Process Variation Aware Performance Modeling and Dynamic Power Management for Multi-Core Systems

Siddharth Garg Diana Marculescu
Dept. of Electrical and Computer Engineering
Carnegie Mellon University
{sgarg1,dianam}@ece.cmu.edu

Sebastian X. Herbert
DC Energy
Vienna, VA
sherbert@ece.cmu.edu

***Abstract*—Emerging multi-core platforms are increasingly impacted by the manufacturing process variations that introduce core-to-core and chip-to-chip differences in their power and performance characteristics. This can result in unacceptable yield loss since a large fraction of manufactured parts may not meet the design specifications. In this work, we present some promising, recently proposed solutions to mitigate the impact of process variations on multi-core platforms that deal with variability aware performance modeling, and static and dynamic power reduction. These solutions demonstrate the significant benefits that can be reaped if variability information is considered at the micro-architecture and system level design abstractions.**

I. Introduction

As the feature sizes of the transistors that are fabricated shrink below the wavelength of light used to print them, it has become increasingly difficult to *precisely* control the semiconductor manufacturing process. This leads to observed variations in the manufacturing process parameters that may occur at various granularities, for example, from wafer-to-wafer, from die-to-die or even within a die. The sources of manufacturing process variations include, among others, random dopant fluctuations (RDF), oxide thickness variations, line-edge roughness, lens focus and aberration [5]. From the perspective of a single fabricated die, process variations are typically characterized as either:

- Die-to-die (D2D) Variations: D2D variations affect each transistor on a given die in exactly the same way, but manifest as variations from one die to another. For example, in a *slow* process corner, the oxide thickness of each transistor on a given die may be larger than the nominal value by exactly the same amount.
- Within-die (WID) Variations: WID variations affect each transistor on a given die differently. The differences from transistor-to-transistor could be completely independent or may exhibit spatial correlations. In the latter case, it has been observed that the correlations between the WID variations from one transistor to another decrease as a function of the distance between the transistors [6].

Due to manufacturing process variations, the power and performance characteristics of each fabricated chip are different not only from the design intent, but also from one chip to another. While designing for the worst-case process variation impact is an option, it can lead to overly conservative design if the magnitude of process variations in large, which is the case at nano-scale technology nodes. In order to truly combat the effect the impact of process variations on semiconductor ICs, it is therefore imperative to adopt a *statistical design methodology*, as opposed to traditional static or worst-case design. Furthermore, to reap maximum benefit, statistical design techniques must be used at *all levels* of design abstraction, including at the micro-architecture and system level.

Process variations lead to differences between the power (i.e., leakage power dissipation) and performance (i.e., frequency) characteristics of otherwise identical cores on a multi-core system (due to WID variations) and, from one multi-core chip to another (due to D2D variations). However, core-to-core variations are exacerbated with technology scaling because of the increasing contribution of random or spatially correlated WID variations. Since WID process variations are typically unpredictable at design time, the most effective technique to deal with these variations involve post-fabrication tuning of parameters. For example, speed binning is a popularly used technique in which a population of fabricated dies is divided into several bins based on their maximum operating frequencies. Compared to the use of a single design frequency, speed binning allows improved yield — by placing a bin frequency near the bottom of the designs frequency distribution — and higher performance — by placing several bins near the top of the designs frequency distribution. Nonetheless, die- or chip-level speed-binning does not address the core-to-core variations in frequency characteristics, for which finer-grained adaptation techniques are required.

Multiple voltage frequency island (VFI) designs have emerged as an appealing solution to address the need for fine-grained tuning multi-core power and performance characteristics to combat the impact of WID process variations. In the most general sense, each island, which could be a single or a collection of cores, in a multi-core multiple VFI system can run at an independent supply voltage, body-bias voltage and clock frequency. In addition, the VFI parameters can be

978-1-4244-8192-7/10/$26.00 ©2010 IEEE

set post-fabrication, or even varied dynamically at run-time in response to process (and other sources of) variation. As these capabilities increasingly begin to appear in both application specific and general purpose multi-core systems, it is important to provide tools and methodologies for performance modeling, static power reduction and dynamic power management, all in the presence of process variations.

II. Process Variation Aware Performance Modeling

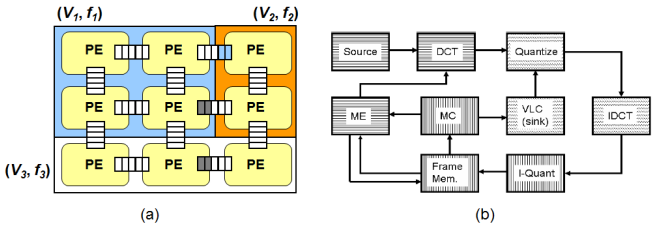

Fig. 1. (a) Multi-core platform with three VFIs and mixed-clock mixed-voltage FIFOs between VFIs. (b) Task graph of video encoder benchmark.

A promising solution to mitigate the impact of process variations, particularly WID variations, on the performance of systems with multiple processing elements is to perform speed binning at a much finer granularity than die-level speed binning, i.e., by dividing the chip into VFIs and allowing each island to run at its own maximum clock frequency. Therefore, in the proposed scheme, even if some processing elements (PE) in the die contain frequency-limiting critical paths due to WID process variations, only those PEs would be required to run at lower frequencies while the others could continue to operate at their optimum clock speeds. Since each clock domain in a VFI design can run at a different clock speed, communication between clock domains must be orchestrated using asynchronous communication interfaces. We note that each asynchronous interface comes with an associated performance and power penalty — it is therefore important to ensure that the system is partitioned into VFIs such that the communication bandwidth requirements *across* clock domains are minimized. Figure 1(a) shows an example of a multi-core that is partitioned into three VFIs.

Figure 1(b) shows the *task graph* of a video encoder, where each box represents a computational task and edges represent data communicated between tasks. Assume that this task graph is mapped and executed on the VFI platform shown in Figure 1(a). In the absence of process variations, each VFI (and core) runs at a fixed frequency and therefore the frame encoding rate, or *throughput*, of the video encoder is represented by a fixed number. On the other hand, if we allow the frequency of each VFI to be determined post-fabrication based on the impact of manufacturing process variations, the frequency of each VFI is represented as a *probability distribution*. As a result, the throughput of the video encoder will also be a distribution which can be used to determine, for a given throughput constraint, what percentage of manufactured dies will actually meet the constraint.

A. Throughput Analysis Under Uncertainty

The throughput of a task graph is determined by the cyclic dependencies (or loops) in the graph — for example, in the video encoder benchmark the ME, Frame Mem. and MC tasks form a loop. More specifically, if we denote C_i as the number of execution cycles of the i^{th} task in the graph and T_i as the cycle time of the PE on which the task is mapped, we can write the throughput, λ^*, as:

$$\lambda^* = \max_{L \in G} \frac{|L|}{\sum_{i \in L} C_i T_i} \quad (1)$$

where $|L|$ represents the number of edges in loop L. Recall that, due to process variations, the cycle times T_i are random variables.

The throughput of a task graph with deterministic cycle times can be computed in polynomial time using Karp's Maximum Cycle Mean (MCM) algorithm [4]. However, in the presence of uncertainty, computing the performance distribution is somewhat more complicated. One option is to perform Monte Carlo simulations — running Karp's MCM algorithm multiple times with different inputs in each run. However, this can be time consuming. In [2], the authors propose a statistical version of Karp's MCM algorithm that can provide the a close approximation of the throughput distribution with significant reduction in run-time over Monte Carlo simulations.

B. Granularity of VFI Partitioning

Finer-grained partitioning of a system into VFIs is expected to provide increased performance in the face of manufacturing process variations, although it could result in increased implementation costs. In addition, the performance overhead of crossing clock domains cannot be ignored. Figure 2 shows the cumulative distribution function *cdf* of throughput for the video encoder benchmark running on a nine core platform with an increasing granularity of VFI partitioning.

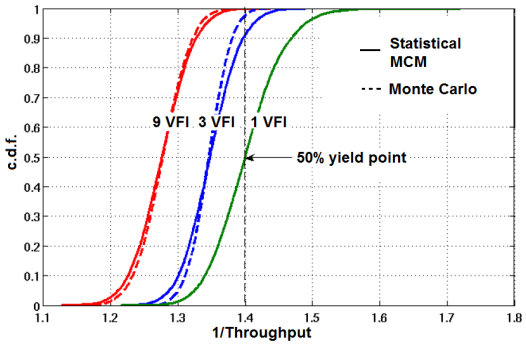

Fig. 2. *cdf* of throughput for the video encoder benchmark with varying granularity of VFI partitioning. The dashed curves represent results from Monte Carlo simulations while the solid curves are obtained from statistical MCM.

As it can be seen, increasing the granularity of VFI partitioning improves the system yield from only 50% to 90% and 98% for the three and nine VFI designs, respectively.

III. Process Variation Aware Static Power Reduction

Besides their impact on the cycle time or operating frequency, process variations have a significant impact on the leakage power dissipation of the processors in a multi-core system. The exponential dependence of the subthreshold leakage current on transistor threshold voltage implies that even small variations in the process parameters that effect threshold voltage will lead to large variations in subthreshold leakage. As a result, designers can no longer set their power budgets based on nominal values of leakage power dissipation, since the yield at the nominal value would be unacceptably low.

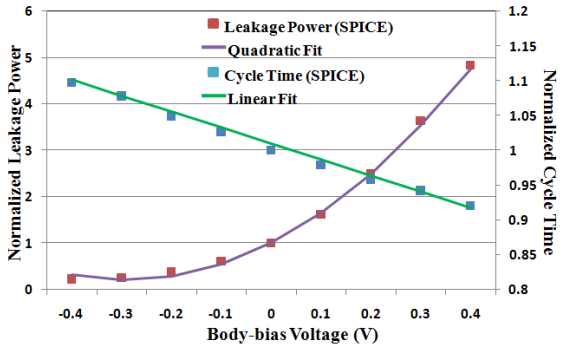

Fig. 3. Impact of body-biasing on inverter delay and leakage power dissipation.

Adaptive Body Biasing (ABB) is one of the most effective *post-silicon* techniques to minimize the variability in leakage power dissipation. The technique is based on the observation that changing the voltage of the body terminal of a transistor modulates its threshold voltage, thereby changing the leakage power dissipation of the transistor. Specifically, a negative (positive) body-to-source voltage applied to an nMOS (pMOS) device, also called Reverse Body Biasing (RBB), reduces its leakage power and increases its delay, while applying a positive (negative) body-to-source voltage, or Forward Body Biasing (FBB), increases its leakage power and reduces its delay. Figure 3 shows the impact of body-biasing on the leakage power dissipation and delay of a 90 nm benchmark circuit — as it can be seen, the impact is well modeled using a quadratic and linear fit, respectively.

Again, due to the increasing impact of WID variations, significant benefits can be obtained by partitioning a multi-core system into VFIs and allowing the body-bias voltage of each VFI, henceforth referred to as a body-bias island (BBI) to be set independently post-fabrication. This allows leakier BBIs on the die to receive a RBB, thereby reducing the leakage, but reducing their performance. On the other hand, less leaky cores would receive a FBB in order to recover the performance lost by the BBIs that received a RBB — since the relationship between leakage and body-bias is strongly non-linear, while that between frequency and body-bias is linear, we can make use of the fine-grained, post-fabrication tuning to reduce leakage without sacrificing performance.

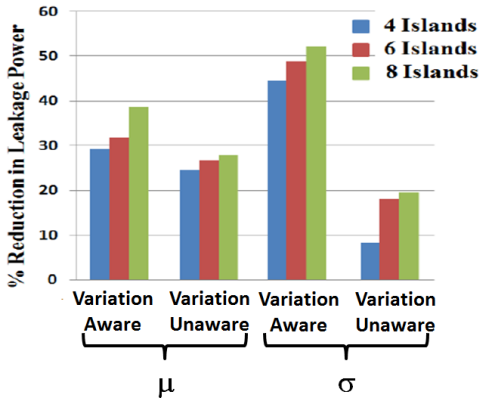

Fig. 4. Reduction in mean and standard deviation of leakage power dissipation using varying granularity of BBI partitioning.

Ideally, allocating each core to a separate body-bias island would maximize the saving in leakage power dissipation but might be impractical from an implementation perspective. For a fixed number of BBIs, [1] shows that the mapping of cores to BBIs has a significant impact on the leakage power savings obtained from post-silicon body-bias tuning. The authors also propose a robust convex optimization based algorithm to determine the best core to BBI mapping at design time. Figure 4 shows the reduction in the mean (μ) and standard deviation (σ) of leakage power dissipation with varying granularity of BBI partitioning, for both variation aware and variation unaware partitioning techniques. As it can be seen, taking into account the variability information results in significant savings in both the mean and standard deviation of leakage power dissipation. In addition, increasing granularity of BBI partitioning, as expected, also provides benefits from a leakage power perspective.

IV. Process Variation Aware Dynamic Power Management

In order to maintain single-thread performance, current general-purpose chip multi-processor (CMP) designs use relatively few high-performance cores as opposed to a large number of less capable cores. However, the highest available level of performance is often not called for by a thread (for example, if a given application phase is highly memory-bound). Dynamic voltage/frequency scaling (DVFS) is a popular method for exploiting this observation and is implemented in virtually all commercial microprocessors. By lowering clock speed and supply voltage during frequency-insensitive application phases, large reductions in power can be achieved with modest performance loss.

In the absence of process variations, the DVFS control algorithm used for each one of the many homogeneous cores in a chip-multiprocessor can be exactly the same. However, in the presence of process variations, the cores vary substantially in their leakage power consumption due to the exponential dependence of subthreshold leakage current on both channel length and threshold voltage. The energy per switching event

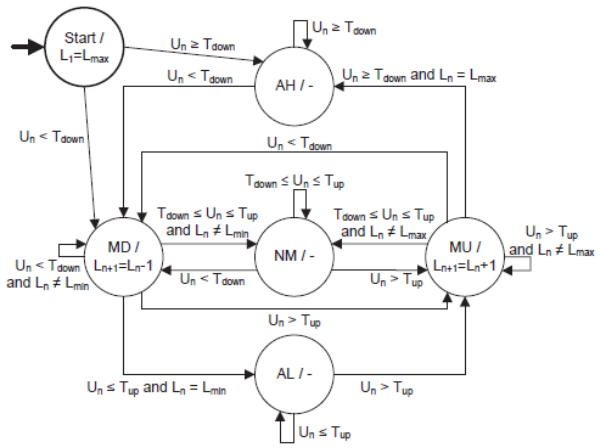

Fig. 5. Variability-unaware Threshold DVFS algorithm.

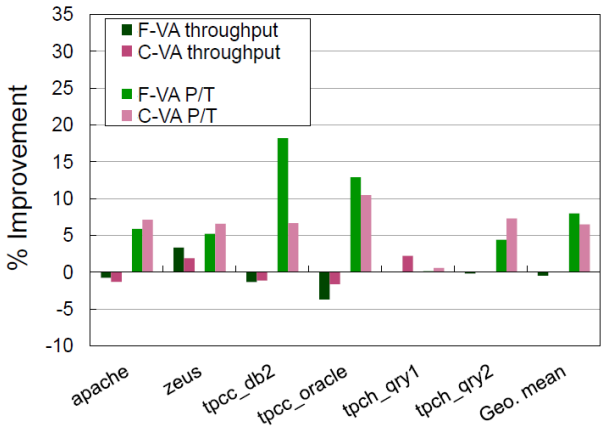

Fig. 6. Throughput and power/throughput (P/T) of variation aware DVFS compared to variation-unaware DVFS for fine-grained (F) and coarse-grained (C) partitioning.

is not significantly affected by variations, so leakier cores will be less energy-efficient than their less leaky counterparts, delivering identical performance for greater power. A recently proposed scheme, Variation Aware (VA) DVFS [3], proposes to explicitly exposes variability information (rather than just power consumption) to the DVFS controller so that it can be taken into account when setting voltage/frequency levels.

VA DVFS works by biasing leaky, energy-inefficient voltage/frequency islands (VFIs) towards lower voltage/frequency (VF) levels and less leaky, energy-efficient VFIs to higher VF levels. Because of the exponential dependence of leakage power on supply voltage and the orders of magnitude difference in leakage between leaky and less leaky VFIs, the large amounts of leakage power saved on leaky VFIs will outweigh the extra leakage power consumed on less leaky VFIs at iso-performance. For example, the variability-unaware threshold algorithm shown in Figure 5 is modified to take into account the impact of process variations by setting *different* thresholds for each core, based upon its classification as either leaky, energy-inefficient core or a less-leaky, energy-efficient core.

VA DVFS is well-suited to environments where threads are not tightly coupled (as some threads will be slowed down while others are sped up) and that contain a large number of machines (so that the overall throughput is unaffected). For example, a large-scale transaction processing environment would benefit significantly from VA DVFS, while a scientific application where threads synchronize at a barrier between phases would not.

Figure 6 shows the improvement in power/throughput (P/T) obtained by using a variation-aware (VA) DVFS algorithm compared to a baseline variation-unaware technique for a typical 16 core chip multi-processor running a number of parallel processing benchmarks. For the 16 core CMP, we considered two scenarios — fine-grained (F) partitioning in which each core is a separate VFI and coarse-grained (C) partitioning in which each VFI consists of four cores. It can be observed that up to 17% (11%) improvement in power/throughput is obtained for the fine-grained (coarse-grained) design by making use of a variation-aware DVFS policy as opposed to a conventional variation-unaware policy.

V. Conclusion

Future generation, massively parallel multi-core systems will need to cope with large scale variations in the power and performance characteristics due to manufacturing process variations, even if the cores are otherwise identical. While circuit-level, design time variability mitigation techniques can certainly help, variability aware post-fabrication tuning techniques at the micro-architecture and system level design abstractions are expected to have the greatest impact. Design tools, methodologies and algorithms are required to aid the adoption and adoption of these advanced variability mitigation techniques — we have briefly outlined in this paper some emerging solutions along this direction. In particular, we have discussed recently proposed variability aware performance modeling, static power reduction and dynamic power management techniques for both application specific and general purpose multi-core platforms.

References

[1] S. Garg and D. Marculescu. System-level mitigation of WID leakage power variability using body-bias islands. In *Proceedings of the 6th IEEE/ACM/IFIP international conference on Hardware/Software codesign and system synthesis*, pages 273–278. ACM, 2008.

[2] S. Garg and D. Marculescu. System-level throughput analysis for process variation aware multiple voltage-frequency island designs. *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, 13(4):1–25, 2008.

[3] S. Herbert and D. Marculescu. Variation-aware dynamic voltage/frequency scaling. In *IEEE 15th International Symposium on High Performance Computer Architecture, 2009. HPCA 2009*, pages 301–312, 2009.

[4] R.M. Karp. A characterization of the minimum cycle mean in a digraph. *Discrete Math*, 23, 1978.

[5] X. Li, J. Le, and L.T. Pileggi. *Statistical Performance Modeling and Optimization*. NOW Publishers, 2007.

[6] J. Xiong, V. Zolotov, and L. He. Robust extraction of spatial correlation. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 26(4):619–631, 2007.