following potential problems.

First of all, turning off the nMOS sleep transistor of a gating structure during sleep periods results in charging the virtual ground (VGND) node of the power gating structure being charged up to a steady-state voltage close to VDD. As a consequence, the data in storage elements is completely lost. A data-recovery process then becomes necessary, significantly degrading system performance. Secondly, the instantaneous discharge current through the sleep transistor, which is operating in its saturation region, creates current surges during the change from sleep to active. Because of the self-inductance of the off-chip bonding wires and the inherent parasitic inductance of the on-chip power rails, these current surges cause voltage fluctuations in the on-chip power distribution network [4], [5]. Ground bounce is a phenomenon that has often been associated with input/output buffers, internal digital circuitry, and clock gating. As the supply voltage in deep-submicrometer technologies has been reduced, the noise margins of CMOS devices have decreased, and minimizing ground bounce has become critical. The most commonly used capacitor is the MOS capacitor, and on-chip MOS decoupling capacitors are extensively adopted to minimize ground bounce noise. However, these capacitors also introduce a large leakage power in scaled technology. For example, a decapacitor leakage power of 26 W (about 12% of total power) has been reported [6].

A virtual power/ground rail clamp (VRC) scheme dynamically reduces the virtual supply voltage across a circuit using two diodes [7], [8]. During standby mode, pMOS and nMOS switches (MP and MN) are turned off by asserting low (high)

Manuscript received August 30, 2006; revised December 1, 2006. The work of S. Kim was supported by the Nano-Systems Institute (NSI-NCRC) program sponsored by the Korea Science and Engineering Foundation (KOSEF). This paper was recommended by Associate Editor V. Kursun.

S. Kim is with the Department of Electrical Engineering and Computer Science, Seoul National University, Seoul 151-744, Korea (e-mail: suhwan@snu.ac.kr).

S. V. Kosonocky, D. R. Kneble, and K. Stawiasz are with the IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA.

M. C. Papaefthymiou is with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109 USA.

Digital Object Identifier 10.1109/TCSII.2007.894428

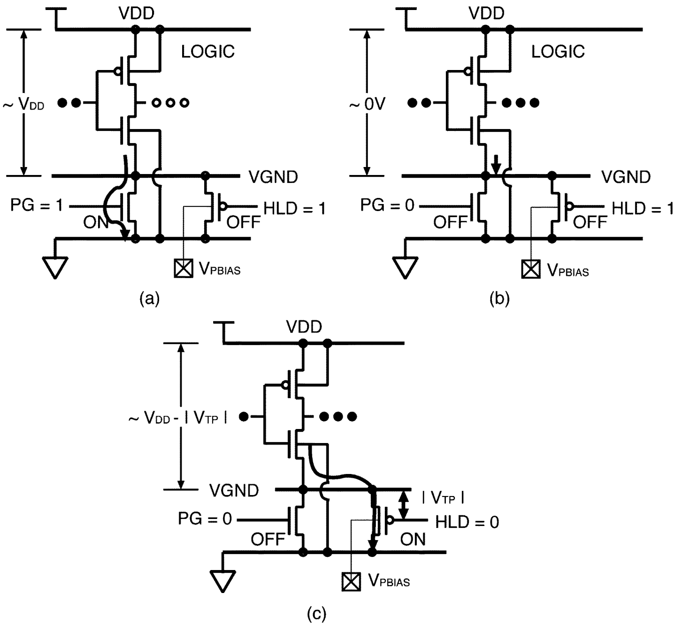

Fig. 1. Tri-mode power gating structure showing the dominant current flow in (a) RUN/IDLE (normal) mode, (b) COLD (cutoff) mode, and (c) PARK (intermediate power-saving and data-retention) mode.

at CS (/CS) and a pair of diodes (DN and DP) clamp the supply voltage to a lower value [9], [10]. VRC allows state retention in the storage elements, thus eliminating the need for state restoration procedures, without requiring high-efficiency, low-noise and low-power regulators or multiple supply voltages (or gate bias voltage generators) [11], [12]. However, in the case of the latest deep-submicron technologies, whose supply voltage is already in the 1-V range in standby mode, this VRC structure may not be able to retain state with VDD-$(2 \times |V_{\mathrm{DN,DP}}|)$.

III. Tri-Mode Power Gating Structure

To address the problems outlined in Section II, we propose adding a single pMOS (not a diode-connected PMOS) in parallel to a power gating structure with nMOS transistors, which leads to a power gating structure that can support an additional intermediate power-saving and data-retaining mode, as well as a power cutoff mode. This new intermediate mode enables data retention and reduces leakage as well as the magnitude of power supply voltage fluctuations during power-mode transitions.

Fig. 1(a) shows our new power gating structure in RUN/IDLE mode. In this mode, PG is asserted high to force the nMOS transistor in the power gating structure into a low-resistance state, while HLD is set high. The nMOS is used to short the virtual ground VGND of the logic circuit to the real ground potential GND, allowing the full supply voltage VDD to be applied across the circuit, and thus permitting high-speed operation.

Fig. 1(b) shows the circuit in COLD mode, which does not retain state. PG is held low while HDL is high, the current path to GND is cut, and the voltage across the logic circuit collapses, suppressing both gate and subthreshold leakage currents. As the VGND node is close to VDD, the sources of the pulldown network transistors are reverse-biased with respect to their body connections which are held at GND. Drain-induced barrier lowering (DIBL) causes an increase in the threshold voltage and reduces the subthreshold leakage currents in the transistors of the pulldown network [13], [14]. In the state-retention or PARK mode, shown in Fig. 1(c), both PG and HLD are asserted low. Consequently, the nMOS device is turned off and the pMOS operates as a source-follower. In PARK mode, the virtual ground rail VGND is held at a voltage $|V_{\mathrm{TP}}|$ above that of the ground rail, where V_{TP} is the threshold voltage of the PMOS. The voltage across the logic circuit becomes VDD-$|V_{\mathrm{TP}}|$, causing a reduction in gate leakage and subthreshold leakage, since these leakages are dependent on the voltage applied to the devices. In this mode, the VGND voltage level is limited by $|V_{\mathrm{TP}}|$. As a result, state is retained, and the ground bounce induced by power mode transitions is smaller than it is in COLD mode. PARK mode can also be used as an intermediate step to reduce the ground bounce induced by the transition from COLD to RUN/IDLE.

IV. Test Chip Design

To demonstrate the effectiveness of the proposed tri-mode power gating structure and intermediate power-saving mode, three differently configured macros were designed and fabricated, using 0.13-μm CMOS bulk technology. To minimize process variation, the three macros were implemented on the same multiple project wafer (MPW). Each macro includes nine identical design-under-test (DUT) modules. The basic components of each DUT module are two linear-feedback shift registers (LFSRs), one 32-bit carry lookahead adder (CLA), and one multiple-input signature register (MISR). The LFSRs generates a sequence of pseudo-random patterns and feeds them to the CLA, while the MISR validates the correct operation of the DUT module. The ground nodes of the CLA and the output register in the first and second macros are connected to GND through the new power gating structure.

Sizing the sleep transistors is one of the major challenges in power gating. If we overestimate their size then we end up wasting silicon area; but if we make them too small, the required performance may not be achieved due to increased resistance to ground. The sleep transistor of each DUT module in the first macro is sized to handle the worst-case current through that module, and the transistor in each DUT module in the second macro is sized for average current. In our test chip, the size of the sleep transistor in each DUT module in the first macro is 2.6% of the total nMOS and pMOS size (7.0% of the total nMOS size) in the CLA and the output register. Similarly, the size of the sleep transistor in each DUT module in the second macro is 1.0% of the total nMOS and pMOS size (2.7% of the total nMOS size) in the CLA and the output register. And the data-retention transistor in the first and second macros is sized at 0.1% of the total nMOS and pMOS size (0.27% of the total nMOS size) in the CLA and the output register. The power-gating structure requires a layout area overhead of less than 4.5%. The third macro is designed to compare our power gating structure with a circuit that has no power gating. Accordingly, the ground nodes of the CLA and the output register are directly connected to GND. To exclude the leakage power consumption of the decoupling capacitors from the leakage of the logic circuitry, on-chip decoupling capacitors are not included in our test chip.

The block diagram and layout in Fig. 2 show the supply and ground rail distribution to the 3 × 3 array of DUT modules in

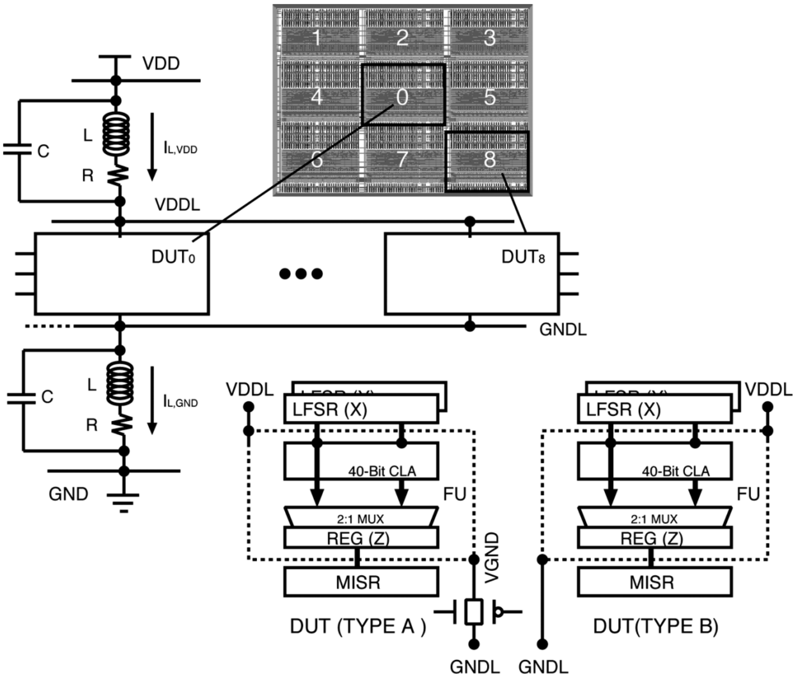

Fig. 2. Block diagram and layout of one of the macros implemented in our test chip.

one of the macros. In the first and second macros, the power gate structures are incorporated individually in each DUT module of type A, thus creating independent power islands with power-mode controls that allow one island to remain running while the others are put into PARK or COLD mode. This arrangement also allows us to observe how voltage disturbances due to the *RLC* networks formed by the VDD and GND rail distribution are coupled to each of the other DUT modules.

V. Experimental Results

Fig. 3 shows the hardware setup used to test and measure our unpackaged chips. When the power gating structure is in RUN or IDLE mode, the DUT module may either be in RUN mode (clocked at the highest operating frequency that does not result in any failure signature) or in IDLE mode (shutting off the clock to the latches of the DUT module).

The following four independent test and measurement scenarios were used to quantify the effectiveness of our power gating implementation, and to compare performance and power consumption in PARK and COLD modes. 1) Measuring the maximum operating frequency of the DUT module and their active power consumption at that frequency. For this test, DUT_0 is in RUN mode, and DUT_1 through $\mathrm{DUT}_8(\mathrm{DUT}_{1-8})$ are in IDLE mode. 2) Comparing the leakage current with all DUT modules in IDLE mode. This test is repeated for the PARK and COLD modes. 3) Measuring the off-chip ground voltage with DUT_0 in IDLE mode and the remaining DUT modules switching from COLD to IDLE modes. This test is repeated with the DUT modules switching from COLD to IDLE through PARK modes. In these tests, DUT_{1-8} create ground bounce noise due to switching of the power gating structures. 4) Measuring the effect of on-chip ground bounce on the performance of nearby logic by putting DUT_0 into RUN mode and switching the remaining DUT modules from COLD direct to IDLE mode and from COLD to IDLE through PARK mode.

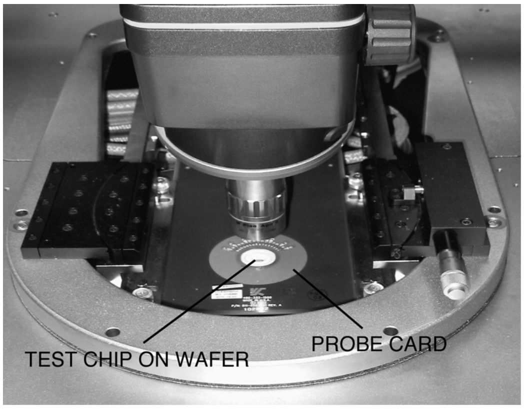

Fig. 3. Hardware setup used for test and measurement of our chip.

In Fig. 4, the performance of a DUT module of type A is compared to our baseline, which is the maximum operating frequency of an otherwise identical 32-bit CLA design without the sleep transistor, across the allowed range of supply voltages. Negligible frequency degradation is observed when the sleep transistor in a type A DUT module is sized at 2.6% of the total pMOS and nMOS of the 32-bit CLA and the corresponding output registers. But if the sleep transistor in the type A DUT module is made a lot smaller, so that it is only 1.0% of the total pMOS and nMOS size, then the frequency degrades by as much as 8.25%. The kink in the upper part of the curves at high voltages is due to the 650 MHz clocking limit of our test and measurement setup.

The leakage consumption of a macro with a type B DUT module was also compared to that of a macro with a type A DUT module in both PARK and COLD modes. The results are shown in Fig. 5. At a supply voltage of 0.9 V, the COLD mode reduces the leakage power by a factor of 43 compared to IDLE mode.

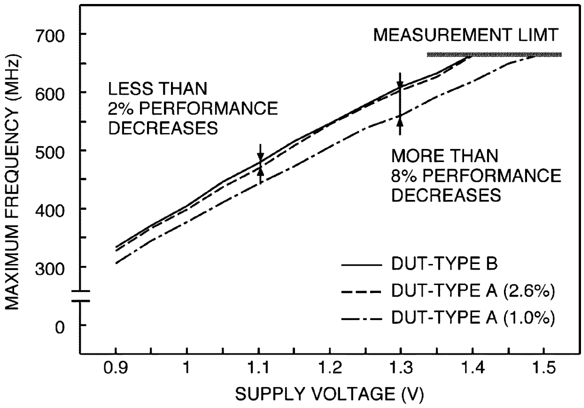

Fig. 4. Performance as a function of supply voltage for type A (2.6%), type A (1.0%), and type B DUT modules.

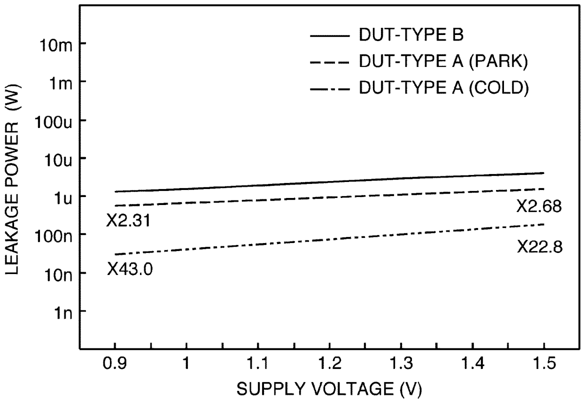

Fig. 5. Leakage savings from the PARK and COLD modes as a function of supply voltage.

The effectiveness of this power supply interrupt gradually decreases as the supply voltage increases, declining to a factor of approximately 23 at a supply voltage of 1.5 V. The pMOS transistor in the power gate switch regulates the leakage reduction in PARK mode across the allowable voltage range. The reduction in leakage power is shown to be approximately 2.6 times less than in IDLE mode.

The ground bounce noise is externally measured on the wafer using the picoprobe and hardware setup shown in Fig. 3. No additional discrete decoupling capacitors are mounted on the probe card, although the inherent capacitance of the probe and PCB are nontrivial. Fig. 6 shows the measured off-chip ground bounce when the power modes of DUT_{1-8} are switched from COLD direct to IDLE, while DUT_0 is held in IDLE mode. The families of curves shown in these two figures were generated by repeating the mode transition and ground rail voltage measurement while varying the supply voltage between 0.9 and 1.5 V, in 0.1-V increments. The value of the supply voltage affects the amplitude of the ground bounce. By scaling down the supply voltage, not only is there a reduction in the charge stored in the parasitic capacitance of the logic circuitry and the VGND node during COLD mode, but also the current flowing inside the logic circuitry rapidly becomes weaker. Both of these effects tend to reduce the ground bounce noise induced by instant turn-on of the sleep or data-retention transistors in the power gating structure. Reducing the supply voltage by 40% (from 1.5 to 0.9 V) reduces the ground bounce in the transition from COLD to IDLE by almost 75%.

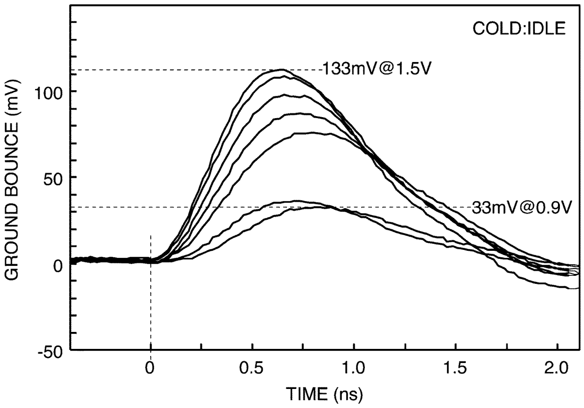

Fig. 6. Measured ground bounce when the power mode is switched from COLD direct to IDLE.

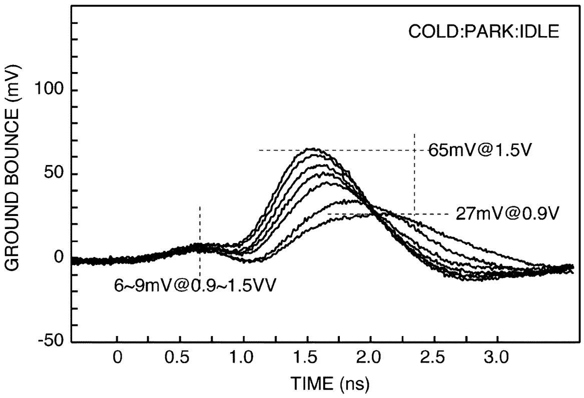

Fig. 7. Measured ground bounce when the power mode is switched from COLD to IDLE through PARK mode.

Fig. 7 shows the measured off-chip ground bounce when the power mode of DUT_{1-8} is sequentially switched from COLD to PARK and from PARK to IDLE. To reduce the ground bounce noise generated by the transition from COLD to PARK, only a quarter of the pMOS devices of power gating structure in each DUT module are initially turned on, while in the nMOS devices are all turned on after a short delay. When we compare these results with those of Fig. 6, we see that the power mode transition from COLD to IDLE through PARK reduces the ground bounce by up to 50%, depending on the supply voltage.

Just measuring the off-chip ground noise may not be enough to understand the impact of on-chip ground bounce on the performance of logic circuitry running in active mode. To investigate the effect of on-chip ground bounce, the maximum operating frequency of DUT_0 is measured for power-mode transitions by $\mathrm{DUT1}_{1-8}$ from COLD direct to IDLE and from COLD to IDLE through PARK mode. The maximum operating frequency of DUT_0 is degraded by the ground bounce noise induced by the power-mode transition noise generated by DUT_{1-8}, coupled with the high-speed clocking noise of DUT_0.

To switch from COLD direct to IDLE, we turn off the nMOS sleep transistor by setting $V_{\mathrm{PG}} = 0$, and also turn off the pMOS data-retention transistor by setting $V_{\mathrm{G,HLD}} = \mathrm{VDD}$. We then wait for 50 μs for the voltage levels at all the internal nodes to

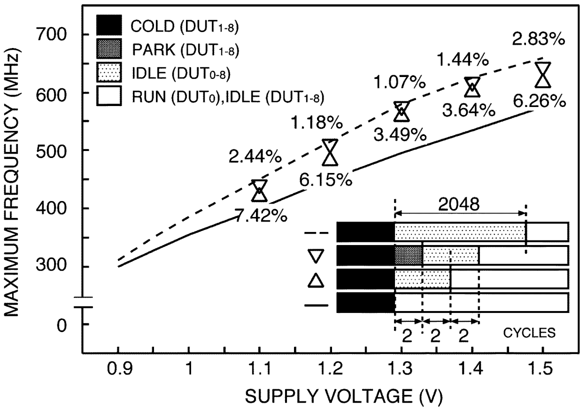

Fig. 8. Effect of ground bounce on the performance of nearby logic DUT_0, when the power mode of DUT_{1-8} is switched from COLD to IDLE.

stabilize. Finally, we turn on the nMOS sleep transistor by setting $V_{\mathrm{PG}} = \mathrm{VDD}$, and wait for 4 clock cycles before measuring the maximum operating frequency of DUT_0. We can compare this result with the maximum operating frequency of DUT_0 on the same macro, with the sleep transistors of DUT_{1-8} always on and the clock off so that DUT_{1-8} are always in IDLE mode, thus effectively excluding the ground bounce noise produced by clocking DUT_{1-8}. A similar procedure is used to analyze the transition from COLD to IDLE through PARK mode. The modules only stay in PARK mode for 2 clock cycles.

Fig. 8 shows the reduction of the maximum operating frequency of DUT_0 in RUN mode that occurs between 0.9 and 1.5 V. Unlike the results shown in Figs. 6 and 7, Fig. 8 shows the internal impact of the ground bounce related to power-mode transitions on the maximum operating frequency of the CMOS logic circuits. If DUT_0 starts operating without waiting for any clock cycles after the power mode of DUT_{1-8} modules is changed from COLD to IDLE, its maximum operating frequency is degraded by up to 14.5%.

VI. Conclusion

We have proposed and evaluated a tri-mode power gating structure with two power-saving modes for deep-submicron technologies operating at a low supply voltage. These modes allow a choice between a large reduction in leakage without state retention and an intermediate level of leakage reduction with state retention and a reduction in the ground bounce noise induced by power mode transitions of the power gating structure. Representative logic circuits with and without power gating circuits were designed and fabricated in 0.13-μm CMOS bulk technology. Test results show that, when a moderate area overhead is dedicated to the sleep transistor in the power gating structure ($<2.6\%$), the maximum operating frequency is decreased by less than 2.0%. The leakage current is dramatically reduced when the ground supply to the logic circuit is interrupted by the small nMOS sleep transistor switch, and is moderately reduced (by slightly more than a factor of two) when the pMOS data-retention switch is used to reduce the rail-to-rail voltage. The ground bounce induced by switching between power modes was measured, together with its effect on the performance of neighboring circuits. The intermediate leakage-saving mode significantly reduces the ground bounce and hence its effect on the performance of neighboring circuits.

If our power gating structure were to be applied to a static random-access memory (SRAM) required to retain data despite process, supply voltage, and temperature (PVT) variations [13], [15], then the VGND node would need to be carefully controlled to maintain sufficient cell stability and avoid potential data loss.

References

[1] H. Mahmoodi-Meimand and K. Roy, "A lekage-tolerant high fan-in dynamic circuit design style," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 51, no. 3, pp. 495–503, Mar. 2004.

[2] S. Mutoh, T. Douseki, Y. Matsuya, T. Aoki, S. Shigematsu, and J. Yamda, "1-V power supply high-speed digital circuit technology with multithreshold-voltage CMOS," *IEEE J. Solid-State Circuits*, vol. 30, no. 8, pp. 847–854, Aug. 1995.

[3] Z. Hu, A. Buyuktosunoglu, V. Srinivasan, V. Zyunban, H. Jacobson, and P. Bose, "Microachitectural techniques for power gating of execution units," in *Proc. Int. Symp. Low-Power Electron. Design*, 2004, pp. 32–37.

[4] S. Kim, S. V. Kosonocky, and D. R. Knebel, "Understanding and minimizing ground bounce during mode transition of power gating structure," in *Proc. Int. Symp. Low-Power Electron. Design*, Aug. 2003, pp. 22–25.

[5] A. Abdollahi, F. Fallah, and M. Pedram, "An effective power mode transition technique in MTCMOS circuits," in *Proc. Design Autom. Conf.*, Jun. 2005, pp. 37–42.

[6] J. Gu, R. Harjani, and C. Kim, "Distributed active decoupling capacitors for on-chip supply noise cancellation in digital VLSI circuits," in *Proc. IEEE Symp. VLSI Circuits*, 2006, pp. 216–217.

[7] K. Kumagai, J. Iwaki, H. Suzuki, T. Yamada, and S. Kurosawa, "A novel powering-down scheme for low Vt CMOS circuits," in *Dig. Tech. Papers IEEE Symp. VLSI Circuits*, 1998, pp. 44–45.

[8] B. R. McDaniel and L. T. Clark, "Integrated Circuit Low Leakage Power Circuitry for Use With an Advanced CMOS Process," U.S. Patent # 6 166 985, 2000.

[9] M. Ferretti and P. A. Beerel, "Low swing signaling using a dynamic diode-connected driver," in *Proc. Eur. Solid-State Circuits*, 2001, pp. 369–372.

[10] A. U. Diril, Y. S. Dhillon, A. Chatterjee, and A. D. Singh, "Pseudo dual supply voltage domino logic," *J. Low Power Electron.*, vol. 1, no. 2, pp. 145–152, Aug. 2005.

[11] L. T. Clark, R. Patel, and T. S. Beatty, "Managing standby and active mode leakage power in deep-submicron design," in *Proc. Int. Symp. Low-Power Electron. Design*, Aug. 2004, pp. 274–279.

[12] K. Agarwal, H. Deogun, D. Sylvester, and K. Nowka, "Power gating with multiple sleep modes," in *Proc. Int. Symp. Quality Electronic Design*, Mar. 2006, pp. 633–637.

[13] K. Osada, Y. Saitoh, E. Ibe, and K. Ishibashi, "16.7-fa/cell tunnel-leakage-suppressed 16-Mb SRAM for handling cosmic-ray-induced multierros," *IEEE J. Solid-State Circuits*, vol. 38, no. 11, pp. 1952–1957, Nov. 2003.

[14] L. T. Clark, M. Morrow, and W. Brown, "Reverse-body bias and supply collapse for low effective standby power," *IEEE Trans. Very Large Scale Integration (VLSI) Syst.*, vol. 12, no. 9, pp. 947–955, Sep. 2004.

[15] K. Zhang, U. Bhattacharya, Z. Chen, F. Hamzaoglu, D. Murry, N. Vallepalli, Y. Wang, B. Zheng, and M. Bohr, "SRAM design on 65-nm CMOS technology with dynamic sleep transistor for leakage reduction," *IEEE J. Solid-State Circuits*, vol. 40, no. 4, pp. 895–901, Apr. 2005.