21st International Conference on VLSI Design

Power Reduction of Functional Units considering Temperature and Process Variations

Deepa Kannan, Aviral Shrivastava
Compiler and Microarchitecture Laboratory
School of Computing and Informatics
Arizona State University, Tempe, AZ 85281
{deepa.kannan, aviral.shrivastava}@asu.edu

Sarvesh Bhardwaj, Sarma Vrudhula
VLSI Electronic Design Automation Laboratory
School of Computing and Informatics
Arizona State University, Tempe, AZ 85281
{sarvesh.bhardwaj, vrudhula}@asu.edu

Abstract

Continuous technology scaling has resulted in an increase in both, the power density as well as the variation in device dimensions (process variations) of the manufactured processors. Both power density and process variations have a significant impact on the leakage power. Therefore, power optimization techniques should be sensitive to the variation in leakage power due to both temperature as well as process variations. Operation to Functional Units Binding Mechanism (OFBM) is the mechanism to dynamically issue operations to Functional Units (FUs) in superscalar processors. We propose a Leakage-Aware OFBM (LA-OFBM), which is both temperature and process variation aware. Our experimental results demostrate that LA-OFBM reduces the mean and standard deviation of the total energy consumption of ALUs by 18%, and 46% respectively, as compared to the traditional OFBM, without any performance penalty.

1. Introduction

The ever increasing performance demands from high-end microprocessors have been one of the most important forces behind continuous technology scaling for the past four decades. As a consequence of incessant technology scaling, leakage power has become a major component of the total power budget of the processors developed in nano-scale CMOS technologies. In fact, according to [19], leakage power in 65 nm technology amounts to almost 40% of modern microprocessors' total power budget.

Leakage power, in contrast to dynamic power is highly sensitive to variations in the operational temperature. Infact, leakage of a CMOS gate increases exponentially with increase in temperature. According to [15] a 10^oC rise in temperature at 35^oC will result in leakage currents going up by 126%.

Another important consequence of technology scaling is a significant loss of control in the lithography as well as channel doping steps during the manufacturing, resulting in large variations in the characteristics of the manufactured devices (ITRS 2003) [7]. This phenomenon, called process variation has significant impact in terms of power consumption, yield, reliability, and design processes. Since leakage power has an exponential dependence on device characteristics, even small variations in the device characteristics result in large variations in the leakage power. Leakage power being a major contributor to the total power consumption in present day microprocessors, the variations in leakage power results in a significant variation in the processor power consumption across dies. The impact of process variations was demonstrated recently in [1] wherein the authors demonstrated 20X variation in leakage power for a 1.3X performance variation in Intel processors.

Given the significant and increasing impact of process variations and temperature variations on the power consumption of processors, it is important for power optimization techniques to consider both these factors, and aim at reducing both the power consumption, and the variation in the power consumption across dies.

Functional units (FUs) such as ALUs and multipliers are significant contributors to the total energy consumption of the processor [4, 3]. In addition, owing to high energy densities of FUs, the effect of process variations on the leakage of FUs is amplified by the exponential dependence of leakage on temperature. *Consequently, reducing both the total power, and the variation in the total power consumption of FUs is an important problem.*

OFBM (Operation to FU Binding Mechanism) is the mechanism by which the ready operations are issued to the FUs in a superscalar processor. Traditional OFBMs statically bind the operations to FUs, without considering the process and temperature variations. They may therefore issue operations to FUs that will leak more, due to either of

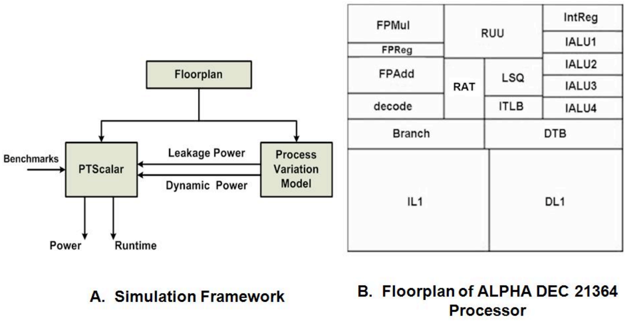
1063-9667/08 $25.00 © 2008 IEEE
DOI 10.1109/VLSI.2008.81

the two effects. Traditional OFBMs therefore result in high FU power, as well as high variation in FU power.

In this paper, we propose to introduce leakage sensors in each FU, and develop LA-OFBM (Leakage-Aware Operation to FU Binding Mechanism) that is cognizant of both process and temperature variations through the leakage sensor. Our experimental results show that LA-OFBM reduces: (i) the average (mean) ALU energy consumption by 18% and (ii) the variation in the total ALU energy consumption (standard deviation) across 1000 die samples by 46%, as compared to previous OFBMs, without any performance overhead.

2. Experimental Setup

Our simulation framework is depicted in Fig. 1A. We perform our experiments on the ALPHA DEC 21364 processor. This is a 4-wide superscalar processor, whose floorplan is shown in Fig. 1B.



A. Simulation Framework

B. Floorplan of ALPHA DEC 21364 Processor

Figure 1. Simulation Setup

The power, performance and temperature modeling of the alpha processor is done using a modified version of sim-outorder of the PTScalar toolset. PTScalar is a coupled power and thermal simulator built over SimpleScalar [2]. The floorplan of the processor, leakage and dynamic powers of all units in the processor are given as input to PTscalar, which then simulates the benchmark to estimate the power, performance and temperature of all the units. We execute several benchmarks from the MiBench [8], and Spec 2000 [16] suite.

Process Variation Model: We model the variations in device features (gate length and threshold voltage) using the stochastic process corresponding to gate length using the Karhunen-Loève Expansion proposed in [5]. The floorplan of the processor is given as an input to the process variation model to accurately model the spatial correlation in the device parameters. The process variation model generates the dynamic power and leakage power values for all the units in the processor corresponding to one die. We generate 1000 such die samples, which are fed into the PTScalar for power, performance, and temperature modeling. The power numbers are scaled to correspond to a 45nm technology.

3. Related Work

The impact of process variations and temperature on leakage, has been extensively researched, and the importance of leakage reduction in FUs has been recognized for long. However, to the best of our knowledge, there are no prior OFBMs, which are aware of temperature and process variations.

3.1. Leakage Reduction of Functional Units

Due to their dominant transistor budget in a processor, earlier research focused on leakage reduction of storage structures [9, 12, 6]. However, recognizing that FUs can be the spots of highest leakage density, recent efforts have focused on mitigating leakage in the FUs.Power gating has been the most researched technique for reducing the leakage of FUs. [10] proposed a mechanism to reduce leakage through power-gating of FUs. [14] detects the idle intervals of FUs dynamically to power gate the FUs and thereby reduce leakage. Talli et al [18] use the profile information to identify the idle periods of the functional units and use the compiler to issue corresponding on/off instructions.

3.2. Operation-to-FUs Binding Mechanisms

Several thermal-aware scheduling approaches to balance temperature distribution across the functional units of a VLIW architecture are proposed in [13]. This is one of the prior works that is closely related to ours. Our approach is different from theirs in the sense that ours is a microarchitecture level binding mechanism unlike their compiler-based approaches. The approaches proposed in [13], though are simple, require recompiling the application, which may not be desirable/possible. The applicability of those techniques are limited by the fact that they insist on the availability of the source code for analysis and recompilation. Another significant difference in our approach is that the technique we propose are for superscalar processors which has the scalability to be extended to other processors, as compared to their techniques which are only for VLIW processors. Also, their techniques do not take into account the process variations. There exists several other such scheduling schemes proposed for functional units. However, no prior work on operation to FU binding mechanisms considers the dependency of leakage power on both temperature and process variations.

4. Prior OFBMs

4.1. Fixed Priority OFBM (FP-OFBM)

Operation of FU binding mechanism (OFBM) is the technique to issue the ready operations to FUs. In the absence of any process or temperature variations, all OFBMs are the same. Consequently a direct mapping or FP-OFBM is usually employed in all existing processors. In FP-OBM, FUs are assigned static priorities at design time, and the priorities do not change over time. In every cycle, ready instructions are distributed to the FUs in order of their priority. A lower priority FU will get an operation iff FUs with higher priority also get an operation. FP-OFBM results in a very high activity in the highest priority FU, increasing it's temperature and leakage.

4.2. Load Balancing OFBM (LB-OFBM)

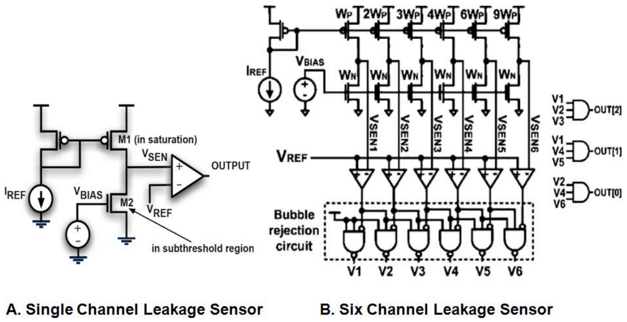
Recognizing the impact of temperature variations on the leakage of the FUs, [13] observed that FP-OFBM causes a skew in the activity of ALUs, i.e., higher priority ALU gets much more operations than a lower priority one. As a result the ALU with the highest priority leaks a lot. Consequently, authors in [13] proposed to reduce leakage by distributing the activity equally among the ALUS, and proposed Load Balancing OFBM (LB-OFBM). In LB-OFBM, operations are issued to FUs in a round robin fashion and since all FUs are equally active, the leakage of any one FU does not rise significantly more than the other FUs.

5. Our Approach: Leakage Aware OFBM (LA-OFBM)

We propose LA-OFBM to reduce the leakage energy and therefore the total energy of processors. To achieve this, we introduce a leakage sensor in each FU, and issue operations to the FUs based on the leakage information of the FUs. In the LA-OFBM, the sensors within the ALUs are used to accurately detect leakage and set the ALU priorities dynamically. The leakage sensor values are continuously read and the FU priorities are updated to be in the decreasing order of increasing leakage. Since the temperature of the ALUs vary over time, the priorities assigned to the ALUs will also change dynamically. LA-OFBM is therefore both process and temperature variations aware OFBM.

5.1. Introducing Leakage Sensor in FUs

We propose to introduce the leakage sensor proposed by Kim et al. [11] inside each FU and continuously measure the FU leakage during the chip operation. A single channel leakage sensor is shown in Fig. 2A. The bias circuits for generating I_{REF} and V_{BIAS} are process variation insensitive. M2 is the only transistor that is sensitive to the variation in leakage of the ALU due to the impact of temperature and process variations. Therefore, the accuracy of the leakage sensor itself is not affected by process and temperature variations.We explicitly model the area, power and inaccuracy introduced when converting the leakage sensor in our experimental setup. The overhead of using leakage sensors accounts to around $3-4\%$ reduction in the total power savings obtained using our LA-OFBM.



A. Single Channel Leakage Sensor

B. Six Channel Leakage Sensor

Figure 2. Single and six Channel Leakage sensors [11]

5.2. Leakage Sensor Placement

If the leakage measured by the sensor is not a good estimate of the total leakage of the ALU, we might not get the correct ordering of the ALUs in terms of their leakage power. This could potentially result in instructions being bound to a higher leakage ALU instead to the lowest leakage ALU, thus eliminating the power savings obtained using our approach. To find a good location for the leakage sensor, we compared the leakage of a device located at various locations (x_i, y_i) inside the ALU, and the average leakage of the ALU ($I_{av} = I_{S,T}/N$). We found that mean of the percentage difference between the average ALU leakage and the leakage of a device located at the center of the ALU for a sample of 1000 dies to be less than 1%. The maximum percentage error over the same set of samples was 7%. Thus a single leakage sensor located at the center of the ALU can provide accurate estimation of the leakage power of the entire ALU.

5.3. Architecture Model

In the LA-OFBM, whose architecture model is shown in Fig. 3, operations are issued to the FUs based on their leakage information. The sensors within the ALUs are used to accurately detect leakage and set the ALU priorities dynamically in the FU Priority Updater. The leakage sensor values

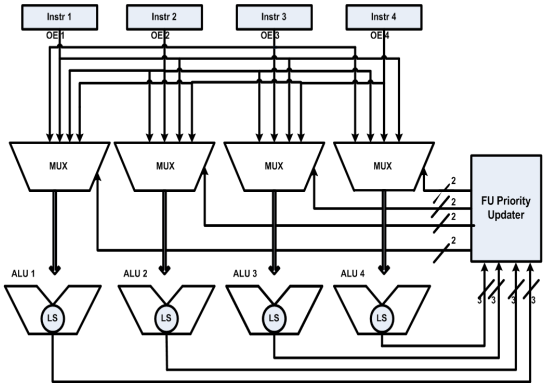


Figure 3. OFBM in 4-issue superscalar processor

are continuously read and the FU priorities are updated after every 10,000 cycles, to be in the decreasing order of increasing leakage. We introduce four 4-to-1 line multiplexors in the operation issue path, to select the ALUs to which the incoming instructions are to be issued. Since the temperature of the ALUs vary over time, the priorities assigned to the ALUs will also change dynamically. LA-OFBM is therefore both process and temperature variations aware OFBM.

Microarchitectural Overheads: We accurately model the impact of microarchitectural enhancements structurally in PTScalar. The leakage sensor is very small, only a few gates, and is not in the critical path of the ALU. Therefore there is no performance impact of the sensor. The multiplexors lie in the critical path of execution, they might cause some extra delay. However this is very small, and in our experiments, we observe that it can be accommodated in the cycle time slack. We synthesized the multiplexors and the FU Priority Update logic using Synopsys design compiler [17]. The energy overhead of multiplexors and the FU priority Updater is less than 0.75 μJ which is very small as compared to the 500 μJ energy of all the 4 ALUs. But we included both their leakage and dynamic powers in the power computation using PTscalar in all our simulations.

6. Experimental Results

6.1. FP-OFBM has a higher total energy consumption as well as variation in total energy

Fig. 4 plots the total energy consumption of all the ALUs in each of the 1000 die samples for FP-OFBM, for the representative susan corners benchmark. It can be observed that due to process variations, there can be upto 25% difference in the total ALU energy consumption between the lowest and the highest power dies.

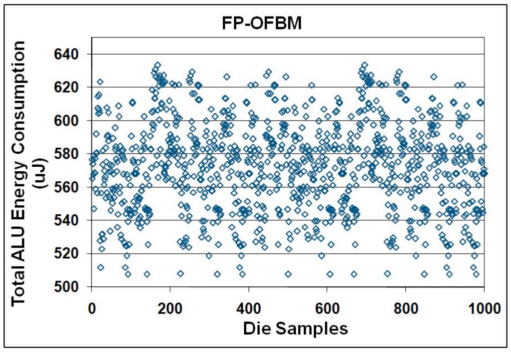


Figure 4. Total ALU Energy consumption for 1000 die samples using FP-OFBM

6.2. LB-OFBM increases total energy consumption but reduces variation in total energy

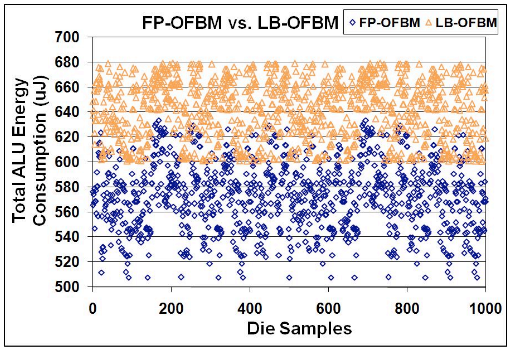


Figure 5. Total ALU Energy consumption for 1000 die samples using LB-OFBM and FP-OFBM

Fig. 5 plots the total energy consumption of all the ALUs in each of the 1000 die samples, for FP-OFBM and LB-OFBM , for susan corners benchmark. LB-OFBM shows 13% increase in the mean leakage, but a reduction of 15% in the variation in the total energy of the ALUs, as compared to FP-OFBM.

Fig. 6 plots the mean of the ALU energy consumption computed over 1000 sample dies, normalized to FP-OFBM, for all the OFBMs, for all the 10 benchmarks. The last set of bars plot the average ALU power reduction over all the benchmarks. This plot shows that the difference in mean energy consumptions between all the OFBMs is consistent over benchmarks.

The increase in the total energy consumption by the LB-OFBM is an important and counter-intuitive result. This is

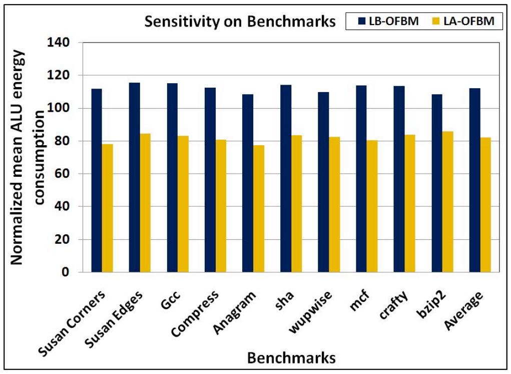


Figure 6. Total ALU energy reduction is consistent across benchmarks

because, it could be argued that FP-OFBM will concentrate the ALU activity on the lowest priority ALU, and increase the temperature of that ALU, which in turn would result in higher energy dissipation. However this is not the observation. On closer investigation, we found that the cooling efficiency improves with the increase in temperature. Therefore after some time it becomes very difficult to increase the temperature. Thus issuing more instructions to the same FU may not increase the temperature and therefore leakage much.

6.3. LA-OFBM reduces total energy consumption

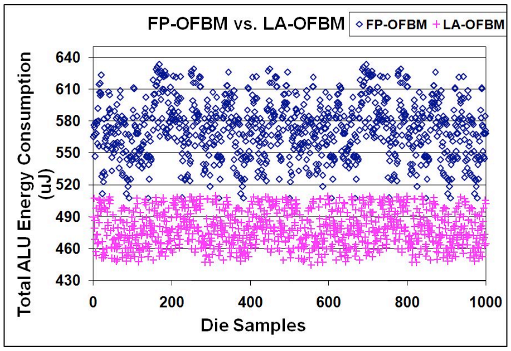


Figure 7. Total ALU Energy Consumption for FP-OFBM and LA-OFBM

Fig. 7 plots the total energy consumption of the ALUs for the FP-OFBM (baseline) and LA-OFBM techniques for 1000 die samples, for susan corners benchmark. The first observation we make from this figure is that all the LA-OFBM points are lower than the FP-OFBM points. As compared to the FP-OFBM, LA-OFBM reduces the total energy consumption of the ALUs by 18%. In terms of leakage energy alone, the LA-OFBM decreases the total leakage energy of all the ALUs by 44% as compared to FP-OFBM. Fig. 6 further bolsters our claim by demonstrating that LA-OFBM consistently reduces the energy consumption across the benchmark spectrum.

6.4. LA-OFBM reduces variation in total energy

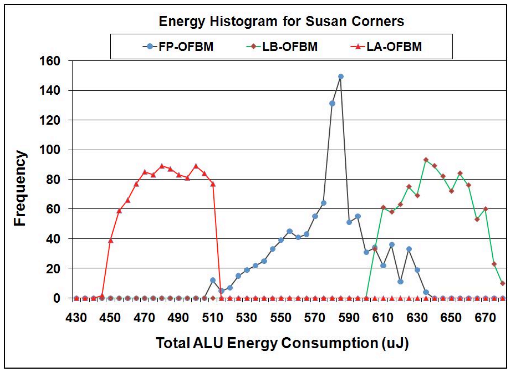


Figure 8. Variation in total energy consumption

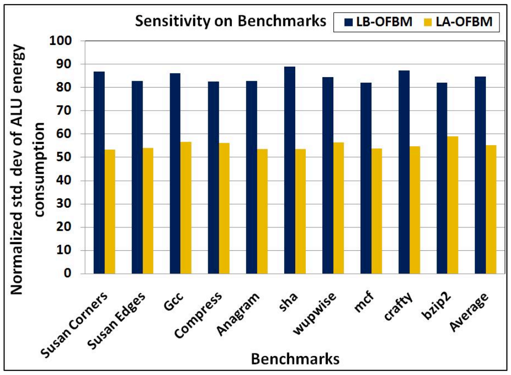


Figure 9. Variation in total ALU energy consumption is consistent across benchmarks

Another observation that we make from Fig. 7 is that the width of the vertical band in which points of LA-OFBM lie is lesser than the width of the band in which the points of FP-OFBM lie. In fact, the standard deviation of the LA-OFBM is just 15 μJ, which is 46% lesser than the standard deviation in the FP-OFBM case. Fig. 9 plots the standard deviations of the ALU energy consumption computed over 1000 sample dies, normalized to FP-OFBM, for all the OFBMs, for all the 10 benchmarks. This plot shows that the reduction in standard deviation in the energy consumptions in the OFBMs is consistent over the benchmarks.

Fig. 8 plots another view of the same data, It plots the energy histogram for each of the OFBMs for susan corners benchmark for 1000 die samples. The second curve from the right (lines connected by circles) corresponds to the energy distribution for FP-OFBM. The rightmost curve (lines connected by diamonds) is the energy histogram of LB-OFBM, which increases the mean energy consumption by 13% but reduces the standard deviation in total energy by 17%. Finally LA-OFBM, as compared to FP-OFBM, reduces the mean and standard deviation in energy consumption by 18% and 46% respectively, as shown by its energy histogram depicted by the leftmost curve (lines joined by triangles). Thus our LA-OFBM reduces the total energy as well as the variation in total energy in the presence of both temperature and process variations.

7. Summary and Future Work

Continuous technology scaling for the last four decades, has lead us to the point, where the leakage energy has become a significant portion of the total energy budget of the processor. Leakage energy is highly sensitive to process and temperature variations. However existing Operation to Functional Unit Binding Mechanisms (OFBMs) do not take the process and temperature variations into consideration. In this paper, we propose to introduce leakage sensor in the ALU and propose a Leakage-Aware OFBM (LA-OFBM), which is sensitive on both the process and temperature variations, and is therefore able to effectively reduce both the total ALU power consumption, as well as the variation in the total ALU power consumption. Our experiments on a 0.45nm, 4-wide superscalar processor demonstrates that LA-OFBM reduces the total ALU energy consumption of the ALUs by 18%, and the variation in the total ALU energy consumption by 46%.

Acknowledgement

We would like to thank Microsoft Corporation for their generous support.

References

[1] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De. Parameter variations and impact on circuits and microarchitecture. *dac*, 00:338, 2003.

[2] D.Burger and T.Austin. The simplescalar tool set version 3.0, 1997.

[3] J. Deeney. Reducing power in high-performance microprocessors. In *International Symposium on Microelectronics*, 2002.

[4] S. Dropsho, V. Kursun, D. Albonesi, S. Dwarkadas, and E. Friedman. Managing static leakage energy in microprocessor functional units, 2002.

[5] S. B. et al. Modeling of intra-die process variations for accurate analysis and optimization of nano-scale circuits. In *Proc. of IEEE/ACM Design Automation Conference*, 2006.

[6] K. Flautner, N. S. Kim, S. Martin, D. Blaauw, and T. Mudge. Drowsy caches: simple techniques for reducing leakage power. In *Proc. of ISCA*, pages 148–157, Washington, DC, USA, 2002. IEEE Computer Society.

[7] P. Friedberg, Y. Cao, J. Cain, R. Wang, J. Rabaey, and C. Spanos. Modeling within-die spatial correlation effects for process-design co-optimization. In *Proc. of ISQED*, 2005.

[8] M. R. Guthaus, J. S. Ringenberg, D. Ernst, T. M. Austin, T. Mudge, and R. B. Brown. MiBench: A free, commercially representative embedded benchmark suite. In *IEEE Workshop in workload characterization*, 2001.

[9] H. Hanson. Static energy reduction techniques for microprocessor caches, in proc. iccd 2001, 2001.

[10] Z. Hu, A. Buyuktosunoglu, V. Srinivasan, V. Zyuban, H. Jacobson, and P. Bose. Microarchitectural techniques for power gating of execution units. In *Proc. of ISLPED*, pages 32–37, New York, NY, USA, 2004. ACM Press.

[11] C. H. Kim, K. Roy, S. Hsu, R. Krishnamurthy, and S. Borkar. A Process Variation Compensating Technique with an On-Die Leakage Current Sensor for nanometer Scale Dynamic Circuits. *IEEE Transactions on VLSI*, 14(6):646–649, 2006.

[12] P. Li, Y. Deng, and L. T. Pileggi. Temperature-dependent optimization of cache leakage power dissipation. In *Proc. of ICCD*, pages 7–12, Washington, DC, USA, 2005. IEEE Computer Society.

[13] M. Mutyam, F. Li, V. Narayanan, M. Kandemir, and M. J. Irwin. Compiler-directed thermal management for vliw functional units. In *Proc. of LCTES*, pages 163–172, New York, NY, USA, 2006. ACM Press.

[14] S. Rele, S. Pande, S. Onder, and R. Gupta. Optimizing static power dissipation by functional units in superscalar processors. In *Computational Complexity*, pages 261–275, 2002.

[15] M. Santarini. Thermal integrity: A must for low-power integrated circuit digital design, 2005.

[16] SPEC2000 Benchmarks, www.spec.org/benchmarks/html, 2000.

[17] Synopsys Design Compiler, www.synopsys.com/products/logic/design-compiler.html.

[18] S. Talli, R. Srinivasan, and J. Cook. Compiler-directed funcitonal unit shutdown for microarchitecture power optimization. In *Proc. of IPCCC*, NewOrleans, LA, USA, 2007.

[19] J. W. Tschanz, S. G. Narendra, Y. Ye, B. A. Bloechel, S. Borkar, and V. De. Dynamic sleep transistor and body bias for active leakage power control of microprocessors. *IEEE Journal of Solid State Circuits*, 38, Nov 2003.