Simultaneous Control of Subthreshold and Gate Leakage Current in Nanometer-Scale CMOS Circuits

Youngsoo Shin‡, Sewan Heo‡, Hyung-Ock Kim‡, and Jung Yun Choi§
‡Department of Electrical Engineering, KAIST, Daejeon 305-701, Korea
§Samsung Electronics, Yongin, Gyeonggi-Do 449-711, Korea

ABSTRACT

Power gating has been widely used to reduce subthreshold leakage. However, its efficiency degrades very fast with technology scaling due to the gate leakage of circuits specific to power gating, such as storage elements and output interface circuits with a data-retention capability. A new scheme called *supply switching with ground collapse* is proposed to control both gate and subthreshold leakage in nanometer-scale CMOS circuits. Compared to power gating, the leakage is cut by a factor of 6.3 with 65nm and 8.6 with 45nm technology. Various issues in implementing the proposed scheme using standard-cell elements are addressed, from RTL to layout. The proposed design flow is demonstrated on a commercial design with 90nm technology, and the leakage saving by a factor of 32 is observed with 3% and 6% of increase in area and wirelength, respectively.

I. INTRODUCTION

Subthreshold leakage current increases exponentially with every process generation, due to the scaling down of the threshold voltage. Reducing subthreshold leakage has therefore been conceived as a key to achieving low standby power. Power gating [1], [2], [3] uses a current switch to cut off a circuit from its power supply rails during standby mode, and has been widely used in the semiconductor industry to reduce subthreshold leakage [4], [5], [6], [7].

Although power gating is efficient in controlling subthreshold leakage, it suffers from a gate oxide direct tunneling current (gate leakage, for brevity). Furthermore, gate leakage grows very fast with CMOS technology scaling, even faster than subthreshold leakage, due to the scaling down of the gate oxide thickness. In fact, for CMOS technology of 60nm and below, gate leakage is expected to exceed subthreshold leakage [8]. The efficiency of power gating does indeed degrade very fast with technology scaling and also with temperature. The reduction in efficiency is due to the gate leakage of circuits specifically associated with power gating, such as storage elements and output interface circuits with a data-retention capability, and current switches.

In order to overcome the efficiency limitation of power gating circuits, we propose a new circuit technique, which we call *supply switching with ground collapse (SSGC)*. This reduces standby gate leakage by switching to a lower-voltage supply, while current switches, through which ground collapses in standby mode, suppress subthreshold leakage. We address various issues in the design of SSGC: choice of a low-voltage standby supply and design of supply switching circuits, design of power network and current switches, and the design of SSGC-specific circuits. We also discuss the design flow for SSGC using standard-cell elements, starting from the RTL description of a circuit and going down to a layout. A range of experiments show that, compared to power gating, SSGC reduces leakage by a factor of 5 to 7 at 65nm and 6 to 11 at 45nm. The proposed design flow is demonstrated on a commercial design, showing the validity of the proposed method.

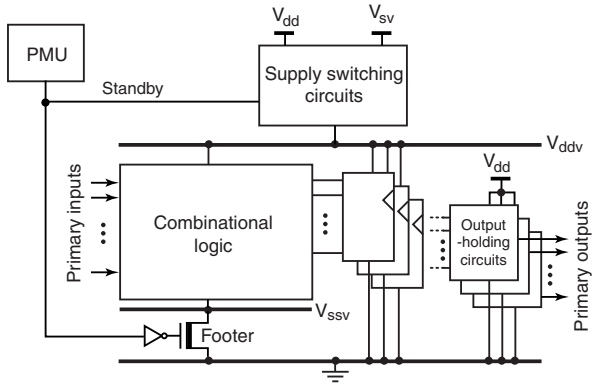

Fig. 1. Supply switching with ground collapse.

The remainder of this paper is organized as follows: in the next section we propose SSGC and deal with various aspects of SSGC design using standard-cell elements. Experimental results are presented in Section III, and the application of SSGC to a commercial design is studied in Section IV. We draw conclusions in Section V.

II. SUPPLY SWITCHING WITH GROUND COLLAPSE

The efficiency of power gating degrades with technology scaling due to the presence of gate leakage in data-retention storage elements and output holding circuits. Using SSGC, the component of gate leakage in storage elements is reduced by dropping the supply voltage, while power gating largely eliminates subthreshold leakage in the combinational circuits.

Fig. 1 shows the SSGC concept. When the circuit is in active mode, the normal supply voltage (V_{dd}) is applied through supply control switches and the footer is turned on. When the

1-4244-0630-7/07/$20.00 ©2007 IEEE.

PMU detects[1] that the circuit is in standby state, it steers the supply control switches so that the standby supply voltage (V_{sv}) is applied to the circuit. At the same time, the footer is turned off and subthreshold leakage from the combinational logic is eliminated. Note that the storage elements bypass the footer and are directly connected to V_{ss}, so that conventional storage elements can be used without any modification. This solves the two main problems of conventional power gating: gate leakage and the overhead of the data-retention storage elements.

The voltage V_{sv} is considerably lower than V_{dd}, significantly reducing the standby gate leakage since gate leakage is proportional to $V_{dd}{}^4$ [9]. The standby voltage V_{sv} should be chosen so that the potential that drives the logic block (the virtual supply voltage V_{ddv} in Fig. 1) is higher than the minimum voltage necessary for the storage elements to retain their states, plus some margin to guarantee state retention in the presence of noise.

A. Choice of the Standby Supply Voltage and Design of Supply Switching Circuits

The key ingredient of SSGC is V_{ddv} in standby mode, which should be as low as possible to reduce gate leakage but, at the same time, high enough to maintain the states of the storage elements. Temperature and process variation can affect the integrity of states at the reduced standby voltage, and need to be taken into account when we determine standby V_{ddv} (note that standby V_{ddv} is only used to choose V_{sv}, and V_{ddv} itself is not constant).

As an example, we took a D flip-flop (based on tristate-inverters and inverters) in commerical 90nm as shown in Fig. 2. We repeated SPICE simulation and determined the lowest V_{ddv} for temperatures between -40°C to 125°C, which we assumed to be a realistic operating range. We repeated the same experiment for different process corners to allow for process variation: each plot in Fig. 2 corresponds to a different process corner. Fig. 2 shows that at least 260mV is required to power this flip-flop reliably in the presence of temperature and process variations. If a design has several types of storage element, these experiments need to be repeated for each type, and the maximum necessary voltage can then be assumed as V_{ddv}.

The supply switching circuits in Fig. 1 can be designed as two MOS switches, as shown in Fig. 3. The normal supply voltage V_{dd} is supplied through M1, which is a PMOS switch with a high threshold voltage. Using a device with a high threshold can reduce the subthreshold leakage of M1, which is turned off in standby mode. In active mode, V_{ddv} is lower than V_{dd} due to the voltage drop across M1, which increases the circuit delay. This implies that the sizing of M1 is important for circuit performance. The wake-up delay, which is the delay in switching from standby to active mode (i.e. the time needed to turn off M2 and turn on M1), is also dependent on the size of M1.

A low threshold voltage is preferred for the size of M2. This may increase the subthreshold leakage, but that has little impact

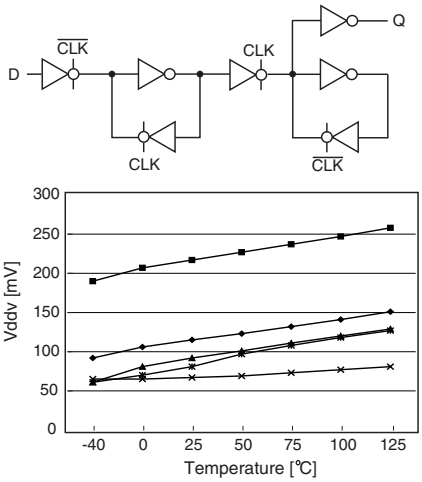

Fig. 2. Low supply voltage for data retention.

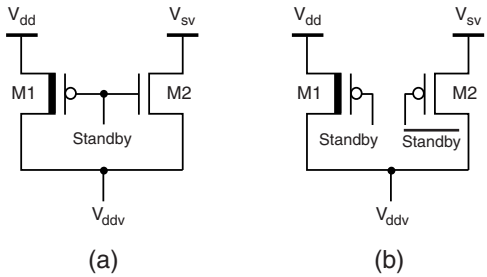

Fig. 3. Supply switching circuits with M2 implemented in (a) NMOS and (b) PMOS device.

since M2 turns off in active mode, while most of the leakage current in the circuit occurs during standby mode. The polarity (NMOS vs. PMOS) and the size of M2 are important determinants of the total leakage current, since they affect the choice[2] of V_{sv}, which in turn determines the gate leakage.

Our experiments show that the use of NMOS switch generally reduces the leakage current, and is therefore preferable for M2. At most temperatures, the total leakage current in the circuit with an NMOS switch is less than half of that with a PMOS switch. This can be understood from the observation that V_{sv} is higher for PMOS transistor at the maximum temperature (recall that V_{sv} is determined at the maximum temperature), which leads to higher values of V_{ddv} at lower temperatures, making the total leakage greater than it would be with an NMOS switch.

The size of the NMOS switch M2 should be chosen to minimize the total leakage, so long as the area overhead from the switch can be tolerated. Simulating a circuit to determine the total leakage while varying the size of M2 can take a significant amount of time. Fortunately, it can be shown through experiment that the total leakage approximately correlates with V_{sv} (e.g. see Fig. 4, which corresponds to the industrial example used in Section IV). Sizing can therefore be performed more economically by changing the NMOS size and deducing V_{sv}

[1]There are many alternative power management interfaces and the full range of possibilities is beyond the scope of this paper. For example, a circuit may have internal logic that detects its own standby state and sends a standby request to the PMU. The PMU may then acknowledge the request, depending on the configuration of the whole system. The same logic can be used to detect the wakeup condition and to interface with the PMU to achieve a return to active mode.

[2]Note that V_{sv} has to be higher than the minimum voltage needed for data retention, since there is a voltage drop across M2 in standby mode.

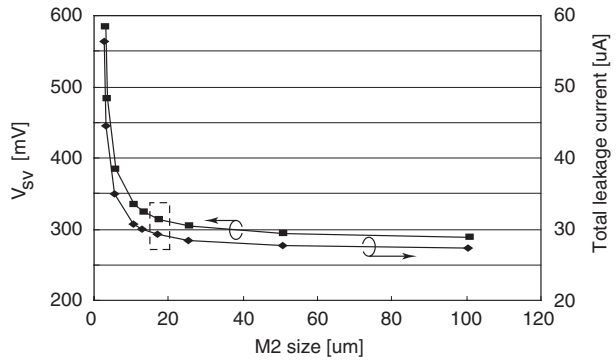

Fig. 4. Correlation of V_{sv} and total leakage current.

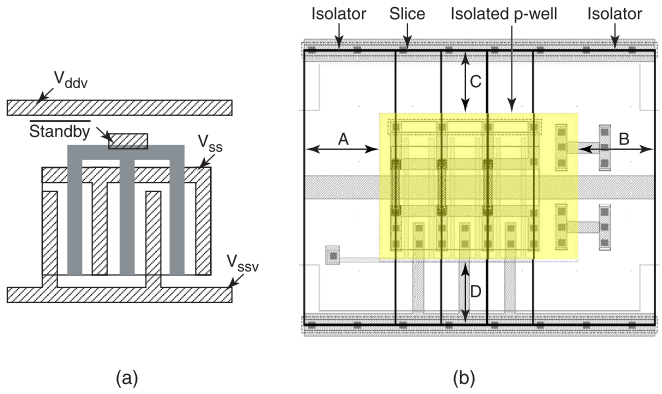

Fig. 6. (a) Conceptual layout of a footer cell and (b) the layout of a footer with slices and isolators.

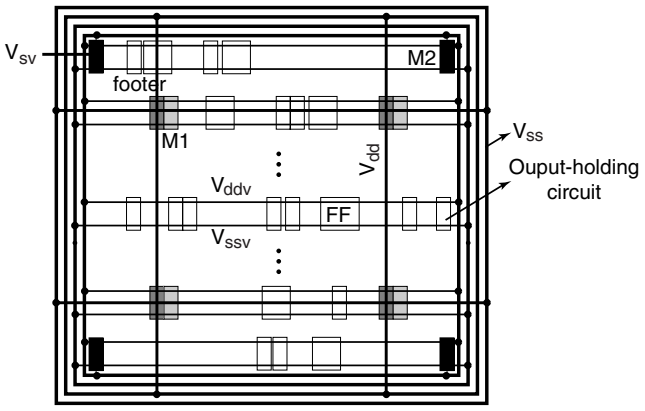

Fig. 5. Power networks for supply switching with ground collapse.

from the average leakage current of the circuit, with V_{ddv} set to the minimum voltage that supports data retention.

B. Design Methodology

In this subsection, we will discuss various issues that arise in applying SSGC to designs based on standard-cell elements.

B.1 Power Networks

Fig. 1 shows that we need additional power networks for V_{ddv} and V_{ssv}, as well as conventional networks for V_{dd} and V_{ss}. To meet this demand, we propose the new power network topology shown in Fig. 5. These networks consist of four power rings and corresponding power rails. The networks providing V_{dd} and V_{ss} are connected to chip-level power networks, while the V_{ddv} and V_{ssv} networks are local. Note also that the V_{ddv} and V_{ssv} rails connect respectively to the VDD and VSS terminals of the cells implementing combinational logic, allowing unmodified conventional standard-cell logic elements to be used.

The locations of the footer and the M1 switch are important, since they affect the circuit operation in active mode. In Fig. 5, for example, they are located in the four quadrants of the placement region. Accurate analysis of the power network may be required, depending on the power delivery requirements (current, IR drop, electromigration, etc) that need to be imposed. M1 receives V_{dd} from a vertical rail that resides in a higher metal layer, and connects to V_{ddv} through its VDD terminal. Similarly, the footer connects to V_{ssv} through its VSS terminal, while connecting to V_{ss} via a horizontal rail in a higher metal layer. The placement of M2 is less important, since it only supplies standby voltage (V_{sv}). In Fig. 5, M2 switches are located in four corners of the placement region.

Storage elements need V_{ddv} and V_{ss}, as shown in Fig. 1. Because we modify the conventional storage elements slightly, to reduce their subthreshold leakage, they require V_{ssv} as well. (The detail is explained in Section B.3) As a result, the elements' VDD and VSS terminals connect to the V_{ddv} and V_{ssv} rails, while the connection to V_{ss} is made through their signal pins. The output-holding circuit receives both V_{dd} and V_{ss} through its signal pins.

B.2 Footer Design

Fig. 6(a) shows a conceptual cell layout of a footer switch. Its source and drain terminals are connected to V_{ss} and V_{ssv} respectively, while its VDD terminal merely serves as a connecting medium for the cells on its left- and right-hand sides.

The body biasing of logic cells is implicit (the body of PMOS to V_{ddv} and the body of NMOS to V_{ssv}), since we do not modify any standard cell layout. However, the body of a footer can be biased either to its source or to its drain. This allows us a trade-off between area overhead and leakage saving. If the body of a footer is connected to its drain (i.e. V_{ssv}), the footer can share its body with logic gates, which makes the layout more compact. However, when a footer is turned off, V_{ssv} approaches V_{ddv}, resulting in a p-n junction current in the footer, which is a disadvantage in terms of standby mode leakage. Conversely, if the body is connected to its source (i.e. V_{ss}), the leakage situation improves, although there is an area overhead due to the need for well isolation. We chose this second option to minimize leakage.

To cope with the area overhead due to well isolation, we build a footer by combining two types of cells, which we call a *slice* and an *isolator*. A slice is a unit footer; when slices are abutted together, they constitute a larger footer. Isolators are placed at both ends of the slices so that there is guaranteed to be enough room between the footer and the logic cells for well isolation. Fig. 6(b) shows a footer constructed by abutting three

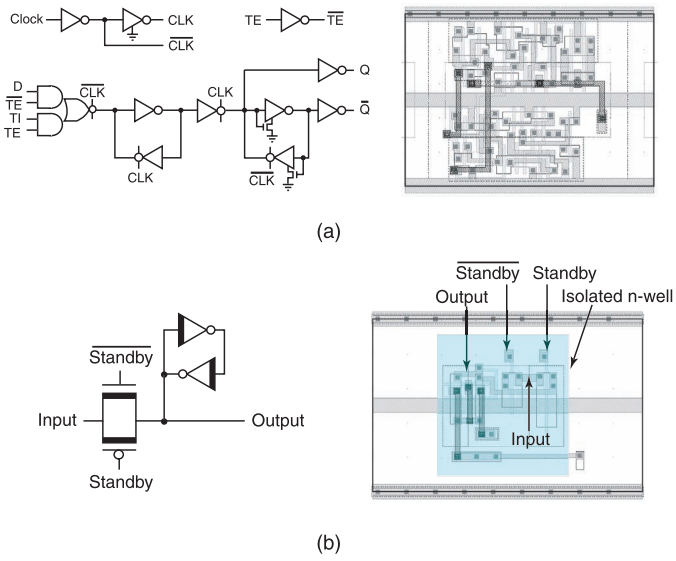

Fig. 7. SSGC cells: (a) flip-flop and (b) output-holding circuit.

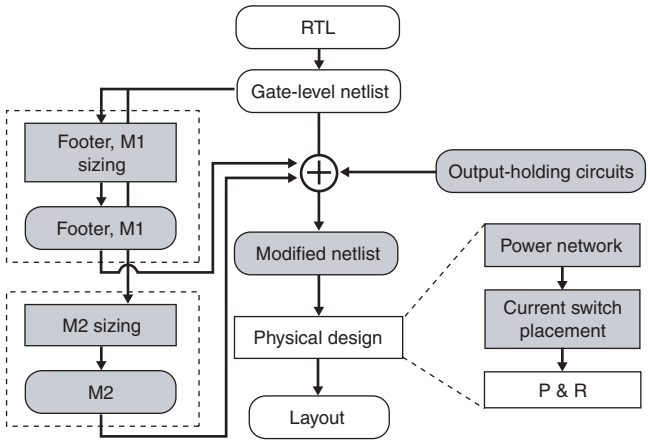

Fig. 8. Design flow.

slices with two isolators. The footer needs to be placed in an isolated p-well, which in turn needs to be inside an n-well for well isolation. The extra spaces required are denoted by A, B, C, and D.

Once the size (width) of a footer has been determined [10] from a given performance requirement, we know the number of slices that need to be placed. In terms of a simple tally of area, the best way to place slices is to abut them all together, since this requires only two isolators. But a single large footer can block placement of the logic cells. Furthermore, the power network (i.e. V_{ssv}) may experience a large IR drop if the logic cells are physically distant from the footer. Instead we place discrete footers in a regular pattern, as shown in Fig. 5.

B.3 SSGC Cells

While power gating requires data-retention storage elements, conventional storage elements can be used in SSGC, since they are not power gated and their standby leakage current is controlled by reducing the supply voltage (V_{sv}). However, if we power-gate a portion of a storage element which is not involved in saving the state during standby mode, we can further reduce the subthreshold leakage. For instance, in the testable flip-flop shown in Fig. 7(a), only a slave latch is connected to V_{ss} through serially connected NMOS switches, while the remainder of the circuit is power gated (i.e. connected to V_{ssv}). Fig. 7(a) also shows a layout of the flip-flop. This choice slightly increases the area of the storage elements but does not affect the wire-length, since these elements are not controlled by the PMU.

Like power gating, SSGC needs an output-holding circuit, since the outputs of a circuit are driven by reduced voltage (V_{sv}) in standby mode, while the blocks that are connected to the outputs may be driven by V_{dd} (either because they are in active mode or because they do not employ SSGC). Fig. 7(b) shows our output-holding circuit and its layout.

B.4 Design Flow

The design flow for SSGC is shown in Fig. 8, where SSGC-specific steps are highlighted. The register transfer level (RTL) design goes through a traditional logic synthesis to create the gate-level netlist. In order to determine the size of M1 and the footer, we first apply random logic patterns to the inputs of the netlist, simulate it with a circuit simulator, which gives us the average current. Combining this result with the target delay penalty[3] and the turn-on resistance of a minimum-size MOS transistor gives us the size of M1 and the footer [10]. The size of M2, together with V_{sv}, can now be determined from the leakage of the circuit at maximum temperature. An output-holding circuit is inserted at each primary output.

In the physical design stage, we first generate the conventional power and ground networks, which serve as V_{ddv} and V_{ssv} networks in SSGC. Combining these with the extra networks for V_{dd} and V_{ss}, we now have our power networks, as shown in Fig. 5. The slice blocks for a footer, and the M1 and M2 switches, are placed in a regular fashion and then fixed in their locations. After the placement of the logic cells, the signal routing and the routing of the standby signal can be performed. The transistor-level netlist is extracted from the layout and simulated with SPICE to estimate the leakage current.

III. Experimental Results

We performed experiments on a set of circuits taken from the ISCAS'89 and ITC'99 benchmarks. In Table I, the second and three subsequent columns show the characteristics of the original circuits. The remaining columns show the leakage savings achieved by power gating (pg) and SSGC (as factors), for implementation in 65 and 45nm predictive technologies[11], all at room temperature. To implement power gating, we used data-retention flip-flops and output-holding circuits. The sizing of the footers was based on the average current, assuming that a 10% increase in delay can be tolerated. For the SSGC implementation, we assumed that the same 10% delay penalty for

[3]We assume that the footer and M1 contribute equally to the delay increase in the circuit. In other words, both footer and M1 take equal responsibility for the delay penalty.

TABLE I
TOTAL LEAKAGE REDUCTION FACTOR. TEMPERATURE = 25°C.

Circuits	Inputs	Outputs	Flip-flops	Gates	65nm		45nm	
					pg	SSGC	pg	SSGC
s344	8	11	15	175	5.3	29.5	4.2	25.6
s1269	18	10	37	659	8.3	58.7	6.2	38.2
s3384	43	26	183	1621	6.2	42.1	5.2	51.4
b03	4	4	30	174	5.0	33.2	4.2	46.4
b14	32	54	245	7108	16.4	93.3	11.6	116.3

sizing M1 and a footer would be allowed, and we used the circuits shown in Fig. 7 for flip-flops and outputs.

The leakage saved by SSGC increases with technology scaling, since the gate leakage now takes a higher proportion of the total leakage current. The leakage is cut by a factor of 51 with 65nm and 56 with 45nm technology, on average. Conversely, and for the same reason, the saving from power gating decreases as the technology is scaled down. On average, compared to power gating, the leakage is cut by a factor of 6.3 with 65nm and 8.6 with 45nm technology. These results demonstrate the efficiency of SSGC as technology scales down, and power gating becomes less effective. The ability of SSGC to save leakage is determined by the number of storage elements and outputs in the original circuit, since they are the main sources of leakage current in standby mode (compare b03 and b14, for example). Therefore, the leakage saving of SSGC is relatively insensitive to the delay penalty, although the area overhead will increase with decreased delay penalty.

We repeated the experiments at different temperatures to explore the temperature dependency of both techniques. For SSGC, the leakage saving improves with decreasing temperature, since the subthreshold leakage gets smaller as the gate leakage becomes a higher proportion of the total. This is in contrast to power gating, where leakage saving decreases as temperature goes down.

IV. CASE STUDY: EMBEDDED TRACE MACROCELL

In order to validate SSGC, we used an embedded trace macrocell (ETM) [12] as a test vehicle. An ETM provides debug and trace facilities for ARM processors, and allows information about the processor's state to be captured both before and after a specific event. The original design used in this experiment consists of 90K gates after mapping on to a commercial 90nm, 1.0V gate library. The design has 124 outputs, and the same number of output-holding circuits are required for an SSGC implementation. There are 5.5K storage elements, made up of four types of flip-flops and a single type of latch. To determine the standby voltage (V_{sv}) of the design, we first repeated a process similar to the one shown in Fig. 2 for each storage element. The highest of the voltages needed for data-retention was 284mV, which was then used to choose the size of M2 (17.6μm) and V_{sv} (314mV) as shown in Fig. 4. The design then follows the flow in Fig. 8. Physical design was done in the flat, with floorplan constraints imposed on two large sub-

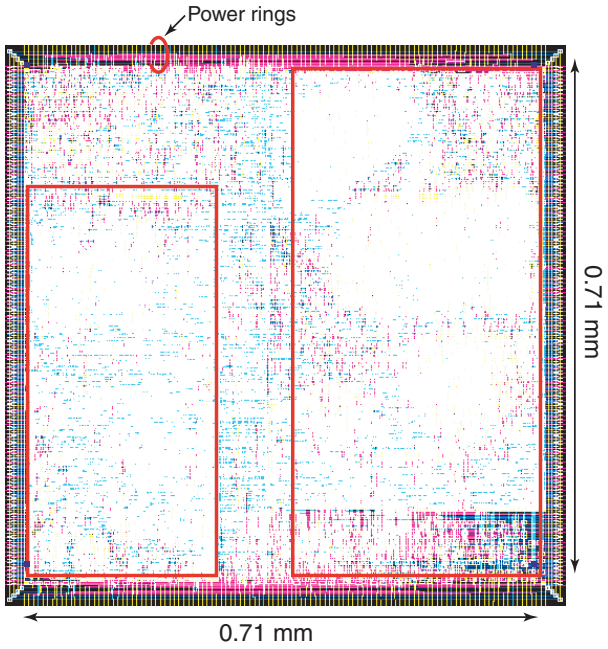

Fig. 9. Final layout of ETM with SSGC.

blocks. The final layout is shown in Fig. 9. Four power rings are also shown in the figure.

SSGC reduces the leakage current by a factor of 32 at 25°C (13μA compared to 410μA with the original design). The saving goes up at reduced temperatures, as we would expect: rising to a factor of about 130 at -40°C. This result is in accordance with the experimental results presented in the previous section, implying that we could expect more savings in 65nm and smaller technologies, where gate leakage takes a higher proportion of the total standby leakage current.

We analyze the contribution to leakage current by different components of the design. The storage elements (SEs) draw 12.8μA and thus are responsible for most of the leakage current. This is understandable since their subthreshold leakage can be significant due to the use of a low V_t (see Fig. 7), although leakage is reduced by a lower supply voltage (V_{sv}) in standby mode. Footers and output-holding circuits take a negligible leakage current due to their use of high V_t transistors. However, it should be noted that the relative contribution of different elements will be different depending on the proportion of gate leakage in the total leakage current.

We have also analyzed the overhead of using SSGC with this layout, in terms of area and wirelength. The area increases by 3%, which is almost negligible. The total wirelength only increased by 6%. This is because SSGC does not use the PMU to control storage elements. In contrast, the data-retention storage elements used with power gating need control signals, which significantly increase the total wirelength.

V. Conclusion

Although power gating has been widely used to reduce subthreshold leakage, its efficiency degrades very fast with technology scaling, limiting its application to nanometer-scale technologies, such as 65 and 45nm. This is due to the gate leakage of circuits specific to power gating, such as data-retention storage elements and output-holding circuits.

In order to overcome this limitation of power gating, we have proposed a new circuit technique called supply switching with ground collapse. We performed a range of experiments to compare the leakage with SSGC and with power gating. SSGC outperforms power gating by a factor of 5 to 7 at 65nm and 6 to 11 at 45nm. We have presented the design flow for applying SSGC to a semi-custom design using standard-cell elements, and we have demonstrated its feasibility on a commercial design using commercial 90nm technology.

References

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

[2] K. Usami, N. Kawabe, M. Koizumi, K. Seta, and T. Furusawa, "Automated selective multi-threshold design for ultra-low standby applications," in *Proc. Int'l Symposium on Low Power Electronics and Design*, Aug. 2002, pp. 202–206.

[3] T. Kitahara, N. Kawabe, F. Minami, K. Seta, and T. Furusawa, "Area-efficient selective multi-threshold CMOS design methodology for standby leakage power reduction," in *Proc. Design, Automat. and Test in Europe Conference and Exhibition*, Mar. 2005, pp. 646–647.

[4] S. Shigematsu, S. Mutoh, Y. Matsuya, Y. Tanabe, and J. Yamada, "A 1-V high-speed MTCMOS circuit scheme for power-down application circuits," *IEEE Journal of Solid-State Circuits*, vol. 32, no. 6, pp. 861–869, June 1997.

[5] S. V. Kosonocky, M. Immediato, P. Cottrell, and T. Hook, "Enhanced multi-threshold (MTCMOS) circuits using variable well bias," in *Proc. Int'l Symposium on Low Power Electronics and Design*, Aug. 2001, pp. 165–169.

[6] H.-S. Won, K.-S. Kim, K.-O. Jeong, K.-T. Park, K.-M. Choi, and J.-T. Kong, "An MTCMOS design methodology and its application to mobile computing," in *Proc. Int'l Symposium on Low Power Electronics and Design*, Aug. 2003, pp. 110–115.

[7] P. Royannez, H. Mair, F. Dahan, M. Wagner, M. Streeter, L. Bouetel, J. Blasquez, H. Clasen, G. Semino, J. Dong, D. Scott, B. Pitts, C. Raibaut, and U. Ko, "90nm low leakage SoC design techniques for wireless applications," in *Proc. Int'l Solid-State Circuits Conf.*, Feb. 2006, pp. 138–139.

[8] S. Sirisantana and K. Roy, "Low-power design using multiple channel lengths and oxide thicknesses," *IEEE Design & Test of Computers*, vol. 21, no. 1, pp. 56–53, Jan. 2004.

[9] R. K. Krishnamurthy, A. Alvandpour, V. De, and S. Borkar, "High-performance and low-power challenges for sub-70nm microprocessor circuits," in *Proc. Custom Integrated Circuits Conf.*, May 2002, pp. 125–128.

[10] S. Mutoh, S. Shigematsu, Y. Gotoh, and S. Konaka, "Design method of MTCMOS power switch for low-voltage high-speed LSIs," in *Proc. Asia South Pacific Design Automat. Conf.*, Jan. 1999, pp. 113–116.

[11] Y. Cao, T. Sato, D. Sylvester, M. Orshansky, and C. Hu, "New paradigm of predictive MOSFET and interconnect modeling for early circuit design," in *Proc. Custom Integrated Circuits Conf.*, May 2000, pp. 201–204.

[12] ARM, "Embedded Trace Macrocell," http://www.arm.com/ products/solutions/ETM.html.