Integrating Dynamic Voltage/Frequency Scaling and Adaptive Body Biasing using Test-time Voltage Selection

Alyssa Bonnoit
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
abonnoit@ece.cmu.edu

Sebastian Herbert
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
sherbert@ece.cmu.edu

Diana Marculescu
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
dianam@ece.cmu.edu

Lawrence Pileggi
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
pileggi@ece.cmu.edu

ABSTRACT

Adaptive body biasing is a promising technique for addressing increasing process variability, but it also provides new opportunities for reducing power when combined with dynamic voltage/frequency scaling. Limitations of existing ABB/DVFS proposals are explored, and a new scheme, test-time voltage selection (TTVS), is presented. By delaying the mapping between frequency and supply voltage until test, variability information can be incorporated into the V_{DD} selection process. For a 16-core chip-multiprocessor implemented in a high-performance predictive 22 nm technology, TTVS results in 18% power savings over independent ABB/DVFS and 11% power savings over the best of several previously proposed ABB/DVFS schemes.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: Design studies

General Terms

Design, Performance

1. INTRODUCTION

The tremendous success of the semiconductor industry has been driven by the scalability of the MOSFET. For the past 30 years, transistor density has been doubling roughly every two years, enabling increases in microprocessor performance and functionality. As device dimensions are scaled, precise control of key physical and electrical parameters becomes increasingly difficult. Traditionally, the resulting variability has been addressed through corner analysis and by speed-binning chips. However, as variability increases, designing to meet specifications in all corners sacrifices increasing amounts of performance or efficiency in the common case.

An example of efficiency lost to worst-case design can be found in the design of a dynamic voltage/frequency scaling (DVFS) system. By lowering clock speed and supply voltage during frequency-insensitive application phases, DVFS achieves large reductions in power with modest performance loss. DVFS requires that the processor design have multiple voltage/frequency operating points defined for each speed bin. Typically, a design-wide set of discrete voltage levels is chosen, and frequencies corresponding to each V_{DD} for each speed bin are set such that reliable operation is assured in the face of worst-case process variation (within a speed bin), thermal variation, and supply voltage variation.

This approach leaves significant headroom in the common case, as most dies could meet the frequency target using lower V_{DD} due to their less-than-worst-case process variations. One method to reclaim some of this excess margin at runtime is adaptive body biasing (ABB). Forward body biasing (FBB) decreases the threshold voltage (V_{TH}) of transistors, increasing both maximum frequency and leakage, while reverse body biasing (RBB) has the opposite effect. A variety of designs have been proposed to set the body biases statically [14] or dynamically [6, 10].

For a given frequency, there are several feasible operating points corresponding to different supply voltages and body biases, and running at the design-time V_{DD} with RBB to reclaim margins may not yield minimal power. For example, a die with very low leakage might benefit from being run at a lower V_{DD} (saving large amounts of dynamic power) with FBB to make up the frequency loss (at a low cost in leakage).

This paper proposes test-time voltage selection (TTVS) to address this shortcoming. The available frequency levels are chosen as usual, but the choice of which V_{DD} level will correspond to each is delayed until test-time and is based on a simple leakage measurement. The function mapping measured leakage values to V_{DD}s is obtained through characterization, with the goal of running each processor core at close to its optimal V_{DD}. This approach uses only the voltage levels which were available in the baseline DVFS

This research has been funded in part by National Science Foundation Award No. CNS 00720529. Sebastian Herbert is supported by an Intel Foundation PhD Fellowship.

ISLPED'09, August 19–21, 2009, San Francisco, California, USA.

scheme and results in 18% power savings over independent ABB/DVFS for a sample 16-core chip multiprocessor design implemented in a predictive 22 nm technology. TTVS is shown to compare favorably with several other proposed ABB/DVFS schemes, reducing power by 11% compared to the best prior technique. Moreover, the power savings are shown to be robust to differences between the activity factor and temperature assumed in the creation of the mapping and those encountered at runtime.

The remainder of this paper is organized as follows. Section 2 discusses previous studies on V_{DD} and body bias optimization. Section 3 describes the models used in this work, and Section 4 proposes test-time voltage selection. Section 5 details the experimental methodology used to obtain the results presented in Section 6. Section 7 concludes.

2. RELATED WORK

Previous work demonstrated that body biasing and DVFS implemented independently can achieve lower power than DVFS alone [5]. In this implementation, the processor specifies frequency and V_{DD} while a body bias controller adjusts the body biases to meet the frequency target. Martin *et al.* suggested using SPICE models to analytically solve for the V_{DD} and body bias combination with lowest power [4], but measuring the full set of physical and electrical parameters required by SPICE for each fabricated core would be prohibitive in terms of test effort and sensitive to random variations in the test structures used to determine these parameters. TTVS overcomes these limitations by making decisions based on a single measurement of a core's total leakage.

Later work considered integrated approaches. The most straightforward is to test all V_{DD}s at each frequency level, with the body bias controller adjusting the body biases to meet the frequency, and then select the V_{DD} with the lowest power for each frequency level [13]. However, in high-volume manufacturing the test time required is prohibitive.

Several papers have proposed using the lowest possible V_{DD} with the maximum FBB. Narendra *et al.* fabricated a test chip in a 150 nm technology and determined that lower power for a given frequency was achieved by using 450 mV of FBB to enable low-V_{DD} operation [7]. Tachibana *et al.* proposed running non-leaky dies at reduced V_{DD}, with an ABB controller used to meet the frequency target, and showed considerable power savings on such dies [12].

The literature on V_{DD}/V_{TH} optimization (where V_{TH} is adjusted with implants rather than body bias) demonstrated that the minimum total power occurs at a fixed ratio of the switching current through V_{DD} to the leakage current through V_{DD}, regardless of the operating frequency and temperature [9]. This inspired Nomura *et al.* to design a controller which sets both V_{DD} and the body biases to meet the frequency requirement while achieving a fixed ratio of these currents [8]. It is unclear that power will be minimized at a fixed current ratio when V_{TH} is modulated with body biases instead of implants for a variety of reasons (*e.g.*, the two have different impacts on short-channel effects).

Most importantly, many of the proposed approaches do not consider variability, showing results only for a "typical" die. While the minimum V_{DD} approach might yield power savings on a specific die due to the reduced dynamic power, it could increase power on a die where a large percentage of the total power is leakage (due to the increase in subthreshold and junction leakage from FBB outweighing the dynamic power savings). Similarly, the ratio of switching current through V_{DD} to leakage current through V_{DD} which gives the lowest power for a given frequency will change as the contribution of junction leakage changes. By delaying the mapping between frequency and V_{DD} until test, variability information for each die can be taken into account, significantly reducing power at iso-performance.

3. MODELING

The processor architecture used in this work is a chip-multiprocessor composed of 16 core tiles and 16 L2 cache tiles, and divided into multiple voltage/frequency islands (VFIs). The L2 cache, network, and memory controller always run at the nominal V_{DD} and the highest frequency level given a chip's speed bin, while each core has its own clock and V_{DD}. Transistor-level models are used to capture the impact of process and environmental variability in a predictive 22 nm technology.

3.1 Parameter Variation Modeling

Variations are considered in three parameters: effective channel length L_{eff}, PFET threshold voltage V_{THp}, and NFET threshold voltage V_{THn}. Die-to-die and spatially-correlated within-die variations are modeled. Due to averaging over the transistors in a path (for delay) or core (for power), uncorrelated within-die variations do not have a significant impact at the core level and above [1], and therefore are ignored. The inverters used to measure the PFET-to-NFET strength ratio in the body bias controller (as described in Section 5.1) are built from wide transistors to make them insensitive to random within-die variations.

Separate normally-distributed die-to-die shifts are modeled for the three parameters, as the primary source of die-to-die variation in each is different (lithography effects for L_{eff}, PFET/NFET channel doping for V_{THp}/V_{THn}). However, spatially-correlated L_{eff} variation is the main source of spatially-correlated V_{TH} variation [3], so the spatially-correlated components are assumed to be perfectly correlated. L_{eff} values are generated on a 80 × 72 grid overlaid on the processor die, and the L_{eff} values at these points are assumed to follow a multivariate normal distribution. The correlation between parameter values at two points is given by a Gaussian function of the distance r between the points:

$$\rho(r) = e^{-\left(\frac{r}{R_L}\right)^2} \quad (1)$$

A correlation distance R_L of 1 mm is used [1].

A $\frac{3\sigma}{\mu}$ of 15% is assumed for each of the die-to-die V_{TH} variation components. For L_{eff} variation, a total $\frac{3\sigma}{\mu}$ of 20% is used, with equal distribution of variance between the die-to-die and within-die components. The variance of the within-die component is assumed to be equally distributed between its spatially-correlated and -uncorrelated components [1].

3.2 Frequency and Power Modeling

The L_{eff}, V_{THp}, and V_{THn} values generated by this model are used to determine how core-level metrics scale across V_{DD}, temperature T, PFET body bias V_{BSp}, and NFET body bias V_{BSn}. Response surface models for maximum frequency, dynamic power, static power, and the output voltage of an input/output connected inverter were obtained by fitting to data obtained used HSPICE with the 22 nm hi-K metal gate high-performance BSIM4 predictive technology

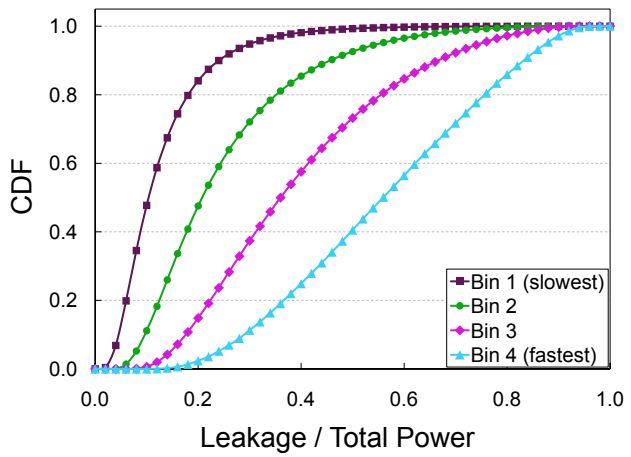

Figure 1: Ratio of leakage to total power

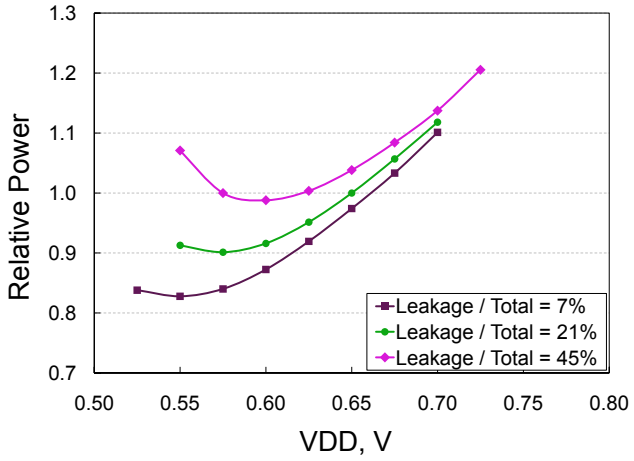

Figure 2: Power vs V_{DD} at fixed frequency

models [15]. Each model is a signomial of the form

$$f(x_1, x_2, ..., x_n) = \sum_{i=1}^{M} \left(c_i \prod_{j=1}^{n} x_j^{a_{ij}} \right) \quad (2)$$

where M is the number of terms. The arguments x_k must be greater than zero, while the model parameters (the c_i and a_{ij}) must be real numbers. All of the models used are 3rd-order, containing exactly once each term corresponding to a combination of 3 or fewer of the arguments.

Fit data were generated on a grid of uniformly-spaced $(V_{DD}, V_{BSp}, V_{BSn}, T, V_{THp}, V_{THn}, L_{eff})$ 7-tuples. V_{DD} values were spaced between the lowest and highest levels in the processor (0.5 V and 0.8 V) while body-biases between 0.5 V RBB and 0.5 V FBB were considered. Temperature was assumed to lie between 45 °C and 100 °C, which are typical values for the ambient temperature and maximum processor temperature, respectively [11]. All variation parameters were simulated over a 6σ range. Test data were generated on a second uniformly-spaced grid with no overlap between values in the fit and test datasets.

The frequency model tracks the frequency of a 13-stage ring oscillator of FO4 inverters. While the ring oscillator might not accurately track wire-dominated paths, the majority of microprocessor critical paths are gate-dominated [2]. The power models predict how dynamic and leakage power scale by fitting separate models to the current drawn through the V_{DD}, V_{BSp}, and V_{BSn} nodes and multiplying by the appropriate voltage. Finally, the inverter output model tracks the output voltage of an I/O-connected inverter.

7^7 fit data points and 8^7 test data points were used. On the test data, the resulting models obtain RMS errors of 0.31%, 0.56%, 3.71%, and 0.43% for frequency, dynamic power, leakage power, and inverter output, respectively.

4. TEST-TIME VOLTAGE SELECTION

4.1 Motivation

To examine the sensitivity of the optimal V_{DD} and body biases to process variation, 100,000 dies were generated and assigned to one of four speed bins, as described in Section 5. Figure 1 shows the distribution of the ratio of leakage power to total power ($\frac{P_{leak}}{P_{total}}$) when running each chip at the highest frequency level for its speed bin and the nominal V_{DD} at a temperature of 75 °C. The average $\frac{P_{leak}}{P_{total}}$ is approximately 30%, but there is significant variation.

Figure 2 examines the interplay between DVFS and ABB. Three dies were selected from Bin 2 (3.2 GHz) at the 10^{th}, 50^{th}, and 90^{th} percentiles of the intra-bin $\frac{P_{leak}}{P_{total}}$ distribution, corresponding to 9%, 21%, and 46% leakage. The figure shows total power versus V_{DD} at an intermediate frequency level corresponding to a V_{DD} of 0.65 V in the baseline design, with the body biases adjusted to keep the frequency constant while balancing the PFET-to-NFET strength ratio.

The least leaky die achieves minimum power with lower V_{DD}, because dynamic power decreases roughly quadratically as V_{DD} decreases and increases slowly as forward bias is applied (due to increasing junction capacitance). The leakiest die has lowest total power with higher V_{DD}. While subthreshold leakage decreases as V_{DD} decreases due to drain-induced barrier lowering (DIBL), this is dominated by the exponential increase in both junction and subthreshold leakage with FBB. Test-time voltage selection exploits these effects by assigning V_{DD}s based on leakage.

4.2 Determination of Mapping

By delaying the mapping of frequency and V_{DD} pairs until test-time, variability information can be exploited. Once per product, a set of dies are characterized to determine a mapping from leakage to V_{DD} for each frequency level of each speed bin. This characterization provides four measurements for each sample core - the speed bin of the chip the core was on, its leakage at standard operating conditions (temperature of 75 °C and the highest V_{DD}), its optimal V_{DD} for each frequency level (at the standard temperature), and its minimum V_{DD} (assuming full FBB) for each frequency level (at the worst-case temperature of 100 °C [11]).

The mapping is complicated by the fact that the correlation between the leakage metric and operating speed is not perfect. As a result, the mapping from leakage to V_{DD} cannot simply be computed based on the optimal V_{DD} ($V_{DD,opt}$) because the optimal V_{DD} for one die could be below the minimum V_{DD} ($V_{DD,min}$) of another die with the same leakage metric value. Moreover, the mapping should try to get cores as close as possible to their optimal V_{DD} at a typical oper-

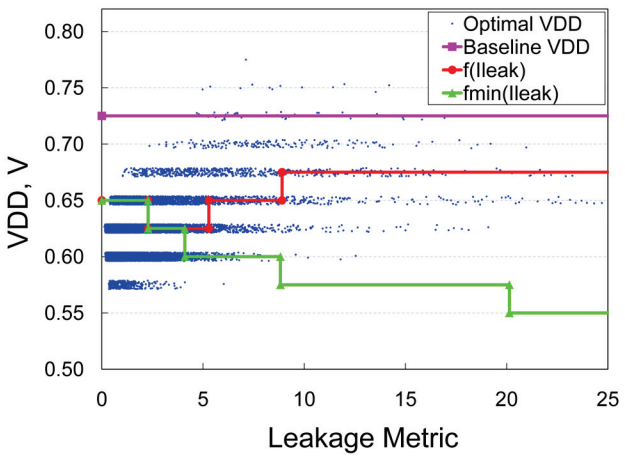

Figure 3: Example mapping from leakage to V_{DD}

ating temperature, while the minimum V_{DD} constraint is based on the maximum temperature.

The minimum V_{DD} constraint function, $f_{min}(I_{leak})$, is defined such that $f_{min}(I_{leak}) \geq V_{DD,min}$ for 99% of the cores. Within each bin, the leakier dies tend to be faster, and thus $f_{min}(I_{leak})$ is defined to be a piecewise-constant, monotonically decreasing function. This approach does not strictly guarantee that no core is assigned too low a V_{DD}, as that would allow outliers to significantly impact the mapping. Instead, a simple test-time extension handles such cases, described in the next subsection. $f_{min}(I_{leak})$ is used to lower-bound the final mapping.

There are two competing effects in determining the final mapping $f(I_{leak})$ - the minimum V_{DD} decreases with increasing leakage, whereas the optimal V_{DD} increases with increasing leakage. As a result, $f(I_{leak})$ is constrained to be a piecewise-constant, convex function. It is determined iteratively by attempting to minimize

$$\epsilon(f) = \sum_{i=1}^{n_{cores}} |V_{DD,opt}(i) - f(I_{leak}(i))| \quad (3)$$

For each iteration, the location of each jump in the function is swept between its left and right neighbors, subject to $f(I_{leak}) \geq f_{min}(I_{leak}), \forall I_{leak} \in [0, \infty)$, and placed at the point which minimizes $\epsilon(f)$. Jumps which "pile up" at the minimum or maximum I_{leak} values are eliminated.

An example of the mapping is shown in Figure 3. Each blue point represents the optimal V_{DD} of a core vs. its leakage metric (noise was added to V_{DD}s to display the density of points). The minimum mapping, $f_{min}(I_{leak})$, is shown in green, and the final mapping, $f(I_{leak})$, is shown in red.

4.3 Application of Mapping

This mapping is used in high-volume manufacturing. Each fabricated die is speed-binned as usual. The leakage of each core is measured at the nominal V_{DD} with no body biases at a typical operating temperature and used to assign a V_{DD} for each frequency level, based on the mapping $f(I_{leak})$. Finally, the die is re-tested using the assigned V_{DD}s and if a core is not able to meet the required frequency at a level, the baseline V_{DD} is assigned to that level. Because these occurrences are extremely rare, negligible benefit is sacrificed.

The test-time overhead of *TTVS* is minimal. Both the leakage measurements and the final per-core test can be performed in parallel for all cores. Furthermore, the extra test adds negligible overhead compared to the testing that must be done for speed-binning.

5. EXPERIMENTAL METHODOLOGY

5.1 Speed Binning

In order to determine speed bins, two million dies were generated via simulation and their maximum frequency computed at the nominal V_{DD} of 0.8 V and worst-case temperature of 100 °C with no body biasing. Four speed bins were created such that each bin's frequency is an integer multiple of a 133 MHz system clock. The speed bins are located at 4.133, 3.733, 3.2, and 2.533 GHz, and contain 10.0%, 16%, 37%, and 35% of dies, resulting in a parametric yield of 98%.

5.2 Schemes Evaluated

Several implementations of ABB/DVFS are considered. The baseline is a traditional DVFS scheme with no body biasing, referred to as *ZBB*. *Independent* adds ABB, but uses the same voltage/frequency mapping as *ZBB*. *Minimum* instead runs each core at the minimum V_{DD} that can be achieved with body biasing, determined at test by sweeping through the available V_{DD}s. *Ratio* is an implementation of the scheme proposed by Nomura *et al.* [8], in which the body biases are determined to meet the frequency constraint and achieve a target ratio of $\frac{I_{sw}}{I_{leak}}$. Finally, *TTVS* is the design proposed in this paper, which chooses the V_{DD} for each frequency based on the leakage power measured at test.

Independent, *Minimum*, and *TTVS* rely on a body bias controller to determine the body biases which both meet the frequency requirement and balance the PFET-to-NFET strength ratio. There is extensive literature on body bias controllers, and the controller used in this work is similar to that presented by Ono and Miyazaki [6,10]. A pair of interlocked feedback loops continually adjust the PFET bias to meet the frequency target while the NFET bias is adjusted to keep the output of an I/O-connected inverter at $\frac{V_{DD}}{2}$.

In *Ratio*, the switching current is assumed to be proportional to $V_{DD} \cdot f$, so body biases are adjusted to meet a target ratio of $\frac{V_{DD} \cdot f}{I_{leak}}$ [8]. The target ratio was determined by running Monte Carlo and choosing the value which yields the lowest average power across all frequency levels.

The mapping functions between leakage and V_{DD} in *TTVS* were determined based on 5,000 simulated dies from each speed bin. Using significantly more points (100,000 from each bin) was not found to improve the average (per-core) value of the error metric from Equation 3.

5.3 Scenarios Evaluated

Both coarse-grained and fine-grained DVFS implementations are considered. Fine-grained DVFS has 13 frequency levels, with V_{DD} levels every 25 mV between 0.5 V and 0.8 V. Coarse-grained DVFS has 5 frequency levels with V_{DD} levels every 75 mV between 0.5 V and 0.8 V. Several distributions of runtime between frequency levels are examined. *Uniform* assumes that the same amount of time is spent at each level. *Skewed-high* assumes that the time spent at a frequency level is proportional to its index i (with the lowest level having $i = 1$ and the highest level having $i = n_{levels}$, while *Skewed-low* assumes that the time is proportional to $(n_{levels} - i)$.

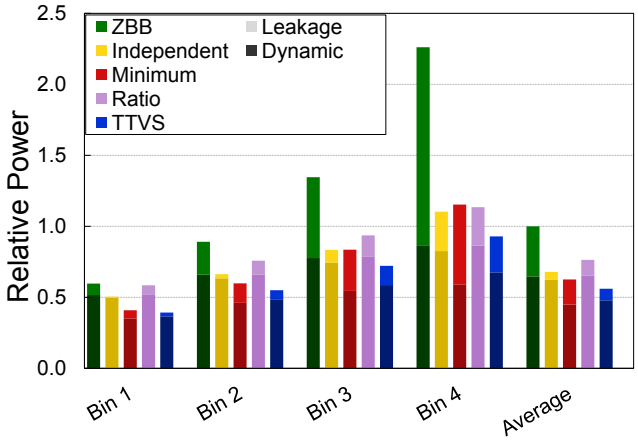

Figure 4: Average power by speed bin

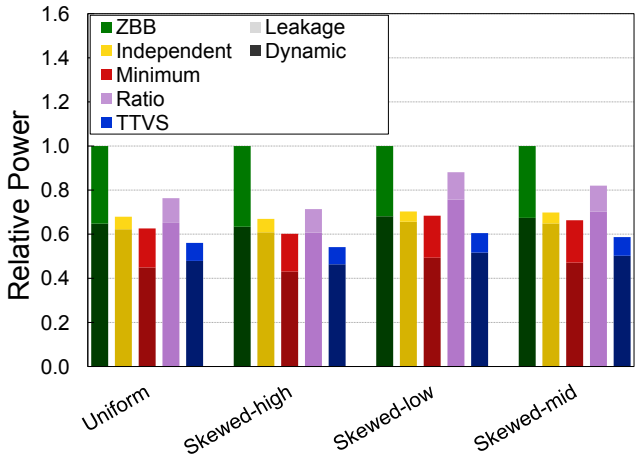

Figure 5: Average power by runtime distribution

Finally, *Skewed-mid* assumes that the time spent at a frequency level is proportional to $min\,(i, n_{levels} - i)$.

The TTVS mapping $f\,(I_{leak})$ was obtained with an average $\frac{P_{leak}}{P_{tot}}$ of 30% at a typical operating temperature of 75 °C. However, there is a significant variation in activity factor among workloads (*e.g.*, based on their compute- versus memory-boundedness). Moreover, processors run in a variety of thermal environments. Therefore, all five DVFS/ABB implementations are evaluated across a range of temperatures and activity factors, with *TTVS* always using the same mapping function. Temperatures of 50 °C, 75 °C, and 100 °C were used, while $\frac{P_{leak}}{P_{tot}}$ was shifted from its average value at 75 °C of 30% to averages of 20% and 40% by scaling the dynamic and leakage power appropriately.

The experimental results ignore the feedback loop between leakage power and temperature. However, it will be shown that this results in a conservative evaluation of TTVS, because it almost always achieves significantly lower power than other schemes, and is within 2.5% of the best scheme across all corners. This would translate to the lowest temperature and thus a slight further power advantage.

6. RESULTS

Figure 4 compares the average power of the five schemes at 75 °C, using fine-grained DVFS levels with a uniform distribution of time between frequency levels. Dynamic power, shown in darker colors, and leakage power, shown in lighter colors, are stacked to yield total power. Results are shown by speed bin as well as an average where the results from each bin are weighted by the percentage of dies in that bin. Each bar is normalized to the total average power of *ZBB*.

TTVS shows consistently lower power than *ZBB* across all speed bins, with a total average power savings of 44%. The results for *Independent* show that adding ABB to the baseline DVFS scheme reduces average power by 32%. *Minimum* is generally able to further reduce power, although the results differ greatly across speed bins (from a 20% reduction in power for the slowest speed bin to 5% increase for the fastest speed bin). This is because slower dies have lower $\frac{P_{leak}}{P_{total}}$, and therefore benefit from the lowest V_{DD} with forward biases. By selecting the V_{DD} at test-time based on the leakage measurement, *TTVS* is able to overcome this shortcoming, reducing average power by a further 11% over *Minimum*. *Ratio* has higher power than *Independent* across all speed bins. This is because the ratio of $\frac{I_{sw}}{I_{leak}}$ for the lowest total power changes with variability.

Figure 5 shows the results across different distributions of runtime between frequency levels at 75 °C using fine-grained DVFS levels. Each bar is the weighted average across all bins. For each distribution of runtime, the powers of the five implementations are normalized to the power of *ZBB*. The trends in power savings are seen to be consistent across the DVFS runtime distributions evaluated, with *TTVS* continuing to achieve the lowest average power.

The results for the low/high temperatures (50 °C and 100 °C) and activity factors (average $\frac{P_{leak}(75\ ^\circ\mathrm{C})}{P_{tot}(75\ ^\circ\mathrm{C})}$ of 20% and 40%) are shown in Figure 6. The following temperature / $\frac{P_{leak}(75\ ^\circ\mathrm{C})}{P_{tot}(75\ ^\circ\mathrm{C})}$ pairs are evaluated: typical/typical (TT), low/low (LL), low/high (LH), high/low (HL), and high/high (HH). *Minimum* achieves 1.8% lower power than *TTVS* in the low/low case. For this corner, the total power is almost all dynamic power and so the lowest V_{DD} is optimal. However, *Minimum* performs considerably worse than all other schemes in the high/high case, where leakage is significant. On the other hand, in the high/high case, *Independent* achieves 2.4% lower power than *TTVS*. At this corner, leakage power is a very large portion of total power, and the greatest reduction in leakage will be achieved with the highest V_{DD} and largest reverse bias. However, *Independent* is worse than *TTVS* in all other cases.

Aside from these two points, *TTVS* always yields the lowest average power. Moreover, it is never far from being the best, while the other schemes can make no such claim. In addition to generally having the lowest power, *TTVS* has the advantage of providing *consistent* power savings which are robust to the processor being run in conditions significantly different from those at which the scheme was calibrated.

Figure 7 shows the average power for each bin for a design with five coarse-grained frequency levels and five corresponding available V_{DD}s, relative to the total average power for *ZBB* on the fine-grained design. While the power results differ by an average of 7.5% and worst case of 17%, the power savings of the schemes are seen to be consistent, with *TTVS* saving 44% of power compared to *ZBB*, 16% compared to *Independent*, and 7% compared to *Minimum*.

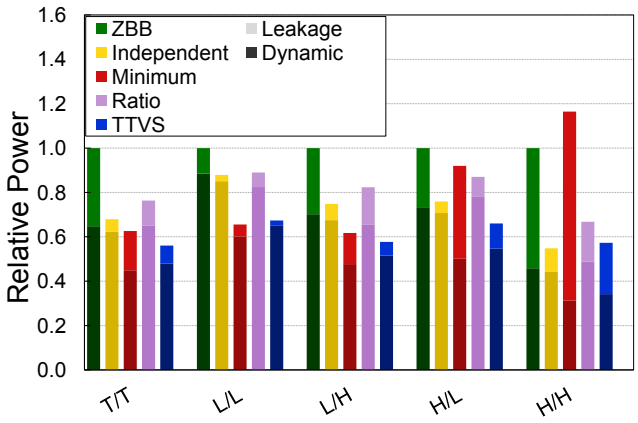

Figure 6: Average power by temperature/leakage power percentage

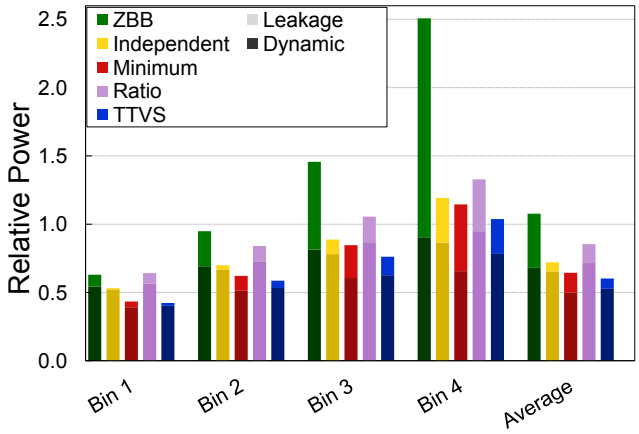

Figure 7: Results from coarse-grained DVFS relative to *ZBB* fine-grained DVFS

7. CONCLUSION

Adaptive body biasing (ABB) is a useful technique for reclaiming margins lost to variability. Significant improvements in microprocessor energy-efficiency can be achieved by integrating ABB with dynamic voltage/frequency scaling (DVFS). For a 16-core chip-multiprocessor implemented in a predictive high-performance 22 nm technology, independent implementation of ABB and DVFS was found to yield a 32% reduction in average power at iso-frequency.

Power can be reduced further by using a cooperative approach. This paper proposed test-time voltage selection, which selects the V_{DD}s for each core at test-time based on a single measurement of the core's leakage. By explicitly considering variability, it is able to yield 18% lower power than independent ABB/DVFS and 11% lower power than the best prior cooperative scheme, which runs every core at the lowest possible V_{DD} given its variations. These results were verified across fine- and coarse-grained DVFS implementations, as well as various distributions of times between DVFS levels. Finally, TTVS was shown to be robust to differences between runtime operating conditions and those at which the leakage-to-V_{DD} mapping is constructed.

8. REFERENCES

[1] Y. Abulafia and A. Kornfeld. Estimation of FMAX and ISB in microprocessors. In *IEEE Transactions on VLSI Systems*, 2006.

[2] K. A. Bowman *et al.* Impact of die-to-die and within-die parameter variations on the throughput distribution of multi-core processors. In *Proceedings of the 2007 International Symposium on Low Power Electronics and Design*, 2007.

[3] Y. Cao and L. T. Clark. Mapping statistical process variations toward circuit performance variability: an analytical modeling approach. In *Proceedings of the 42nd annual Design Automation Conference*, 2005.

[4] S. Martin *et al.* Combined dynamic voltage scaling and adaptive body biasing for lower power microprocessors under dynamic workloads. In *Proceedings of the 2002 IEEE/ACM International Conference on Computer-aided Design*, 2002.

[5] M. Miyazaki *et al.* A 175 mV multiply-accumulate unit using an adaptive supply voltage and body bias (ASB) architecture. In *IEEE International Solid-State Circuits Conference Digest of Technical Papers*, 2002.

[6] M. Miyazaki *et al.* A 1.2-GIPS/W microprocessor using speed-adaptive threshold-voltage CMOS with forward bias. In *IEEE Journal of Solid-State Circuits*, 2002.

[7] S. Narendra *et al.* Forward body bias for microprocessors in 130-nm technology generation and beyond. In *IEEE Journal of Solid-State Circuits*, 2003.

[8] M. Nomura *et al.* Delay and power monitoring schemes for minimizing power consumption by means of supply and threshold voltage control in active and standby modes. In *IEEE Journal of Solid-State Circuits*, 2006.

[9] K. Nose and T. Sakurai. Optimization of VDD and VTH for low-power and high speed applications. In *Proceedings of the 2000 conference on Asia South Pacific design automation*, 2000.

[10] G. Ono and M. Miyazaki. Threshold-voltage balance for minimum supply operation. In *IEEE Journal of Solid-State Circuits*, 2003.

[11] K. Skadron *et al.* Temperature-aware microarchitecture. In *Proceedings of the 30th annual International Symposium on Computer Architecture*, 2003.

[12] F. Tachibana *et al.* A process variation compensation scheme using cell-based forward body-biasing circuits usable for 1.2 V design. In *IEEE Custom Integrated Circuits Conference*, 2008.

[13] R. Teodorescu *et al.* Mitigating parameter variation with dynamic fine-grain body biasing. In *Proceedings of the 40th annual ACM/IEEE International Symposium on Microarchitecture*, 2007.

[14] J. Tschanz *et al.* Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage. In *IEEE Journal of Solid-State Circuits*, 2002.

[15] W. Zhao and Y. Cao. New generation of predictive technology model for sub-45nm design exploration. In *Proceedings of the 7th International Symposium on Quality Electronic Design*, 2006.