ACKNOWLEDGMENT

The authors would like to thank the National Semiconductor Corporation for the use of their process and their friend, Soumya, for her help to test the chip.

REFERENCES

[1] K. D. T. Ngo and R. Webster, "Steady-state analysis and design of a switched-capacitor DC-DC converter," in *Proc. IEEE PESC*, 1992, pp. 378–385.
[2] D. Maksimovic and S. Dhar, "Switched-capacitor DC-DC converters for low-power on-chip applications," in *Proc. PESC*, 1999, pp. 54–59.
[3] Y. K. Ramadass and A. P. Chandrakasan, "Voltage scalable SC DC-DC converter for ultra-low-power on-chip applications," in *Proc. IEEE PE Specialists Conf.*, 2007, pp. 2353–2359.
[4] G. Patounakis, Y. W. Li, and K. L. Shepard, "A fully integrated on-chip DC-DC conversion and power management system," *IEEE J. Solid-State Circuits*, vol. 39, no. 3, pp. 443–451, Mar. 2004.
[5] K. Battacharya and P. Mandal, "A low voltage, low ripple, on chip, dual SC based hybrid DC-DC converter," in *Proc. VLSI Design Conf.*, 2008, pp. 661–666.
[6] J. Han, A. von, and J. G. C. Temes, "A new approach to reduce output ripple in switched-capacitor based step-down DC-DC converters," *IEEE Trans. Power Electron.*, vol. 21, no. 6, pp. 1548–1555, Nov. 2006.
[7] M. Dongsheng, "Robust multiple-phase switched-capacitor DC-DC, converter with digital interleaving regulation scheme," Integr. Syst. Design Lab., Univ. Arizona, Tempe, 2006.
[8] L. Hanh-Phuc, M. Seeman, S. R. Sanders, V. Sathe, S. Naffziger, and E. Alon, "A 32 nm fully integrated reconfigurable switched-capacitor DC-DC converter delivering 0.55 W/mm^2 at 81% efficiency," in *Proc. ISSCC*, 2010, pp. 210–211.
[9] P. V. R. Kumar, K. Bhattacharyya, T. Das, and P. Mandal, "Improvement of power efficiency in SC DC-DC converter by shoot-through current elimination," in *Proc. ISLPED*, 2009, pp. 81–86.
[10] P. Favrat, P. Deval, and M. J. Declercq, "A high-efficiency CMOS voltage doubler," *IEEE J. Solid-State Circuits*, vol. 33, no. 3, pp. 410–416, Mar. 1998.
[11] H. Lee and P. K. T. Mok, "Switching noise and shoot-through current reduction techniques for SC voltage doublers," *IEEE J. Solid-State Circuits*, vol. 40, no. 5, pp. 1136–1146, May 2005.
[12] A. Maiti, R. Raghavendra, and P. Mandal, "Design of a low power voltage regulator for high dynamic range of load current," *Int. J. Electron.*, vol. 94, no. 8, pp. 743–757, 2007.
[13] G. A. Rincon-Mora and P. E. Allen, "A low-voltage, low quiescent current, low drop-out regulator," *IEEE J. Solid-State Circuits*, vol. 33, no. 1, pp. 36–44, Jan. 1998.
[14] Y. Ramadass, A. Fayed, B. Haroun, and A. Chandrakasan, "A 0.16 mm^2 completely on-chip switched-capacitor DC-DC converter using digital capacitance modulation for LDO replacement in 45 nm CMOS," in *Proc. ISSCC*, 2010, pp. 208–209.
[15] M. D. Seeman and S. R. Sanders, "Analysis and optimization of switched-capacitor DC-DC converters," *IEEE Trans. Power Electron.*, vol. 23, no. 2, pp. 841–851, Mar. 2008.
[16] M. D. Seeman, "A design methodology for switched-capacitor DC-DC converters," Ph.D. thesis, Dept. Elect. Eng. Comput. Sci., Univ. California at Berkeley, Berkeley, 2009.

Maximizing Frequency and Yield of Power-Constrained Designs Using Programmable Power-Gating

Nam Sung Kim, Abhishek Sinkar, Jun Seomun, and Youngsoo Shin

***Abstract*—A large spread of leakage power due to process variations impacts the total power consumption of integrated circuits (ICs) substantially. This in turn may reduce frequency and/or yield of power-constrained designs. Facing such challenges, we propose two methods using power-gating (PG) devices whose effective width can be adjusted during a post-silicon tuning process. In the first method, we consider processors exhibiting substantial core-to-core frequency and leakage power variations while only a global voltage/frequency domain is supported. Since each core in a processor often has its own PG device, the total width each PG device and the global voltage are tuned jointly to maximize the global frequency for a given power constraint. Our experiment demonstrates that the maximum frequency of 2-, 4-, 8-, and 16-core processors is improved by 5%–21%. In the second method, we take rejected dies due to excessive leakage power. We adjust the width of PG devices such that the dies satisfy their given power constraint. Our experiment shows that 88%–98% of discarded dies violating their power constraint are recovered.**

***Index Terms*—Power constraint, power-gating devices, process variations, yield.**

I. INTRODUCTION

As CMOS technology is scaled below 65 nm, substantial variations of maximum operating frequency ($F_{\max}$) and leakage power consumption (P_{leak}) have been observed both across dies (i.e., die-to-die (D2D) variations) and within each die (i.e., within-die (WID) variations). Moreover, the increasing spatially correlated WID variations, for example, lead to considerable variations in core-to-core (C2C) $F_{\max}$ and P_{leak} as individual cores in multi-core processors are becoming very small.

Traditionally, an $F_{\max}$ constraint has determined the yield of manufactured dies. However, P_{leak} has increased exponentially with technology scaling, which has begun to affect the yield of power-constrained designs notably [1]. Although many dies are fast enough to satisfy their $F_{\max}$ constraint, they can be discarded due to excessive active P_{leak}. This becomes worse as the spread of P_{leak} (e.g., 20 × even in 0.13-μm technology [1]) and its proportion in total power consumption (P_{tot}) increases with technology scaling.

To minimize standby P_{leak}, a power-gating (PG) device, placed between an IC and its power supply rails, is commonly used [2]. A PG device is comprised of many current switches in parallel, and all the switches are turned off to cut off the power supply of the IC, thereby reducing the P_{leak}. When a PG device is turned on, however, some resistance in each constituent switch affects the *virtual* V_{DD} (VV_{DD}) (and thus both $F_{\max}$ and active P_{leak}) of an IC. On the other hand, adjusting the on-resistance (i.e., total number of enabled switches or total effective width) through a programmable PG (PPG) device, we can modulate $F_{\max}$ and active P_{leak} of an IC within a limited range after manufacturing [3].

Manuscript received September 06, 2010; revised January 14, 2011, April 27, 2011; accepted May 05, 2011. Date of publication September 15, 2011; date of current version July 19, 2012.

N. S. Kim and A. Sinkar are with the Department of Electrical and Computer Engineering, University of Wisconsin, Madison, WI 53706 USA (e-mail: nam.sung.kim@gmail.com; sinkar@wisc.edu).

J. Seomun and Y. Shin are with the Department of Electrical Engineering, Korea Advanced Institute of Technology (KAIST), Daejeon 305-701, Korea (e-mail: jseomun@dtlab.kaist.ac.kr; youngsoo@ee.kaist.ac.kr).

Digital Object Identifier 10.1109/TVLSI.2011.2163533

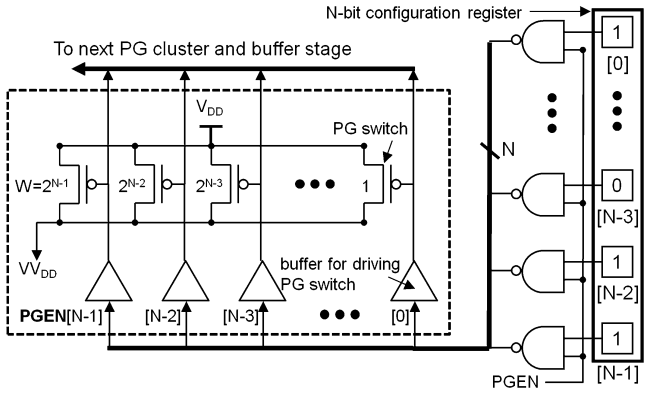

Fig. 1. Implementation of a PPG device. The PG switches and buffers in a dotted box represent one of distributed PG switch groups.

To maximize the $F_{\max}$ and yield of power-constrained designs, we first present an analysis on how P_{leak}, $F_{\max}$, and VV_{DD} in active mode can be affected by varying the total effective width (i.e., on-resistance) of a PPG device. We also analyze the impact of varying the total effective width on the power consumption of the PPG device. Second, we present a method to improve $F_{\max}$ of power-constrained ICs with multiple power-gating domains. In a multi-core processor, for example, some cores exhibit higher $F_{\max}$ but consume more active P_{leak} than other cores due to WID process variations. Meanwhile, the slowest core determines the $F_{\max}$ of a multi-core processor with a global voltage/frequency domain. To reduce the excessive active P_{leak} of the fast cores in a processor die, we adjust the total effective width of their PPG devices such that the $F_{\max}$ of all the cores match that of the slowest core. Such adjustments may reduce P_{tot} well below a power constraint. As a result, we can increase the global supply voltage (and thus $F_{\max}$) until the power constraint is just satisfied again. Finally, we provide a method to improve the yield by recovering dies that are discarded due to excessive active P_{leak} under a power constraint. We consider two different cases: 1) application-specific IC (ASIC)-type fixed P_{leak} and 2) microprocessor-type variable P_{leak} constraints. The discarded fast-but-leaky dies can be recovered by adjusting the total effective width of their PPG devices until each die can satisfy the power constraint (but within the frequency constraint).

II. PPG Device

A. Concept

Fig. 1 illustrates an implementation of a PPG device. The pMOS header (or nMOS footer) switches are connected to the *PG enable* (PGEN) signal through the NAND gates. The other input of each NAND gate is connected to a configuration bit programmed by a programmable fuse. As the value of the configuration register decreases, fewer switches will be turned on in active mode. As a result, the decreased total effective width (i.e., increased on-resistance) of the PPG device reduces VV_{DD}. This in turn reduces both $F_{\max}$ and active P_{leak} of a circuit connected to the PPG device. To provide a fine-grain control of the total effective width with a minimum number of configuration bits, we can arrange the switches as shown in Fig. 1. Each switch in turn may consist of several smaller switches connected in parallel. This allows the sum of header switch width to be proportional to the binary value of the configuration bits. This facilitates easier programming and requires less configuration bits than the original scheme proposed in [3].

The header switches and buffers within the dotted-line box in Fig. 1 is one of distributed PG switch groups, and they are the required resource to implement a conventional PG device. The PGEN propagates through multiple buffer stages, each of which is responsible to control a fraction of PG switches, in a daisy-chain fashion to minimize the rush current for turning on the PG switches [4]. Since the PGEN drives a large amount current and travels across the entire circuit IC connected to the PG device, a wide metal line (or multiple parallel metal lines) must be used; for a PPG device, the N PGEN signals derived from the original PGEN signal and the configuration bits will run across the entire circuit as a group in the same manner as a conventional PG device design. The configuration bits of the PPG device can be set after a post-manufacturing characterization of $F_{\max}$ and P_{leak}, which can be performed like any other variation compensation techniques such as adaptive body biasing (ABB) or adaptive voltage scaling (AVS).

B. Impact on Delay, Leakage, and VVDD

To analyze the impact of applying PPG on $F_{\max}$, P_{leak}, and VV_{DD} in a 32 nm technology node, a PPG device is initially sized to provide 25 mV drop from V_{DD}, which is 0.9 V at the nominal process corner and 100 °C. Under such a condition, we assume that the ratio between dynamic current (I_{dyn}) and active leakage current (I_{leak}) of an IC connected to the PPG device is 7:3 in P_{tot} [5]. We model I_{dyn} and I_{leak} with a dummy circuit as illustrated in [6]. P_{tot} is directly measured at the V_{DD} node to include the power consumption of the PPG device itself.

Fig. 2(a) and (b) show active I_{leak} and delay ($1/F_{\max}$) while the total width of disabled PG switches is varied; the figures are normalized to those of an IC in which all switches are enabled. As the fraction of disabled switches increases to 10%, 20%, and 30% at the nominal process corner, active I_{leak} decreases by 3.7%, 7.9%, and 12.7%, while the delay increases by 0.5%, 1.0%, and 1.8%, respectively. Active I_{leak} reduction is more significant at the fast process; it decreases by 5.0%, 10.3%, and 16.2% while the delay increases by 0.6%, 1.4%, and 2.3%, respectively. It is the fast corner in which the proposed technique will be mainly applied to dies, because the dies are often unnecessarily fast thereby consuming too much P_{leak}.

Fig. 2(c) and (d) show the VV_{DD} normalized to that of an IC in which all switches are enabled, as well as the proportion of power consumption of the PPG device (P_{PG}) in P_{tot}. As the fraction of disabled switches increases to 10%, 20%, and 30%, the resistance across the PPG switches increases, thereby reducing VV_{DD} by 10%, 22%, and 37% (relative to the initial 25 mV drop), respectively. This also reduces P_{dyn} of the connected IC. As the resistance increases, on the other hand, more power is consumed by the PPG switches. However, the portion of P_{PG} in P_{tot} is very small as shown in Fig. 2(b); it is only 8% at the fast corner although 50% of the PPG switches are disabled. Note that the proposed PPG device is very similar to low-drop-output (LDO) linear voltage regulators, in which the power loss due to the regulator is very small when the amount of the voltage drop is not significant.

III. $F_{\max}$ Improvement of Power-Constrained Designs Supporting Multiple Power-Gating Domains

A large design often consists of several blocks where each of them has its own PG device to minimize standby P_{leak}. As an example of a quad-core processor, three cores can be disabled using the associated PG devices when only one core is sufficient to serve workload demand. Meanwhile, some cores are faster than others due to WID C2C $F_{\max}$ variations in a multi-core processor. When all cores are forced to operate at the same global frequency, the $F_{\max}$ of the processor is determined by the slowest core. A power constraint (P_{totmax}) is often imposed when all cores are running simultaneously at a maximum sustainable performance point. This limits the increase of V_{DD} and $F_{\max}$. We use V_{DDPC} and F_{maxPC} to denote the V_{DD} and $F_{\max}$ which the processor can reach under the given power constraint.

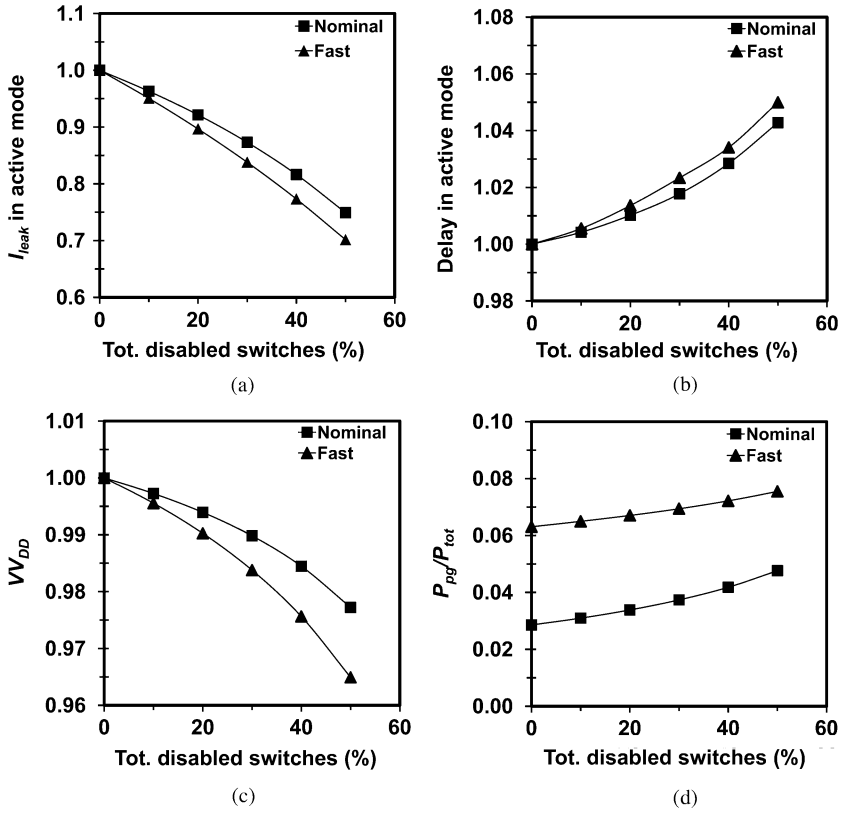

Fig. 2. Normalized (a) leakage current, (b) delay, (c) $VV_{\rm DD}$, and (d) $P_{\rm pg}/P_{\rm tot}$ versus the fraction of off PPG switches.

A. Problem Formulation

In a multi-core processor, assume that we have a PPG device per core. We can then set the configuration bits of the PPG devices such that the $F_{\max}$ of cores can be made equal to that of the slowest core. As a result, active $P_{\rm leak}$ of the cores decrease while the overall $F_{\max}$ remains unchanged. This in turn lets the $P_{\rm tot}$ of the multi-core processor be under its $P_{\rm totmax}$. Consequently, we can increase the global $V_{\rm DD}$ and $F_{\max}$ of the processor as long as power and other constraints including *maximum* $V_{\rm DD}$ ($V_{\rm DDmax}$) are not violated.

Let $VV_{\rm DD}$ of PPG domain i, $VV_{\rm DDi}$ be represented by $VV_{\rm DDi} = v_i(W_i)\cdot VV_{\rm DD}$, where W_i is the total effective width of the PPG device in domain i; v_i is a function that returns the $VV_{\rm DD}$ scaling factor of domain i for given W_i. Note that $VV_{\rm DD}$ is a strong function of W_i and its value is always lower than $VV_{\rm DDi}$ in PPG. However, if we increase the global $V_{\rm DD}$, $VV_{\rm DD}$ can become higher than its initial $V_{\rm DD}$. Note that modulating the strength of a PPG device affects the $VV_{\rm DD}$ of the corresponding domain alone while scaling the global $V_{\rm DD}$ affects the $VV_{\rm DD}$ of all the domains. The objective of the problem we address is to maximize the $F_{\max}$ of a given die while $P_{\rm totmax}$ and $V_{\rm DDmax}$ are satisfied.

Objective:

$$\text{Maximize}\,(F_{\max}\,(V_{\rm DD}, V_{DD1}, \ldots, V_{\rm DDN})). \tag{1}$$

Constraint:

$$P_{\rm tot} = \sum_{i=1}^{N} P_{\rm toti}(VV_{\rm DDi}, F_{\max}) \le P_{\rm tot\,max}, V_{\rm DD} \le V_{\rm DD\,max} \tag{2}$$

where $F_{\max i}$ and $VV_{{\rm DD}i}$ are $F_{\max}$ and $VV_{\rm DD}$ of the circuitry in P domain i, respectively; $F_{\max}$ is *min*$\{F_{\max 1}(VV_{\rm DD1}), \ldots, F_{\max N}(VV_{\rm DDN})\}$; N is the number of PPG domains; and $P_{{\rm tot}i}$ corresponds to the total power consumption that includes dynamic ($P_{{\rm dyn}i}$) and static ($P_{{\rm leak}i}$) components in PPG domain i.

B. Results

Fig. 3(a) shows the average active $P_{\rm leak}$ reduction versus the number of cores per die after the first PPG tuning step is applied to 100 die samples used in [7]. With more cores per die, we have more relative $P_{\rm leak}$ and $F_{\max}$ spread between cores as noted in [8]. This provides a better opportunity in reducing active $P_{\rm leak}$ for fast cores and improving the overall $F_{\max}$ of a die as a result. The experimental result shows that $P_{\rm leak}$ decreases by 9%–41% for 2-, 4-, 8-, and 16-core processors after the first tuning step. We assume that: 1) a processor consumes $P_{\rm tot} = P_{\rm totmax}$ at $V_{\rm DD} = 0.8$ or 0.9 V (i.e., $V_{\rm DDPC}$) and 2) $P_{\rm leak}$ is responsible for 30% of $P_{\rm totmax}$ before the first PPG tuning step. Since $P_{\rm leak}$ scales more substantially at higher voltage, processors with higher $V_{\rm DDPC}$ can provide more $P_{\rm leak}$ reduction opportunities; $V_{\rm DDPC} = 0.9$ V offers 2%–6% more $P_{\rm leak}$ reduction, potentially leading to more improvement in $F_{\max}$.

When $V_{\rm DDPC}$ is 0.8 V and $P_{\rm leak}$ is 40% of $P_{\rm totmax}$ at $V_{\rm DDPC}$, we can improve $F_{\max}$ by 5%–21% on average for 2-, 4-, 8-, and 16-core processors as shown in Fig. 3(b). The percentage of $P_{\rm leak}$ in $P_{\rm totmax}$ should also impact the $F_{\max}$ improvements since $P_{\rm leak}$ can change more dramatically than $P_{\rm dyn}$ for adjusting the PPG device and $V_{\rm DD}$. However, as illustrated in Fig. 3(b), increasing the percentage of $P_{\rm leak}$ in $P_{\rm tot}$ from 20% to 40% results in only 1%–5% difference in $F_{\max}$ improvement for 2, 4, 8, and 16 cores. This is because $P_{\rm leak}$ scales at a similar rate as $P_{\rm dyn}$ when $V_{\rm DD}$ is around the $V_{\rm DDPC}$ region.

For a given $P_{\rm totmax}$ constraint, the $F_{\max}$ improvement can be affected by $V_{\rm DDPC}$ in two ways: 1) we can have more $P_{\rm leak}$ scaling at

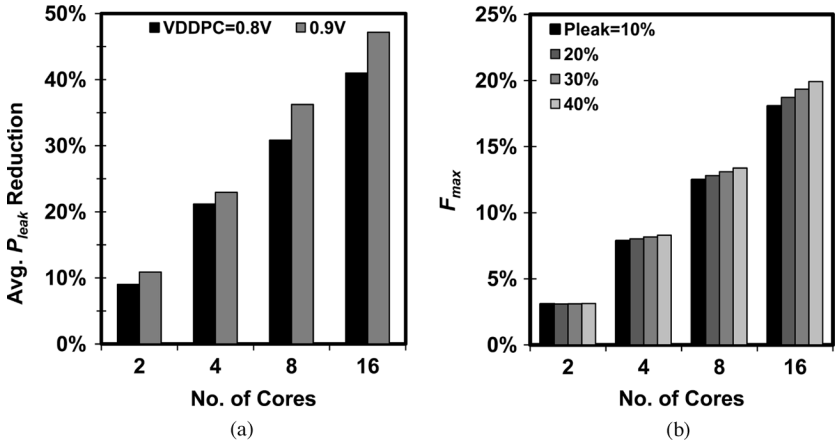

Fig. 3. (a) Average $P_{\rm leak}$ reduction after the first PPG tuning step. (b) $F_{\rm max}$ improvement after the second PPG + VDD tuning step. In (a) and (b), $P_{\rm leak}$ is responsible for 30% of $P_{\rm totmax}$ and $V_{\rm DDPC}$ is 0.8 V, respectively.

higher $V_{\rm DDPC}$ as shown in Fig. 3(a), but 2) we have less power headroom to improve $F_{\rm max}$ because a design with lower $V_{\rm DDPC}$ is assumed to have higher power than one with higher $V_{\rm DDPC}$ at the same $V_{\rm DD}$. Hence, the difference of $V_{\rm DDPC}$ should not affect the average $F_{\rm max}$ improvement significantly; $V_{\rm DDPC} = 0.9$ V provides less than 1% difference in $F_{\rm max}$ improvement for 2-, 4-, 8-, and 16-core processors even when $P_{\rm leak}$ is 40% of $P_{\rm totmax}$.

Finally, in the experiments shown in Fig. 3, we do not constrain the voltage drop across the PPG devices. However, the increasing voltage drop may need to be limited to consider noise issues and reliability problems at low voltage. Thus, we repeat the same experiment while limiting the voltage drop to 100 mV, which leads to 0%–3% less $F_{\rm max}$ improvement depending on the number of cores per chip or the initial fraction of $P_{\rm leak}$ in $P_{\rm tot}$.

IV. D2D-Variation-Aware Yield Improvement of Power-Constrained Designs

Since dies in a fast corner have shorter $L_{\rm eff}$ and lower $V_{\rm th}$ for their transistors, they exhibit much more $P_{\rm leak}$. As a result, they violates their $P_{\rm totmax}$ constraint, and thus the yield is reduced. In this section, we present a method to improve the yield of power-constrained designs using PPG devices for two scenarios: designs with: 1) fixed and 2) variable $F_{\rm max}$ and $P_{\rm leak}$ constraints, respectively. To improve the yield of dies violating the $P_{\rm totmax}$ constraint, their $P_{\rm leak}$ can be reduced using PPG such that their $P_{\rm totmax}$ constraint is satisfied, usually for fast-but-leaky dies, with no or minimum impact on $F_{\rm max}$.

A. Designs With Frequency Target: Fixed $P_{\rm leak}$ Constraint

ASICs are usually designed for *frequency* (F) and *power consumption* (P) targets. Specifically, $F_{\rm max}$ and $P_{\rm tot}$ of each die must be satisfied while the yield is maximized.

Objective:

$$\text{Maximize}(\text{Yield}(F_{\rm max}, P_{\rm tot})). \tag{3}$$

Constraint:

$$P_{\rm tot} = P_{\rm dyn}(VV_{\rm DD}) + P_{\rm leak}(VV_{\rm DD}) \le P, F_{\rm max} \ge F. \tag{4}$$

In this section, $F_{\rm max}$ is not an operating frequency but a maximum frequency that can be achieved by a particular die; unless F exceeds $F_{\rm max}$, dies will operate at F. Fast-but-leaky dies often satisfy the frequency constraint but not the power constraint. Since $F_{\rm max}$, $P_{\rm dyn}$, and $P_{\rm leak}$ are functions of $VV_{\rm DD}$, such dies can be forced to satisfy the power constraint by adjusting $VV_{\rm DD}$ (i.e., resistance) of PPG devices. In (4), $P_{\rm dyn}$ can be considered to be constant if: 1) operating frequency is fixed at F and 2) $VV_{\rm DD}$ after adjusting the PPG takes the value that is not significantly different from the initial $VV_{\rm DD}$. This allows us to consider the power constraint in (4) simply as a leakage target

$$P_{\rm leak}(VV_{\rm DD}) \le P' \tag{5}$$

where P' is the difference between P and $P_{\rm dyn}$ (i.e., $P_{\rm leak}$ budget). In fact, (5) is conservative since $P_{\rm dyn}$ becomes smaller after adjusting $VV_{\rm DD}$. In other words, (4) is always satisfied if (5) is satisfied.

B. Designs With Frequency Binning: Variable $P_{\rm leak}$ Constraint

In Section IV-A, we considered a fixed frequency target, which is typical for ASIC designs. In high-performance processor designs, on the other hand, a list of frequency targets, $F_1 < F_2 < \cdots < F_N$, are provided. The $F_{\rm max}$ of a die is compared to the targets and then it is put into an appropriate bin, i.e.,

$$f(F_{\rm max}) = \left\{ \begin{array}{c} F_i \text{ if } F_i \le F_{\rm max} < F_{i+1}, i = 1, 2 \ldots, N-1 \\ F_N \text{ otherwise} \end{array} \right\} \tag{6}$$

where f is a function that assigns an operating frequency; this process is called frequency binning. Different bins are associated with different $P_{\rm dyn}$. Since the sum of $P_{\rm dyn}$ and $P_{\rm leak}$ has to be smaller than a fixed power constraint, each bin has its own $P_{\rm leak}$ constraint.

We continue to apply PPG to improve yield as we did in Section IV-A, except that we now have varying $P_{\rm leak}$ constraints. Since the bins of higher frequencies are preferred, our new objective is to maximize the operating frequency of each die while a given power constraint is satisfied.

Objective:

$$\text{Maximize}(f(F_{\rm max})). \tag{7}$$

Constraint:

$$P_{\rm tot} = P_{\rm dyn}(f(F_{\rm max}), VV_{\rm DD}) + P_{\rm leak}(VV_{\rm DD}) \le P. \tag{8}$$

Although $P_{\rm dyn}$ is dependent on $VV_{\rm DD}$, we assume that $P_{\rm dyn}$ is a function of $F_{\rm max}$ alone since $P_{\rm dyn}$ is constant for a fixed frequency. Let the nominal $VV_{\rm DD}$ be the $VV_{\rm DD}$ before we change PPG configuration bits. Then (8) can be simplified to

$$P_{\rm tot} \approx g(VV_{\rm DD}) \cdot P_{\rm dynnom} + h(VV_{\rm DD}) \cdot P_{\rm leaknom} \le P \tag{9}$$

where $g(VV_{\rm DD})$ is a function that returns the $F_{\rm max}$ (normalized to the nominal $F_{\rm max}$ as a reference) of a particular die at $VV_{\rm DD}$; $P_{\rm dynnom}$ is

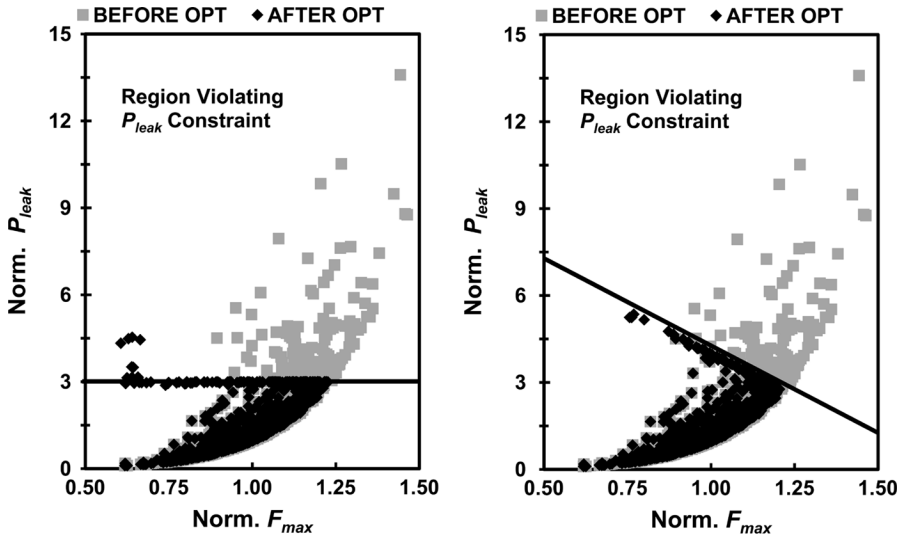

Fig. 4. Normalized $P_{\rm leak}$ and $F_{\rm max}$ distribution before and after applying the proposed method to maximize the yield under (a) fixed and (b) variable $P_{\rm leak}$ constraints for ISCAS85 C432.

the $P_{\rm dyn}$ at the nominal $F_{\rm max}$; $h(VV_{\rm DD})$ is a function that returns $P_{\rm leak}$ of a die at $VV_{\rm DD}$; and $P_{\rm leaknom}$ is the $P_{\rm leak}$ at the nominal $VV_{\rm DD}$.

C. Results

To assess how much yield can be improved, we take ISCAS85 and some of OpenCore circuits. The header switches in Fig. 1 are all connected (i.e., all the configuration bits are initially set to "1"). We perform Monte-Carlo simulations using SPICE with a predictive 45-nm technology model to apply D2D and WID process variations to each die sample of a circuit; we apply 3σ variations (i.e., 0.4 V and 10 nm) to the nominal $V_{\rm th}$ and $L_{\rm eff}$ of nMOS and pMOS, respectively. Initially, PPG switches at the nominal corner are sized such that the maximum voltage drop across the switches does not exceed 50 mV for the *peak current consumption* ($I_{\rm DDmax}$) of each circuit; $I_{\rm DDmax}$ is estimated by applying 10 000 vectors at 100 °C and selecting a pair of vectors causing the worst-case current consumption. For each die sample of a circuit, we measure $P_{\rm leak}$ and $F_{\rm max}$ using SPICE as we keep varying the configuration bits until the die satisfies the (5) and (9).

The $P_{\rm leak}$ and $F_{\rm max}$ of each ISCAS85 C432 die are shown as scatter plots in Fig. 4 for: (a) fixed and (b) variable $P_{\rm leak}$ constraints. In Fig. 4(a), first, P' in (5) is set to $3\times$ of the *nominal* $P_{\rm leak}$ ($P_{\rm leaknom}$) corresponding to $P_{\rm leak}$ of a die without process variations. Second, F is set to -3σ of the *nominal* $F_{\rm max}$ ($F_{\rm maxnom}$), where σ is standard deviation of $F_{\rm max}$ in 1000 dies. The accepted dies, satisfying (4), are the ones below $3 \times P_{\rm leak}$ boundary; 116 dies in the example circuit are rejected before the optimization since they do not satisfy (4). For each of rejected dies due to excessive $P_{\rm leak}$, we tried to change its configuration bits (by setting some of them to 0) so that it is brought under the $3 \times P_{\rm leak}$ boundary. The results are shown as another scatter plots in Fig. 4(a). Only 12 dies are rejected now after the method is applied, and the yield is improved by recovering 104 dies out of 116 rejected ones.

In Fig. 4(b), P is equal to $P_{\rm dynnom} + 4 \times P_{\rm leaknom}$, where $P_{\rm dynnom}$ takes about 60% of P. Before we apply the proposed method to maximize the yield, the dies above the diagonal line are discarded because their $P_{\rm tot}$, as described by (9), exceeds P. As explained earlier, the dies exhibiting higher $F_{\rm max}$ have less $P_{\rm leak}$ budget. We change the configuration bits of each of rejected dies such that it is brought under the diagonal line. The results are shown as diamond shapes in Fig. 4(b). Initially 106 dies are rejected from C432. However, after the method is applied, only 1 die is rejected, recovering almost all the dies in this particular example.

We repeat the same experiments for a subset of ISCAS85 (C499, C880, C1355, C1908, C2670, and C3540) and OpenCore (i2c, pcm_slv, ps2, sasc) circuits, as well as a synthetic circuit that mimics $P_{\rm dyn}$, $P_{\rm leak}$, and $F_{\rm max}$ of a big microprocessor. On average, we could recover 88% and 90% of discarded fast-but-leaky dies when the fixed $P_{\rm leak}$ constraints are set to $3\times$ and $4 \times P_{\rm leaknom}$, respectively. Relaxing the fixed $P_{\rm leak}$ constraint yields fewer initial violations, but it also provides more opportunities to recover more discarded dies, which results in a similar or higher percentage of yield improvement for the given $P_{\rm leak}$ constraints.

To set the power constraint P for a variable $P_{\rm leak}$ constraint, we assume that $P_{\rm totmax}$ is $P_{\rm dynnom} + 4 \times P_{\rm leaknom}$ at the nominal $F_{\rm max}$, and that the ratios between $P_{\rm dynnom}$ and $4 \times P_{\rm leaknom}$ at the nominal $F_{\rm max}$ are: 1) 6:4 and 2) 7:3 to analyze the sensitivity on the percentage of $P_{\rm leak}$, in P. On average, we recovered 98% of the discarded dies when the $P_{\rm leak}$ constraints at the $F_{\rm maxnom}$ point are $0.3 \times P$ and $4 \times P$, respectively. Note that less $P_{\rm leak}$ budget (e.g., $P_{\rm leak} = 0.3 \times$ P at the nominal $F_{\rm max}$) incurs more violations than more $P_{\rm leak}$ budget ($P_{\rm leak} = 0.4 \times$ P at the nominal $F_{\rm max}$) before applying the method due to less $P_{\rm leak}$ headroom for $P_{\rm leak}$ variations at higher $F_{\rm max}$.

Finally, when we limit the voltage drop across PPG devices to 100 mV, the proposed technique still recovers 1) 24%–71% and 2) 29%–88% of discarded ISCAS85 and OpenCore dies for 1) fixed and 2) variable $F_{\rm max}$ and $P_{\rm leak}$ constraints, respectively. The circuits with low activity factors lead to very high $P_{\rm leak}$ in $P_{\rm tot}$ while D2D variations can increase $P_{\rm leak}$ by orders of magnitudes. Thus, limiting the voltage drop across the PPG devices also limits the maximum $P_{\rm leak}$ decrease for fast-but-leaky dies, impacting the percentage of recoverable dies. However, when we consider the synthetic circuit whose nominal $P_{\rm leak}$ fraction in $P_{\rm tot}$ is similar to commercial processors, 88%–93% of dies can be recovered although the voltage drop is limited to 100 mV.

V. Conclusion

We have proposed two methods to improve the maximum operating frequency and yield of power-constrained designs by using

programmable power-gating devices. The first method improves the maximum operating frequency of designs implemented with multiple power-gating domains by adjusting the strength of power-gating devices, domain by domain, which is followed by scaling global supply voltage for higher operating frequency. Our experimental results showed that the proposed method improved the maximum operating frequency by 5%–21% for 2-, 4-, 8-, and 16-core processors.

The second method recovers discarded dies due to excessive active leakage power; a necessary amount of active leakage power is reduced by the strength of power-gating devices until the dies are brought back into the acceptable operating region. To demonstrate the effectiveness of the method, we examined two different design scenarios: 1) ASIC-type fixed leakage power and 2) processor-type variable leakage power constraints. Our experiments demonstrated that about 88% and 98% of discarded dies could be recovered by the proposed methods in two design scenarios, respectively.

References

[1] S. Borkar, T. Karnik, and V. De, "Design and reliability challenges in nanometer technologies," in *Proc. Design Autom. Conf. (DAC)*, 2004, pp. 75–75.

[2] K. Shi and D. Howard, "Challenges in sleep transistor design and implementation in low-power design," in *Proc. Design Autom. Conf. (DAC)*, 2006, pp. 113–116.

[3] H. S. Deogun, D. Sylvester, R. Rao, and K. Nowka, "Adaptive MTCMOS for dynamic leakage and frequency control using variable footer strength," in *Proc. SOC Conf. (SOCC)*, 2005, pp. 147–150.

[4] Synopsys, Mountain View, CA, "Synopsys power-gating design methodology based on SMIC 90 nm process," 2010. [Online]. Available: http://www.synopsys.com.cn/information/snug/2007-2008-collection/synopsys-power-gating-design-methodology-based-on-smic-90nm-process

[5] K. Aygun, M. J. Hill, K. Eilert, K. Radhakrishnan, and A. Levin, "Power delivery for high-performance microprocessor," *Intel Technol. J.*, vol. 9, no. 4, pp. 273–283, Nov. 2005.

[6] A. Sinkar and N. S. Kim, "AVS-aware power-gate sizing for maximum performance and power efficiency of power-constrained processors," in *Proc. Asia South Pacific Design Autom. Conf. (ASP-DAC)*, 2011, pp. 725–730.

[7] N. S. Kim, J. Seomun, A. Sinkar, J. Lee, T. H. Han, K. Choi, and Y. Shin, "Frequency and yield optimization using power gates in power-constrained designs," in *Proc. Int. Symp. Low Power Electron. Design (ISLPED)*, 2009, pp. 121–126.

[8] S. R. Sarangi, B. Greskamp, R. Teodorescu, J. Nakano, A. Tiwari, and J. Torrellas, "VARIUS: A model of process variation and resulting timing errors for microarchitects," *IEEE Trans. Semicond. Manuf.*, vol. 21, no. 1, pp. 3–13, Feb. 2008.

Functional Test-Sequence Grading at Register-Transfer Level

Hongxia Fang, Krishnendu Chakrabarty, Abhijit Jas, Srinivas Patil, and Chandra Tirumurti

***Abstract*—We propose output deviations as a surrogate metric to grade functional test sequences at the register-transfer level without explicit fault simulation. Experimental results for the open-source Biquad filter core and the Scheduler module of the Illinois Verilog Model show that the deviations metric is computationally efficient and it correlates well with gate-level coverage for stuck-at, transition-delay and bridging faults. Results also show that functional test sequences reordered based on output deviations provide steeper gate-level fault coverage ramp-up compared to other ordering methods.**

***Index Terms*—Defect, functional test, output deviation, register-transfer level (RTL), test grading.**

I. Introduction

Functional test is commonly used in industry to target defects that are not detected by structural tests [1]. An advantage of functional test is that it avoids overtesting since it is performed in normal functional mode. In contrast, structural test is accompanied by some degree of yield loss [2].

Given a large pool of functional test sequences (for example, design verification sequences), it is necessary to develop an efficient method to select a subset of sequences for manufacturing testing. Since functional test sequences are much longer than structural tests, it is time-consuming to grade functional test sequences using traditional gate-level fault simulation methods.

The evaluation of functional test sequences is a daunting task if we consider the sheer number of cycles one may have to simulate to evaluate the fault coverage (assuming we know what fault model we are going to grade against). For example, consider a functional test sequence that is equivalent to one second of runtime on a processor with a 3 GHz clock frequency. This means that we have to simulate 3 billion cycles to evaluate the fault coverage. Even for stuck-at or transition fault models, simulation of the order of a billion cycles on a small fault sample on a reasonably large server farm can take months to complete. Furthermore, for system-level tests that are often created to catch circuit marginality and timing (speed path) errors, if we try to grade these tests on delay fault models, the time taken would be orders of magnitude more than the time needed to grade on transition or stuck-at fault models.

To quickly estimate the quality of functional tests, a high-level coverage metric for estimating the gate-level coverage of functional tests is proposed in [3]. However, this approach requires considerable time and resources for the extraction of coverage objects. In particular, experienced engineers and manual techniques are needed to extract the best

Manuscript received December 22, 2010; revised April 22, 2011; accepted July 04, 2011. Date of publication September 12, 2011; date of current version July 19, 2012. This work was supported in part by the Semiconductor Research Corporation under Contract 1588. A preliminary version of this paper was published in the Proceedings on IEEE VLSI Test Symposium, pp. 264-269, 2009.

H. Fang and K. Chakrabarty are with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708 USA (e-mail: hf12@ee.duke.edu; krish@ee.duke.edu).

A. Jas, S. Patil, and C. Tirumurti are with the Intel Corporation, Austin, TX 78746 USA (e-mail: abhijit.jas@intel.com; srinivas.patil@intel.com; chandra.tirumurti@intel.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVLSI.2011.2163651