Workshop on Power Aware Computing and Systems (HotPower ’08)

December 7, 2008
San Diego, CA

Summarized by Alva L. Couch (alva@usenix.org) and Kishore Kumar (kishoreguptaos@gmail.com)

HotPower ’08 depicts a very different approach to computing from that to which the average USENIX member may be accustomed. In a power-centric view of computing, units of measurement are translated into units representing power requirements. “Execution times” are converted into their corresponding power requirements, measured in watts. “Execution cycles” are converted into their corresponding energy requirements, measured in nano-joules. Power is (of course) energy over time: One joule is one watt/second. This energy-aware view of computing—while quite enlightening—takes some getting used to.

The goal of energy-aware computing is not just to make algorithms run as fast as possible, but also to minimize energy requirements for computation, by treating energy as a constrained resource like memory or disk. From a power/energy point of view, a computing system (or ensemble of systems) is “energy proportional” if the amount of energy consumed by the system is proportional to the amount of computational work completed by the system. True energy proportionality is impossible because of the baseline energy cost of keeping systems running even when idle, but one can come close to energy proportionality by powering up servers and/or subsystems only when needed and keeping them powered down (or, perhaps, running at a slower speed or power level) otherwise.

One thing that makes energy-aware computing challenging is that there is a straightforward inverse relationship between energy requirements and execution time, which often requires making a time/energy tradeoff. It is acceptable for some tasks to take longer times so that they can in turn require very little energy to accomplish (e.g., in embedded sensor systems that harvest power from RFID readers). In other cases, for time-critical tasks it is appropriate to balance task completion delay against energy requirements.

Similarly (but perhaps less obviously), there is also an inverse relationship between energy requirements and reliability. Reliability is usually implemented through hardware redundancy, and redundancy means in turn more power consumption. This redundancy can take subtle forms, such as whether a disk is powered up or its data and changes are cached in volatile memory instead.

HotPower is a gathering point for a diverse community of many kinds of researchers, ranging from software experts concentrating on algorithms for reducing power consumption to hardware designers and testers studying the effects of hardware design choices. This community has in a very short time developed its own acronyms and specialized language which can be difficult for a newcomer to grasp. For example, DVFS stands for Dynamic Voltage/Frequency Scaling, which represents the ability to run a CPU or subsystem at several different speeds and/or voltages with varying power requirements. Required background for understanding the papers includes the functional relationships among computing, power consumption, and cooling, as well as the basics of energy transfer including, for example, the relationship between the energy in a capacitor and the observed voltage difference between its contacts. [Editor’s note: Rudi van Drunen’s article in this issue discusses power in electrical terms.]

SCHEDULING AND CONTROL

- ***Memory-aware Scheduling for Energy Efficiency on Multicore Processors***
  Andreas Merkel and Frank Bellosa, University of Karlsruhe

Andreas Merkel and Frank Bellosa presented an energy-efficient co-scheduling algorithm to avoid memory contention problems in multi-core systems. Memory access power requirements depend upon processor architecture, including whether a set of processor cores shares one L2 cache. One approach to avoiding contention is to schedule tasks with different characteristics (memory-bound and CPU-bound) on each core. Another approach, called "sorted scheduling," is to reorder program blocks for processes so that only one memory-intensive block is scheduled at a time. Scheduling algorithms were tested by comparing their performance to DVFS with SPEC CPU2006 on Linux. DVFS performed better only for memory-bound tasks.

- ***Delivering Energy Proportionality with Non Energy-Proportional Systems—Optimizing the Ensemble***
  Niraj Tolia, Zhikui Wang, Manish Marwah, Cullen Bash, Parthasarathy Ranganathan, and Xiaoyun Zhu, HP Labs, Palo Alto

Niraj Tolia et al. showed that it is possible to use optimized techniques to approximate energy-proportional behavior at ensemble level. An "ensemble" is a logical collection of servers and could range from a rack-mount enclosure of blades to an entire datacenter. One can approach energy proportionality by using a virtual machine migration controller that powers machines up or down, in addition to dynamic voltage and frequency scaling in response to demand changes. A power- and workload-aware cooling controller optimizes the efficiency of cooling equipment such as server fans. A case study examines the balance between server power and cooling power and compares several energy-saving approaches, including no DVFS, DVFS alone, and DVFS with simulated annealing for service consolidation. The last approach shows significant improvement over the former two, with some counterintuitive results, including that the cooling effect from a fan is not a linear function of power input to the fan; for optimal efficiency, one must run the fan at about 30% of its peak load.

MODELING

- ***A Comparison of High-Level Full-System Power Models***
  Suzanne Rivoire, Sonoma State University; Parthasarathy Ranganathan, Hewlett-Packard Labs; Christos Kozyrakis, Stanford University

Suzanne et al. used a common infrastructure to evaluate high-level full-system power models for a wide range of workloads and machines. The machines (8-core Xeon server, a mobile file server, etc.) that they used span three different processor families: Xeon, Itanium, and Turion. Several models were compared, including a linear model based on CPU utilization, a linear model based on CPU and disk utilization, and a power-law model based on CPU utilization. To evaluate the models, they used SPEC CPU, SPEC JBB, and also memory-stress and IO-intensive benchmarks. The results of these tests illustrate tradeoffs between simplicity and accuracy, as well as the limitations of each type of model. Performance-counter-based power models give more accurate results compared with other types of power models, though these are processor-specific and thus nonportable. Counterintuitively, using a parameter in a model that is not utilized (e.g., disk in a memory-intensive application) leads to overprediction and error.

- ***Run-time Energy Consumption Estimation Based on Workload in Server Systems***
  Adam Lewis, Soumik Ghosh, and N.-F. Tzeng, University of Louisiana

Adam Lewis et al. showed statistical methods to develop system-wide energy models for servers. They developed a linear regression model based on DC current utilization, L2 cache misses, disk transactions, and ambient and die (CPU) temperatures. When evaluated with the SPEC CPU2006 benchmark programs, their model exhibited prediction errors between 2% and 3.5%. This technique shows promise, but the audience questioned whether this would apply as accurately to uncontrolled, real-world loads. Results suggest that additional performance data—beyond the performance counters that are provided by a typical processor—are needed to get a more accurate prediction of system-wide energy consumption.

POWER IN EMBEDDED

- ***Getting Things Done on Computational RFIDs with Energy-Aware Checkpointing and Voltage-Aware Scheduling***
  Benjamin Ransford, Shane Clark, Mastooreh Salajegheh, and Kevin Fu, University of Massachusetts Amherst

A *computational RFID unit* (CRFID) is a computational unit with no battery that utilizes power harvesting from RFID readers to accomplish computational tasks. CRFIDs utilize extremely low-power hardware, such as the Intel WISP, which consumes 600 micro-amperes when active and 1.5 micro-amperes when sleeping. Units such as the WISP can accomplish useful work in multiple steps—as power becomes available—by dynamic checkpointing and restore. The checkpointing strategy assumes a linear relationship between input voltage and available power, such as that from a capacitor used as a power storage device. A voltage detector senses remaining power and checkpoints the processor's current state to flash memory when the power available drops below a given threshold. This strategy has promise in several application domains, including medical electronics, sensor networks, and security.

- ***The True Cost of Accurate Time***
  Thomas Schmid, Zainul Charbiwala, Jonathan Friedman, and Mani B. Srivastava, University of California, Los Angeles; Young H. Cho, University of Southern California

Maintaining a highly accurate concept of wall clock time for otherwise autonomous wireless nodes has a high power cost. To reduce that cost, a hybrid architecture is proposed in which a relatively higher-power but highly accurate crystal clock circuit is paired with a low-power, low-frequency oscillator on a single chip. The more accurate clock sleeps much of the time and is polled to reset a less accurate LFO clock when needed. The result is a low-power clock chip in a 68-pin configuration that has about 125,000 gates, consumes 20 microwatts on average, and has a 1.2-volt core voltage.

POWER IN NETWORKS

- ***Greening the Switch***
  Ganesh Ananthanarayanan and Randy H. Katz, University of California, Berkeley

Network switches are often provisioned for peak loads, but this is not power-efficient. To reduce power requirements, one can selectively power-down idle switch ports, utilize a separate "shadow port" to accept traffic from a set of powered-down ports, or utilize a "lightweight switch" as a slower, low-power alternative to a fast, higher-power-consumption main switch. If a port is powered down, one loses incoming traffic and queues outgoing traffic on the port. A "shadow port" can receive data from other powered-down ports. A shadow port can reduce but not eliminate data loss, because it can only receive one packet at a time from a group of ports. By contrast, a "lightweight alternative switch" replaces the regular switch and allows it to be completely powered down during nonpeak times. The authors compare these two strategies through trace-driven simulation of the results of the strategy on seven days of network traces from a Fortune 500 company. Lightweight alternative switches turn out to be the most cost-effective of these two strategies, saving, according to the simulation, up to 32% of power.

- ***Hot Data Centers vs. Cool Peers***
  Sergiu Nedevschi and Sylvia Ratnasamy, Intel Research; Jitendra Padhye, Microsoft Research

Should a service be provided in a datacenter or as a peer-to-peer application? This paper analyzes the power requirements of peer-to-peer versus centralized service provisioning in a novel way, by considering the "baseline cost" of the existing systems before the service is added. If one is going to be running underutilized desktop computers anyway, then the added power and cooling requirements from a peer-to-peer application are shown to be cost-effective. The assumptions of the paper were quite controversial to the audience, however; for example, why are the underutilized desktops still powered up when there is nothing to do?

POSTERS

- ***Analysis of Dynamic Voltage Scaling for System Level Energy Management***
  Gaurav Dhiman, University of California, San Diego; Kishore Kumar Pusukuri, University of California, Riverside; Tajana Rosing, University of California, San Diego

Dynamic Voltage/Frequency Scaling (DVFS) is commonly used to save power by lowering the clock rate of a processor at nonpeak periods. DVFS does better at saving power than putting any idle CPU into its lowest-power operating state, but it might not save as much power as shutting down an idle processor and putting memory into self-refresh mode. This perhaps counterintuitive result is predicted by trace-driven simulation.

- ***Energy Aware Consolidation for Cloud Computing***
  Shekhar Srikantaiah, Pennsylvania State University; Aman Kansal and Feng Zhao, Microsoft Research

Energy-aware consolidation is the process of migrating applications and/or services to a small number of physical servers to allow excess servers to be shut down. Simulations demonstrate that packing applications into servers at higher than 50% cumulative CPU load is actually less energy-efficient than keeping the effective load below 50% of peak, due to wasted energy from server thrashing. Similarly, co-locating services so that disk utilization exceeds 70% of peak load leads to energy loss.

- ***Energy-Aware High Performance Computing with Graphic Processing Units***
  Mahsan Rofouei, Thanos Stathopoulos, Sebi Ryffel, William Kaiser, and Majid Sarrafzadeh, University of California, Los Angeles

Low-power energy-aware processing (LEAP) can be applied to code running inside a graphics processor on the video board of a desktop computer. Power savings for CPU-bound applications (e.g., convolution) can be as high as 80%, as demonstrated via trace-driven simulation.

- ***Augmenting RAID with an SSD for Energy Relief***
  Hyo J. Lee, Hongik University; Kyu H. Lee, Purdue University; Sam H. Noh, Hongik University

A solid-state disk (SSD) can be used as a read/write cache for a log-structured filesystem on a RAID disk array. The read-write flash cache is flushed when 90% full. Simulations of this architecture predict power savings of 14% at peak load and 10% at low load.

- ***Workload Decomposition for Power Efficient Storage Systems***
  Lanyue Lu and Peter Varman, Rice University

The traditional definition of "quality of service" (QoS) defines thresholds for response time that cannot be exceeded without penalty. By redefining QoS in statistical terms, one can reduce power requirements for service provision by 50% to 70%. The new definition of QoS allows response times for some percentage of requests to exceed each QoS threshold. Power savings arising from this change are estimated via trace-driven simulation.

- ***CoolIT: Coordinating Facility and IT Management for Efficient Datacenters***
  Ripal Nathuji, Ankit Somani, Karsten Schwan, and Yogendra Joshi, Georgia Institute of Technology

CoolIt is a temperature-aware virtual architecture. A sensing subsystem monitors activity of the virtual architecture, while a cooling control subsystem solves a linear program to optimally control cooling fans. This approach is implemented for Xen in an "ambient intelligent load manager" (AILM).

POWER IN STORAGE

- ***On the Impact of Disk Scrubbing on Energy Savings***
  Guanying Wang and Ali R. Butt, Virginia Polytechnic Institute and State University; Chris Gniady, University of Arizona

Guanying et al. proposed a new metric called the "energy-reliability product (ERP)" to capture the combined performance of energy saving and reliability improving approaches of disks. This metric is a product of energy savings (by spinning down the disk) and reliability improvement in terms of "mean time to data loss." The authors used trace-driven simulations of enterprise applications, such as the Mozilla Web browser, and studied the effects of disk scrubbing and energy management on these applications. Finally, through this study, they showed that ERP can help to identify efficient ways to distribute disk idle time for energy and reliability management.

- ***Empirical Analysis on Energy Efficiency of Flash-based SSDs***
  Euiseong Seo, Seon Yeong Park, and Bhuvan Urgaonkar, Pennsylvania State University

Euiseong et al. analyzed the power consumption pattern of solid-state disk drives (SSDs) with a microbenchmark (using the "DIO tool" workload generator) to show the characteristics for read and write operations at the device level, as well as a macro-benchmark "filebench" to measure real-world behavior of the device. The authors measured differences in terms of power consumption between SSDs and hard-disk drives (HDDs) and also common characteristics shared by SSDs. One audience member asked about the role of logical block lookup tables in improving the reliability of SSDs (by minimizing erasures), and how that affects power requirements. In particular, if a traditional filesystem is written to a SSD, the superblock is not especially vulnerable, because it is logically moved each time it is updated.

CHALLENGES PANEL

Moderator: Feng Zhao, Microsoft Research
Panelists: James Hamilton, Microsoft Research; Randy Katz, University of California, Berkeley; Jeffrey Mogul, Hewlett-Packard Labs

James Hamilton (Microsoft) began his presentation with the question, "Where does power go and what to do about it?" Power losses are easier to track than cooling. Seven watts of each server watt are lost from translation inefficiency. Cooling systems employ a large number of conversions: a "catastrophically bad design." About 33% of cooling power cost is due to mechanical losses. Pushing air 50 feet is catastrophically bad. A secondary problem is evaporative water loss from cooling systems, estimated at 360,000 gallons of water a day for a site. Several creative approaches to the problem include "air-side economization" (open the window!) and cooperative expendable micro-slice servers, with four times the work per watt of current servers.

Randy Katz (Berkeley) asked, instead, "What if the energy grid were designed like the Internet?" Current energy grid technology is a remnant of the machine age, and expertise in power distribution has largely disappeared from academia. As a fresh approach, we can apply principles of the Internet to energy. First, we push intelligence to the edges and concentrate on lower-cost incremental deployment. Enhanced reliability and resilience arise from the same sources as Internet reliability and resilience. The result is the "LoCal-ized datacenter" that is based upon DC distribution rather than AC and contains battery backups (or other forms of power storage) in each rack. This allows flexible use of any kind of power with minimal conversion loss, including stored energy and solar power.

Jeff Mogul (HP) encouraged us to "look between the street lamps" for the next generation of power Ph.D. thesis topics. The street lamps include component power, control theory, and moving work around. These areas are well-explored. There are many topics that fall "between the street lamps," including tradeoffs between reliability and power use, matching customer needs to theoretical solutions, and making it easier to write energy-aware programs. Key challenges to understanding include the boundaries between areas, as well as energy inputs beyond the computer's power supply, including the energy cost of building and disposing of computing hardware.

A spirited discussion ensued in which there were many contributors.

A key principle is "Do nothing well." In other words, stop trying to optimize the effect of every joule going into our hardware; instead, look for median approaches to the problem.

One possible approach is detouring work: Instead of paying for peak energy load, store energy from nonpeak times. However, it remains very difficult to store energy. Innovative approaches include energy harvesting and even compressed-air storage.

Building more power plants to satisfy datacenter demand is not the only way to deal with increasing power demand. We don't know yet how to produce an application-independent layer that does that, and programmers may have difficulties with the resulting level of abstraction.

Another challenge is that of sharing data for mutual benefit. Data privacy is a major problem, but if someone could define what an interesting power trace might be, smaller players could contribute. Alternatively, using open-source applications such as Hadoop allows one to collect power data for one's own application.

There is also a seeming contradiction in the way people and lawmakers react to cooling strategies. If one puts heat into a river, environmentalists are concerned. If one puts heat into the air instead, no one seems to care.

Another potential savings strategy is to reevaluate how datacenters are cooled. We may not want to cool the whole datacenter to 62 degrees. We may want to cool everything to 89 degrees. But then there's no margin for error. In raising the total machine room temperature, we would be operating "nearer to the edge of the hardware function envelope" and any failure of cooling might lead to massive failures of hardware.

The recent rise of cloud computing poses its own power challenges. If everybody outsources storage to Amazon and everyone gets a surge of traffic (e.g., the day after Thanksgiving), do our computer systems have a credit meltdown? What if the whole "ecosystem" undergoes the same set of unforeseen changes?