RC24221 (W0703-099) March 21, 2007
Computer Science

IBM Research Report

Benchmarking for Power and Performance

Heather Hanson*, Karthick Rajamani, Juan Rubio,
Soraya Ghiasi, Freeman Rawson
IBM Research Division
Austin Research Laboratory
11501 Burnet Road
Austin, TX 78758

*Also with University of Texas at Austin

Benchmarking for Power and Performance

Heather Hanson[†*], Karthick Rajamani[*] Juan Rubio[*], Soraya Ghiasi[*] and Freeman Rawson[*]
[*]IBM Austin Research Lab
[†] The University of Texas at Austin
{karthick,rubioj,sghiasi,frawson}@us.ibm.com, hhanson@cs.utexas.edu

***Abstract*— There has been a tremendous increase in focus on power consumption and cooling of computer systems from both the design and management perspectives. Managing power has significant implications for system performance, and has drawn the attention of the computer architecture and systems research communities. Researchers rely on benchmarks to develop models of system behavior and experimentally evaluate new ideas. But benchmarking for combined power and performance analysis has unique features distinct from traditional performance benchmarking.**

In this extended abstract, we present our experiences with adapting performance benchmarks for use in power/performance research. We focus on two areas: the problem of variability and its effect on system power management and that of collecting correlated power and performance data. For the first, we have learned that benchmarks need to expose at least three sources of power/performance variability – workload intensity, workload behavior and component-level differences. Benchmarks not only have to test across all of these forms of variability, but they also must capture the dynamic nature of real workloads and real systems. The workload and the system's response to it change across time, and how fast and well the system responds to change is an important consideration in evaluating its power management capabilities. In the second area, we have developed tools for collecting correlated power and performance data, and we briefly discuss our experience with them.

I. Introduction

There has been a tremendous increase in focus on power and cooling of computer systems both from design and management perspectives. At the same time, there is an increasing interest in understanding the performance effects of power management actions. Most power management actions such as dynamic voltage and frequency scaling have performance consequences, and understanding the trade-offs between power and performance is important at all levels of system design. Historically, performance measurement has concerned itself solely with peak sustained throughput or minimal sustained latency. The idea has been to apply a sustained load that drives the system to its peak performance. Power consumption has significant unique characteristics that need to be understood in relation to a broad spectrum of system loads. Systems are not always run at full capacity, and designers and users need to understand the power consumption and the performance that they are capable of delivering across the entire range of load intensity, from idle to peak capacity. At the same time, benchmarking for power and performance presents a more complex measurement problem than has existed for performance benchmarking alone. Power and performance data must be correlated with each other through time. Simple end-to-end numbers, t transactions consuming j Joules over s seconds do not provide enough information about dynamic runtime behavior. Thus, benchmarking for power and performance requires a broader approach to benchmark creation and execution than traditionally adopted for performance comparison purposes alone.

Our primary concern is with the development and evaluation of techniques for system- and data center-level power management. We have made extensive use of existing performance benchmarks, supplemented with power measurements, to develop models of system behavior and to evaluate our ideas and prototype implementations. Based on our experience, we believe that power and performance are affected by three types of variability. To serve the computer systems design and management communities, benchmarks must capture the **time-varying nature**, of real workloads, including the fact that the rate of time variation can itself change during the lifetime of the benchmark. To capture the real impact of the joint management of power and performance **on systems**, benchmarks need to address variability in activity among different workloads (for example, the differences between high-powered compute-bound programs and lower-power programs that stall often waiting for data from memory). Lastly, the impact **of systems** needs to be captured by exposing the variability in power consumption characteristics among system components.

Not only do benchmarks have to capture variability in workloads, across different components and at different levels of intensity, but they also must deal with the complexities of collecting and reporting two distinct types of quantities – power and performance.

II. Power Variation with Workloads

The difference in power consumption due to workload has been recognized by several benchmarking organizations. As part of the Energy Star program, the U.S. Environmental Protection Agency developed a protocol to measure power under different performance settings [1]. This document recommends that system vendors provide curves that show the power consumption under different loads. The process for producing this curve consists in running the workload applying the maximum load to the system (100%), followed by runs with reduced loads until the system under test reaches an idle state (0%). It is expected that system consumers can use this curve to estimate their overall energy consumption by multiplying their average system utilization with the appropriate point in the curve.

A. Sources of Variation

System utilization is a measure of the percentage of time that non-idle tasks run on the processor. Unlike benchmarks designed to test the peak performance of a system, real workloads exhibit a wide range in workload intensity. Figure 1 shows the variation in web server load from traces collected at two web-sites. Systems can incorporate power management solutions like dynamic voltage and frequency scaling, such as Intel's *demand-based switching*, that exploit this variation for higher power efficiencies. It is important that benchmarks designed for system power comparison capture real-world variation in workload intensity to differentiate among the power management capabilities of different systems. An approach to integrating real-world load variation with a peak-performance benchmark is presented in [2] where the TPC-W [3] benchmark is modified with the load profiles from real web-sites for developing a workload to evaluate power management solutions.

Figure 2 shows the system utilization of the NYSE profile in more detail. The NYSE website shows little activity during non-core business hours and a strong spike in usage during the day. Traditional allocation of tasks to processors in the system attempts to load balance between all available processor. Dynamic voltage and frequency scaling can take advantage of periods of low utilization by slowing the processors uniformly. An alternative approach is shown in Figure 3. In this case, the running tasks associated with the TPC-W workload are coalesced onto the minimum number of processors required to support the workload without showing a reduction in performance. During periods of low activity, only a single processor is powered and the remaining are put into a deep power-saving mode [4].

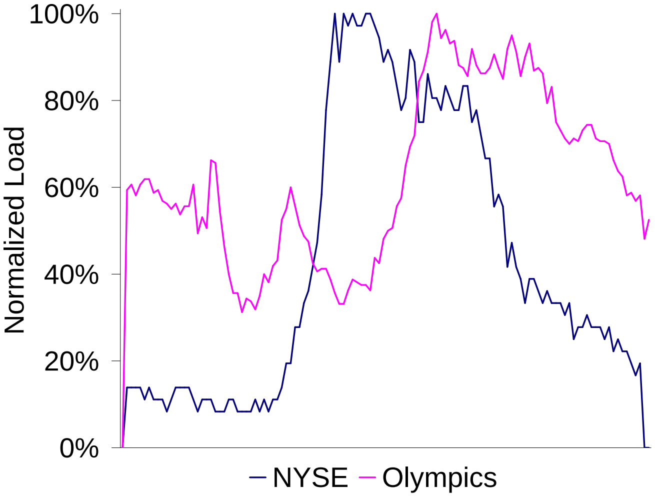

Fig. 1. Server utilization through a 24-hour period.

The EPA approach to power management does not currently capture this type of dynamic behavior. If data center operators provisioned power to account for average utilization of the NYSE profile (∼43%), periods of peak activity would present such a large draw that systems would fail. An assumption of constant utilization ignores the realities of many workloads deployed in the real world. Figure 1 shows the global nature of the Internet does not eliminate the variation seen over time as evidenced by the variation in system utilization for a trace collected on the 1998 Nagano Winter Olympics web site. Any new power and performance benchmarks should reflect the time varying nature of system utilization.

While system utilization has an enormous impact on power consumption it is not the only workload characteristic that must be considered. How a workload uses the resources of the processor also has significant impact on the power consumption of the system. Figure 6 shows the performance variation, as measured by Instructions per Cycle (IPC) , while Figure 5 shows the variation in processor chip power as different SPEC CPU2000 benchmarks are executed on the processor. During this execution the system perceived utilization is 100% - all

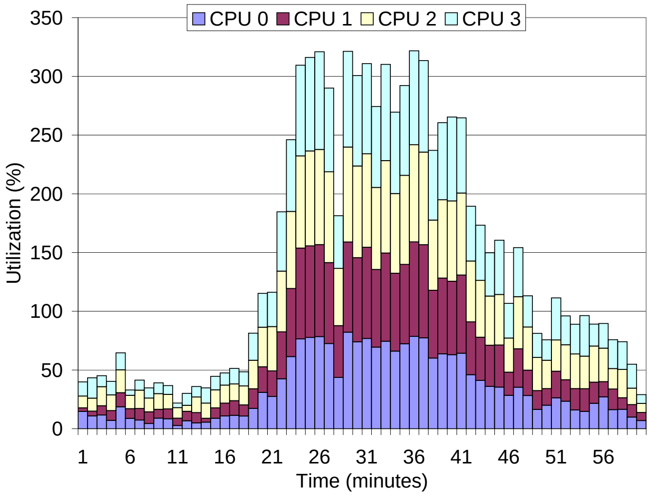

Fig. 2. Typical processor allocation would load balance the NYSE requests over all available processors, even when the extra processors are not needed to meet performance guarantees.

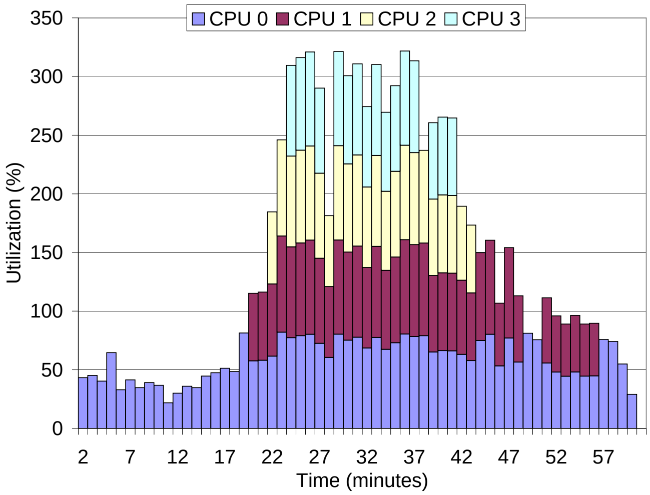

Fig. 3. The NYSE requests can be packed onto a smaller number of processors, allowing additional power savings modes to be utilized.

the variability in power consumption can be traced to the difference in microarchitecture-level activity among the workloads and can be potentially exploited in intelligent power management solutions [5]. Similar variation in power consumption has been observed even in earlier generation (Pentium 4) processors [6] and will be increasingly evident on newer processors as they employ finer activity-oriented circuit-level power-reduction techniques like clock gating and power gating more extensively.

The current EPA approach to power management measurement does not address this type of variation. A Pentium M system under identical system utilization, but with different workloads running can consume power that varies by almost a factor of two. It is important that new benchmarks that are designed for system-power-comparison studies capture this range of microarchitecture-level activity in order to represent the inherent variability in real workloads. A good power/performance benchmark needs to measure the response of the system under test to this type of variability.

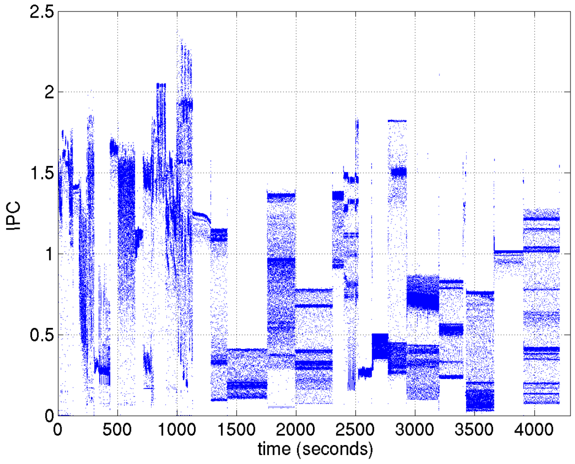

Fig. 4. IPC variation for SPEC CPU2000 on a 2GHz Intel Pentium M.

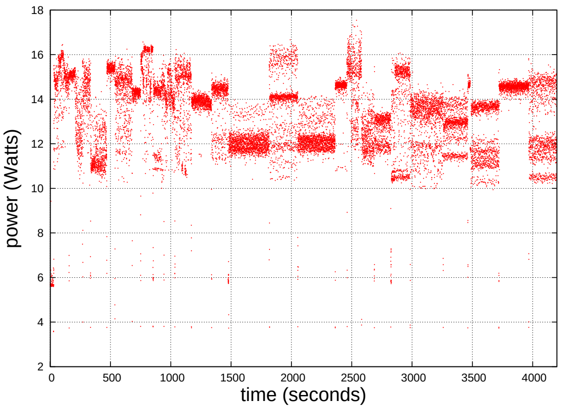

Fig. 5. Power variation for SPEC CPU2000 on a 2GHz Intel Pentium M.

A characteristic of both system utilization and microarchitectural utilization is the rate of change of activity or intensity. Workloads vary over time and the time scales at which they vary can be very different. Some things change quickly, others more slowly. Benchmarks must be sensitive to the change over time AND the rate of change over time. The time function and its derivative both matter.

The rate of change can dictate what power management mechanisms can be adopted without harming performance more than a customer is willing to tolerate. For example, at the circuit-level clock gating has a much lower overhead than power gating. At the microarchitecural-level and system-level clock or pipeline throttling and

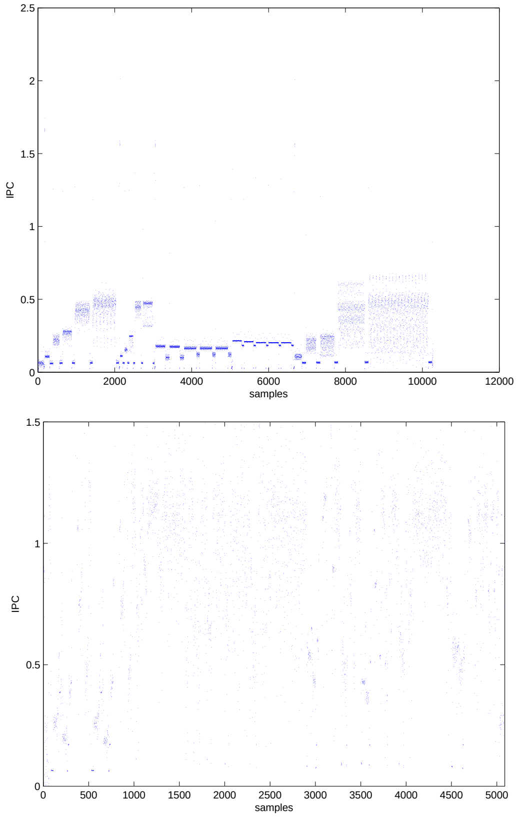

Fig. 6. Measured IPC for selected SPEC CPU2000 benchmarks: (a) `gzip`, (b) `gcc`.

frequency scaling are significantly faster to initiate and have lower overheads than supply voltage scaling, which in turn has lower overheads than powering modules or sub-systems on or off. Figure 6 shows two applications – `gzip` which exhibits a number of different stable phases and `gcc` which has much more chaotic behavior. The techniques applicable to `gzip` can have start and end overhead, while `gcc` limits the power management responses to only those which can respond rapidly.

The current EPA approach does not consider the rate of change in the workloads or system utilization. A technique such as nap mode may work very well for a current-generation transactional web benchmark, but may fail miserably when a new dynamic page is added that requires running a highly variable workload, one with execution characteristics similar to `gcc`, to generate data. The importance of this time-varying nature of workloads is discussed in [2] for a power management solution for a server cluster. Thus, a power/performance-benchmarking workload should strive to capture realistic temporal variation in workload activity to contrast the different power management solutions/capabilities of different systems.

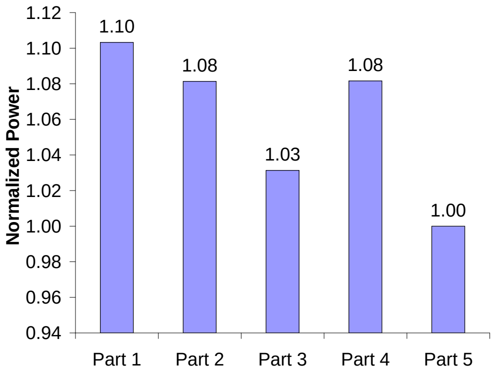

Fig. 7. Power variation across five nominally identical Intel Pentium M processor chips.

B. Variation in Computer System Components

We have considered the impact of variation in system utilization and workload characteristics and indicated why it is important for these to be captured by power/performance benchmarks. There is an additional source of variation that benchmarks run on a single system can not capture. There is a significant amount of variation in the amount of power consumed by the different components of a system. This variation is hard to capture with benchmarks as they are currently designed and suggest that each system with its constituent components should be tested, or power/performance benchmarks should be done on systems with worst case power consumption if they are to be used for provisioning.

For most lower-end systems like desktops, workstations, server blades and small, rack-mounted servers, processors can be significant power consumers among the computing components. Figure 7 shows the chip power consumption in an arbitrary selection of five Intel Pentium M processors with the same specifications. The figure shows the measured power consumption under the same constant workload at identical chip temperature. Chip designers foresee increased process and device variability with continued technology scaling and a consequent power variability [7], [8].

For higher-end or larger SMP systems, DRAM memory sub-system power can become comparable to the processor power consumption [9]. Figure 8 shows the range in the Max-Active (IDD7) and Max-Idle (IDD3N) currents in the datasheets for 512Mb DDR2-533 parts from five different vendors. IDD7 is somewhat analogous to the thermal design point (TDP) for processors. A given vendor typically sells multiple parts with same performance, but different power characteristics. Also, as the fabrication process is tuned over time, manufacturers

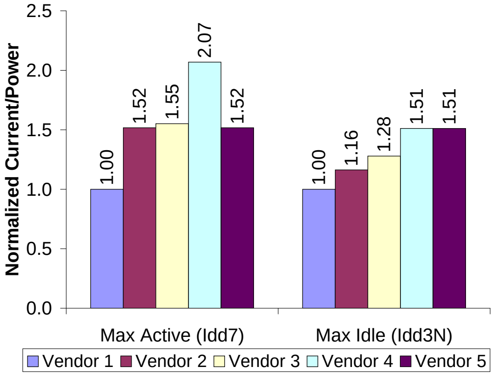

Fig. 8. Power variation across five 512 MB DDR2-533 4-4-4 DRAM parts.

are typically able to lower the power consumption for the same performance, a phenomenon that holds true for processors as well.

Depending on the system configuration, other components like disks, chipsets, graphics sub-system, etc., can also be significant power consumers with somewhat similar inherent variability in their power consumption for the same capability/performance. An illustrative example using the processors and memory from Figures 7 and 8 highlights the variation in total power that components can have. A data center administrator starts with the lowest powered parts and then replaces them as they fail over time. A system which originally consumed at most 20 Watts, may now consume 22 Watts. A single system consuming 2 additional Watts may not be problematic, but a 10% increase in power consumption across the entire data center may exceed power and cooling capacities.

Another source of variability in the power drawn by the system comes from the efficiency of the power supplies and converters used in the system. A survey of AC-DC supplies by Calwell and Mansoor [10] arguing for more usage of efficient power supplies shows that currently marketed power supplies range in efficiency from around 60% to over 85%.

As a consequence of the variability in power consumption for same performance, assessing relative power-efficiencies of two systems becomes more difficult. One must account for or neutralize the difference in component variabilities (even if using ones with identical specifications) between two systems before using their measured power to determine the difference from fundamental design differences. Also if wall-power/AC is measured for the two systems, but comparison is sought between the efficiencies of the systems after the DC conversion by the power supply, one may need to account for the potential difference in efficiencies between the power supplies. This problem is difficult not only because different power supplies have different efficiencies, but also because a single power supply has different efficiencies at different loads.

III. Power / Performance Measurement Techniques

We have used a number of different platforms and data-collection schemes in joint power/performance experiments. The most sophisticated method is described here to indicate what can be done, and how we have been able to collect correlated power/performance data traces for modeling and evaluation.

The experimental infrastructure uses a Pentium M 755 (90 nm Dothan) with Enhanced SpeedStep, which supports dynamic voltage and frequency scaling. A Radisys system board [11] with high-precision sense resistors between the processor and each of two voltage regulators is used to measure the power consumption of the processor. Figure 9 shows a diagram of the configuration. The power measurement system is non-intrusive and has a peak sampling capability of 333 K samples/s. The current, calculated from measured voltage across the sense resistor, and the supply voltage are filtered, amplified, digitized and collected periodically with a National Instruments SCXI-1125 module and National Instruments PCI-6052E DAQ card. The data are processed by a custom LabView program executing on a separate system to form an output trace of power values.

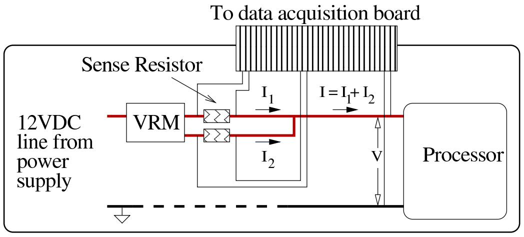

Fig. 9. Experimental platform: system under test with sense resistors and data acquisition probes to measure processor power.

We have written a program, `log`, that drives our data collection. `Log` is an application that

- monitors processor performance
- monitors environmental information
- controls operating conditions.

When started, the application configures the performance counters, clock throttling, operating frequency and voltage of the processor. To get a synchronized trace of application performance and processor power consumption, we use an I/O pin on the south-bridge of the system board as a marker for the beginning and end of the benchmark. Immediately before `log` starts a benchmark, it raises the voltage of the pin and creates an entry in its internal performance data-collection buffer. As soon as `log` starts running the benchmark, it begins periodically sampling the value of the performance counters in an internal trace buffer. Sampling periods can be specified, but we have observed that selecting a value in the range of 5 ms results in no statistically perceivable impact over our class of applications. Once the benchmark completes, `log` lowers the I/O pin and outputs a performance trace file. Once the application completes, we can synchronize both traces – the power trace generated by the LabView program and the performance trace generated by `log` – using post-processing scripts.

An alternate method is to send power measurements and the I/O pin voltage from the data-acquisition system to the primary system via UDP network packets. There is a delay involved in receiving packets, and the sampling rate must be slower to accommodate the extra overhead; the benefit of this approach is a consolidated output file with power and performance data for experiments with low-resolution sampling.

In either case, synchronization of the traces is often imperfect and achieving proper alignment is critical for measuring accurate power and performance. We have used a variety of post-processing techniques, trimming off the ends of both traces and counting the number of samples in each, to get approximate time correlations. Better temporal correlation of the power and performance data is an on-going challenge.

IV. Conclusions

Power and energy concerns are now on par with performance concerns for computer system design and management. As computer system designs evolve toward high-efficiency from purely high-performance, systems are now adopting active power management mechanisms like DVFS whose power and performance impact is determined by workload characteristics.

In order to model systems and evaluate power-management options, researchers need benchmarks that evaluate power and performance together. This paper discusses the three sources of variability, all of which must be exposed by the benchmarking process and then describes some practical issues we encountered while adding power measurement to the benchmarking process.

As new benchmarks specifically designed for joint power and performance evaluation are developed, designers need to bear in mind several considerations. First, variability in intensity over time, in workload type, and across individual systems is growing; researchers and users need benchmarks which expose all forms of variability. Second, producing correlated, meaningful measurements of two different quantities – power and performance – is much more difficult than measuring just one of them alone. Looking forward, we believe that standardized power-performance benchmarks should provide a measure of power-management effectiveness, as well as sheer performance or power, to facilitate comparisons between system power management options.

V. Prior Publication

This paper was previously presented at the 2007 SPEC Benchmarking Workshop, held on January 21, 2007, in Austin, Texas.

References

[1] U.S. Environmental Protection Agency, “Server energy measurement protocol,” Nov. 3 2006. http://www.energystar.gov/ia/products/downloads/Finalserverenergyprotocol-v1.pdf.

[2] K. Rajamani and C. Lefurgy, “On Evaluating Request-Distribution Schemes for Saving Energy in Server Clusters,” in *IEEE International Symposium on Performance Analysis of Systems and Software*, pp. 111–122, March 2003.

[3] W. D. Smith, “TPC-W: Benchmarking an ecommerce solution.” The Transaction Processing Performance Council, Feb. 2000.

[4] S. Ghiasi and W. Felter, “Cpu packing for multiprocessor power reduction,” in *Power Aware Computer Systems* (B. Falsafi and T. N. Vijaykumar, eds.), Springer-Verlag, December 2003.

[5] K. Rajamani, H. Hanson, J. Rubio, S. Ghiasi, and F. Rawson, “Application-Aware Power Management,” in *Proceedings of the 2006 IEEE International Symposium of Workload Characterization (IISWC-2006)*, October 2006.

[6] C. Isci and M. Martonosi, “Runtime power monitoring in high-end processors: Methodology and empirical data,” in *36th Annual ACM/IEEE International Symposium on Microarchitecture*, December 2003.

[7] K. Bernstein, D. J. Frank, A. E. Gattiker, W. Haensch, B. L. Ji, S. R. Nassif, E. J. Nowak, D. J. Pearson, and N. J. Rohrer, “High-performance CMOS variability in the 65-nm regime and beyond,” *The IBM Journal of Research and Development*, vol. 50, pp. 433–449, July/September 2006.

[8] A. Devgan and S. Nassif, “Power variability and its impact on design,” in *Proceedings of the 18th International Conference on VLSI Design*, pp. 679–682, January 2005.

[9] C. Lefurgy, K. Rajamani, F. Rawson, W. Felter, M. Kistler, and T. W. Keller, “Energy Management for Commercial Servers,” in *IEEE Computer Special Issue on Power- and Temperature-Aware Computing*, December 2003.

[10] Chris Calwell and Arshad Mansoor, “AC-DC Server Power Supplies: Making the Leap to Higher Efficiency,” in *Applied Power Electronics Conference and Exposition (APEC)*, March 2005.

[11] Radisys Corporation, "Endura LS855 Product Data Sheet." http://www.radisys.com/oem_products/ds-page.cfm?productdatasheetsid=1158, Oct. 10 2004.