

On Quantifying the Figures of Merit of Power-Gating for Leakage Power Minimization in Nanometer CMOS Circuits

A. Sathanur[‡] A. Calimera[‡] A. Pullini[‡] L. Benini* A. Macii[‡] E. Macii[‡] M. Poncino[‡]
[‡]Politecnico di Torino
Torino, ITALY 10129
^{*}Università di Bologna
Bologna, ITALY 40136

Abstract—Power-gating has proved to be one of the most effective solutions for reducing stand-by leakage power in nanometer-scale CMOS circuits, and different strategies and algorithms for its application have been proposed recently. Unfortunately, power-gating comes with its own set of costs: Performance degradation, area increase, dynamic power increase and routing congestion. When a decision to power-gate a design has to be taken, pros and cons of power-gating have to be properly weighted to achieve optimal results. In this paper, we define “Figures of Merit” (FoMs) for power-gating, which can be used by designers to better understand the benefits and costs of power-gating, thereby allowing them to achieve optimal results. We then quantify the FoMs by applying a state-of-the-art, industry-strength power-gating flow on a set of designs implemented onto an industrial 65nm CMOS process, and provide insightful discussion on how optimum power-gating can be achieved.

I. INTRODUCTION

Leakage power is a major concern in sub-90nm CMOS technologies and numerous design techniques have been introduced to reduce stand-by leakage in digital circuits [1], [2]. Out of the available solutions, power-gating has proved to be very efficient in reducing the stand-by leakage currents in the idle mode. In its most common exploitation, shown in Figure 1, power-gating relies on the insertion of a high-threshold NMOS transistor between the source terminal of a cell/block and the ground rail. The insertion point of the sleep transistor is called *virtual ground line* (VGND).

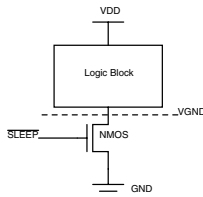


Fig. 1. A Logic Block with NMOS Sleep Transistor.

A sleep transistor can be added on a cell-by-cell basis or can be shared among a number of cells belonging to the same block/cluster. Clustered sleep transistor insertion (STI) implies a smaller sleep transistor area compared to adding a sleep transistor to every single cell, due to discharge current balancing of cells connected to it. In addition, it ensures a lower congestion overhead for sleep signal routing and lower buffering overhead required for sleep signal driving. These are the main reasons why clustered STI has emerged as the preferred choice for industrial application of power-gating. In this paper, we will thus consider clustered STI as the reference power-gating methodology.

Clustered power-gating offers a multitude of challenges and optimization opportunities. The major one concerns the placement of the sleep transistors and the routing of the virtual ground lines. STI is done as a post-layout step; therefore, care should be taken to guarantee little perturbation of the original layout to achieve final design closure. An equally important issue is related to sleep transistor sizing to ensure compliancy with timing constraints. More specifically, since STI intrinsically results in performance degradation due to slowing down of the cells connected to the NMOS footers, the sleep transistor should be up-sized to minimize the virtual ground voltage (V_{VGND}), thus the timing overhead. On the other hand, large sleep transistors would lead to an increase in the overall sleep transistor area and the associated buffers, long power-mode transition times (turn on/off times) and a non-negligible energy spent during the power-mode transition. Depending on the timing criticality of the design and the ratio between stand-by time and active time, one can choose how large the sleep transistors should be to achieve optimal power savings.

To make sure that the total power/area/performance overhead due to power-gating does not overwhelm the benefits of power-gating, a careful investigation of the *Figures of Merit* (FoMs) for power-gating has to be considered to achieve optimal results. This is exactly the main goal of this paper. In particular, we first define a set of FoMs which are suitable for a thorough assessment of the effectiveness of power-gating; next, we discuss how such FoMs can provide useful guidelines to designers for the best exploitation of clustered power-gating. We will refer to realistic data collected on circuit layouts implemented with a state-of-the-art, CMOS process (i.e., the 65nm library by STMicroelectronics), and to an industry-strength design flow for clustered power-gating. To the best of our knowledge, this is the first time such a comprehensive exploration of the pros and cons of power-gating is published for real-life technologies and design flows; therefore, we believe the contribution of this paper will go beyond a simple theoretical value, and it will provide to design practitioners a valuable instrument for supporting their daily work.

The paper is organized as follows. Section II defines various Figures of Merit for power-gating. Section III explains the reference power-gating flow used for conducting the experiments. Section IV presents the results of our trade-off analysis applied to a set of six benchmark circuits of varying size. Finally, Section V concludes the paper.

II. FIGURES OF MERIT (FOMs) FOR POWER-GATING

Recently, a few works have specifically addressed the issue of quantifying the benefits and costs of power-gating. In [3], the authors discuss system level aspects of power-gating and provide insightful discussions on various costs and benefits of this technique. They conclude that, for sub-90nm technologies, the benefits achieved by power-gating are quite substantial compared to the overhead due to it, thus making it a very attractive option to achieve power savings. However, the authors did not address the issues at design/block level and on how one can achieve optimum savings by trading different parameters, such as performance degradation and area. In [4], the authors define the efficiency of power-gating and show how the efficiency varies with different design metrics, such as sleep transistor area, length and also throw light on methodologies, which can be used for sleep transistor power-mode transition. However, the authors do not provide substantial results on real circuits.

As anticipated, this paper takes a step further in this direction by defining various FoMs for power-gating, by benchmarking them against industry-strength design implementations and, finally, by providing useful design tips for applying power-gating in practice. The remainder of this section is devoted to the definition of the FoMs.

A. Figures of Merit

1) *Performance Degradation (α)*: This is a design specification or an input specification, and defines the maximum allowable delay increase in the design/logic block. This is, in fact, the increase in the original critical path delay permitted when the design/logic block is power-gated. We denote this parameter by α , expressed as the percentage increase in the original critical path delay. This metric quantifies how much degradation, in speed or performance, can be tolerated when power-gating a design.

2) *Sleep Transistor Size (W_{sleep})*: Sleep transistor size depends mainly on two parameters: (i) The virtual ground voltage, V_{VGND} ; (ii) The peak discharge current of the design/logic block, I_{peak} . For a given peak current value I_{peak} , one can have a lower V_{VGND} value to obtain a lower speed degradation of the cells in the power-gated design, thus resulting in a larger sleep transistor. On the other hand, a higher value of V_{VGND} would lead to a smaller sleep transistor, but to a higher delay degradation of the cells in the power-gated design. The relationship between W_{sleep} and V_{VGND} can be written approximately as:

$$W_{sleep} = \frac{L}{\mu_n * C_{ox} * R_{on} * (V_{dd} - V_{th})} \quad (1)$$

where L and V_{th} denote the channel length and the threshold voltage of the sleep transistor, respectively, μ_n represents the mobility of the electrons and C_{ox} represents the cell capacitance per unit area of the sleep transistor. Finally, R_{on} denotes the channel resistance of the sleep transistor operating in the linearity region.

While a smaller sleep transistor results in higher performance degradation of a design/block, a larger sleep transistor leads to an increase in chip area, in the number of associated buffers to drive the input sleep signal of the sleep transistor and in the energy dissipated during power-mode transition. Apart from these, as indicated by [3], as the percentage of gated area on a chip increases, more decap insertion in the active circuits is needed to make sure that the ground bounce is under the maximum allowable limits, which again increases leakage and total chip area. Thus, sleep transistor size is a very important design metric in power-gating.

3) *Leakage Power Savings (Savings)*: The main benefit of power-gating is how much leakage power can be saved. As discussed earlier in the paper, performance degradation and sleep transistor area are opposing design constraints and hence one has to trade these metrics very carefully to achieve optimum power savings. For given α and W_{sleep} constraint, one can partially power-gate a design, thus leaving out timing critical cells from being power-gated, thereby satisfying both the constraints while still achieving considerable leakage savings.

4) *Power mode transition time (PM_t)*: Given a power-gated design/block, we define the *turn-off time* as the time required by the design/block to go from the active to the stand-by mode, and the *turn-on time* as the time required by the design/block to make the opposite mode transition. In some cases, a design/block is active most of the time, and it only provides small windows of stand-by time; the availability of short turn-off and turn-on times may enable leakage reductions corresponding to such short stand-by periods. Turn-off and turn-on times mainly depend on the size of the sleep transistor. In particular, the turn-off time depends on the total input capacitance of the sleep transistor and on the time required to pull-up the virtual ground to a stable value. On the other hand, the turn-on time depends on the total capacitance of the virtual ground line and on the net charge flow from the power-gated design/block to the ground. The virtual ground line capacitance has two components, namely: (i) The virtual ground line parasitic capacitance; (ii) The output capacitance of the sleep transistor. We observe that we do not consider the turn-off time of the sleep transistor as an important parameter since, although it is not negligible, it does not introduce any performance penalty in the context of power-gating because it overlaps with the stand-by period. For this reason, in this paper we provide results only for the turn-on time (denoted by PM_t in the remainder of the paper).

5) *Power-mode transition energy (PM_e)*: This quantity denotes a non-negligible energy dissipation during turn-on/off of the sleep transistor. The energy loss occurs in the charging and discharging of the virtual ground line capacitance and in the buffers which drive the sleep transistor. This parameter is strongly coupled with the size of the sleep transistor. A large sleep transistor would mean higher energy dissipation during turn-on/off and vice versa. As indicated in [3], the total energy overhead due to STI should be very low compared to the leakage savings to achieve benefits from power-gating.

III. POWER-GATING FLOW

To quantify the FoMs of power-gating, we adopted a flow based on the row-based methodology of [5]. The methodology realistically assumes that one or more rows are dedicated to the insertion of the sleep transistors and are appended at the end of the existing layout. The methodology guarantees marginal layout changes, that is, it is very much “layout friendly”; as such, it is easily usable within industrial design flows.

A. Clustering and Sizing

As discussed in [5], the clustering algorithm is based on using rows of the layout as atomic clustering objects. It takes two input parameters, namely, the performance (α) and the area (W_{sleep}) constraints. W_{sleep} can be mapped onto an optimal V_{GND} , which sets the individual cell delay degradation. Then, depending on α , the algorithm finds an optimal sub-set of rows of the design that can be power-gated, thus satisfying both the size and the performance constraints.

Sleep transistor sizing requires an accurate estimation of the peak discharge current (I_{peak}) of the design/block to which power-gating is applied. We use the technique of [6] for this task, which builds temporal current waveforms for each cell based on the possible time intervals during which the cell can switch. Using this waveform for each cell, the peak current waveform for the entire design/cluster can be constructed and the actual peak current value can then be obtained.

B. Turn-On/Off Methodology

To measure the PM_t and PM_e , we use the following methodology. We implement a n -stage power-mode transition scheme, as shown in Figure 2. Given a sleep transistor size, we design an optimal inverter chain for driving the sleep transistor input with the goal of achieving minimum turn-off time. For turn-on, we implement a parallel sleep transistor architecture where each sleep transistor is optimally sized to reduce the turn-on time while making sure that the ground bounce is lower than a specified maximum value to avoid sudden current surge on the supply rails. We set a maximum current equal to I_{peak} of the design as the allowable peak current during turn-on. Since different circuits will have different I_{peak} , the PM_t of the circuits also vary accordingly. Note that in the figure, the VGND pull-up transistor is used to pull the virtual ground and the internal nodes of the power-gated design/block to a logical value since some internal nodes become floating when a design/block goes into the power-down mode.

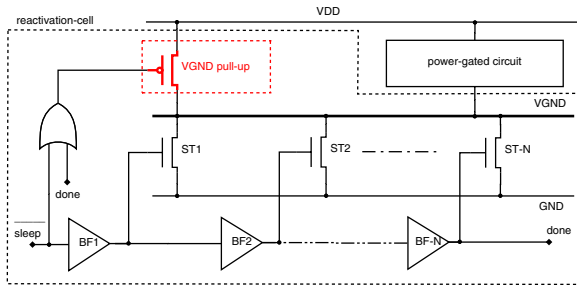


Fig. 2. Sleep Transistor Cell Architecture.

IV. EXPERIMENTAL RESULTS

A. Experimental Set-Up

We applied the power-gating flow to a total of six designs; three of them are public-domain benchmarks taken from the ISCAS85 and ISCAS89 suites, while the remaining three are circuit modules belonging to an industrial SoC provided by STMicroelectronics.

Each design was synthesized, placed and routed onto a 65nm CMOS technology library from STMicroelectronics using Synopsys Physical Compiler for optimal timing. Table I summarizes circuit names and sizes (in terms of number of standard cells).

Benchmark	# Gates
c5315	1459
adder_128	2305
c6288	2368
router	3332
adder_256	4683
s38417	9923

TABLE I. DESIGNS USED FOR THE EXPERIMENTS.

The three experiments we performed targeted the exploration of the achievable leakage savings (i.e., *Savings*), of the power mode transition time (i.e., PM_t), and the power-mode transition energy (i.e., PM_e) as functions of α and W_{sleep} .

For sweeping the α parameter, we first compute the critical path delay for each design and then we set α as the percentage increase in critical path delay ranging from 1% to 10%. For this setting, the sleep transistor size is kept constant at 25% of $W_{sleep-max}$.

For sweeping the W_{sleep} parameter, we first set α at a constant value of 5%, then we compute the total sleep transistor size required to power-gate the entire design with this α value. Let's denote this as $W_{sleep-max}$. We then vary the W_{sleep} from 10% to 100% of $W_{sleep-max}$.

B. Results

Figures 3 and 4 show the leakage power savings we achieved by varying the two input constraints, namely α and W_{sleep} , for the six benchmark designs. As done in [5], a 100% savings indicate that the entire design is power-gated.

From the chart of Figure 3 we can clearly evince that, when the allowed performance degradation increases, the amount of savings increases as well since, for the same area constraint (i.e., for the same sleep transistor), more cells can be power-gated. The variation of leakage savings with W_{sleep} follows the same trend. If we can accommodate more area for the placement of the sleep transistors, we can achieve savings for a given performance degradation. The achievable savings clearly depend also on the type of circuit. For example, circuit c5315 has very many cells with a good amount of slack, and hence we can obtain more savings even with a small area overhead. However in some cases, increasing W_{sleep} might not generate enough slack to power-gate more and hence the savings will not increase. This effect is very evident in design s38417 for W_{sleep} values of 60, 70, 80 and 90%.

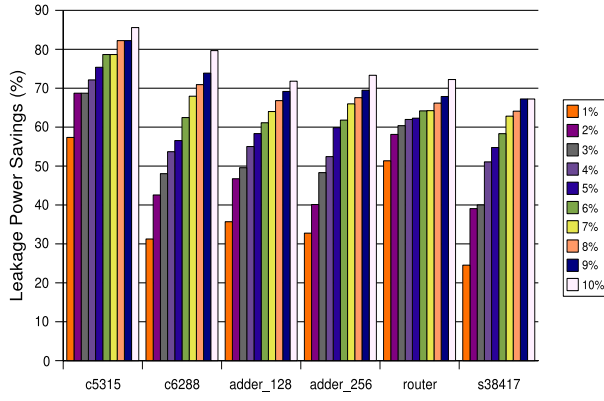


Fig. 3. Leakage Power Savings vs. α .

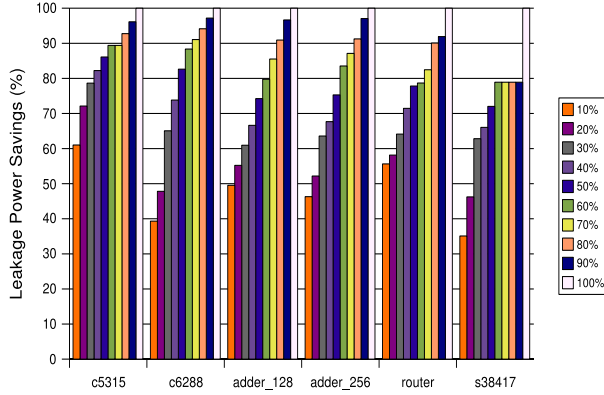


Fig. 4. Leakage Power Savings vs. W_{sleep} .

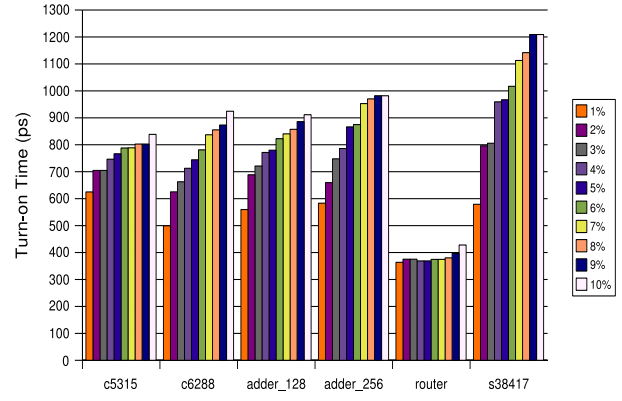


Fig. 5. Turn-on Time vs. α .

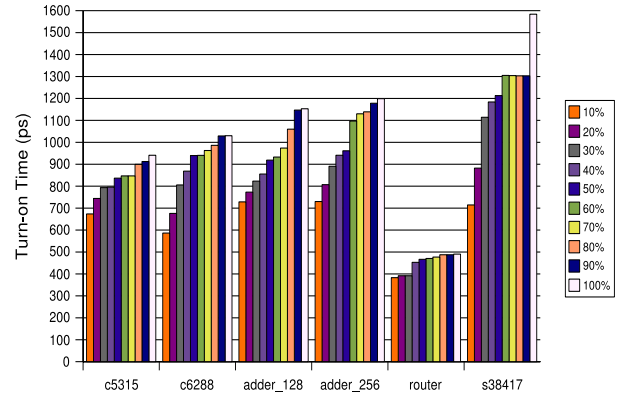


Fig. 6. Turn-on Time vs. W_{sleep} .

Regarding the PM_t parameter, we recall that this mainly depends on the amount of cells that are power-gated and on the size of the sleep transistor. Figure 5 confirms that the transition time increases as we increase the allowed performance degradation. This is due to the fact that, when we increase α , we increase the amount of cells that can be power-gated, thereby increasing the charge flow from the power-gated circuit to the ground. A similar behavior is observed in the diagram of Figure 6 for what concerns the value of PM_t as W_{sleep} changes. The transition time increases as we increase the available area. This is due primarily to the fact that a larger sleep transistor implies a higher capacitance on the virtual ground line.

As anticipated in Section II, in this paper we do not report the results for the power-mode transition energy (i.e., PM_e) due to space reasons. However, the data we have collected demonstrated that the variation of PM_e is in accordance to that of PM_t .

The trade-off curves shed a lot of light on different aspects of power-gating. For instance, let us consider the `adder_256` design. Let the required α be 5% or less. Let the requirement for PM_t be 1000ps or less. From the trade-off curves in Figure 6, we see that for such a requirement, the size of the sleep transistor should be around 50% of $W_{sleep-max}$. So, with this area requirement, from Figure 4, we see that we can power-gate and hence achieve up-to 75% leakage power savings.

V. CONCLUSIONS

Power-gating has proved to be a very efficient way to reduce leakage power in stand-by mode, but its full exploitation entails a good understanding and weighing of the costs and benefits that are attached to this technique. In this paper, we have carried out a detailed investigation of the pros and cons of power-gating by defining Figures Of Merit (FoMs) for power-gating. We have applied a power-gating methodology to a set of designs synthesized in the 65nm CMOS technology, we have studied how the FoMs vary for different design constraints, and we have derived some guidelines for orienting the designers in the use of the power-gating technology.

REFERENCES

- [1] F. Fallah, M. Pedram, "Standby and Active Leakage Current Control and Minimization in CMOS VLSI Circuits," *IEICE Trans. on Electronics, Special Section on Low-Power LSI and Low-Power IP*, Vol. E88-C, No. 4, pp. 509-519, Apr. 2005.
- [2] K. Roy, *et al.*, "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits," *Proceedings of the IEEE*, Vol. 91, No. 2, pp. 305-327, Feb. 2003.
- [3] H. Jiang, M. Marek-Sadowska, S. R. Nassif, "Benefits and Costs of Power-Gating Technique," *ICCD-05*, pp. 559-566, Oct. 2005.
- [4] K. Shi, D. Howard, "Challenges in Sleep Transistor Design and Implementation in Low-Power Designs," *DAC-06*, pp. 113-116, Jun. 2007.
- [5] A. Sathanur, A. Pullini, L. Benini, A. Macii, E. Macii, M. Poncino, "Timing Driven Row-Based Power Gating," *ISLPED-07*, pp. 104-109, Aug. 2007.
- [6] A. Sathanur, A. Calimera, L. Benini, A. Macii, E. Macii, M. Poncino, "Efficient Computation of Discharge Current Upper Bounds for Clustered Sleep Transistor Sizing," *DATE-07*, pp. 1-6, Apr. 2007.