2010 Workshop on Power Aware Computing and Systems (HotPower '10)

October 3, 2010
Vancouver, BC, Canada

Impact of Hardware Trends

Summarized by John McCullough (jmccullo@cs.ucsd.edu)

Dynamic Voltage and Frequency Scaling: The Laws of Diminishing Returns

Etienne Le Sueur and Gernot Heiser, NICTA and University of New South Wales

Etienne Le Sueur observed that dynamic voltage and frequency scaling is a technique commonly used to reduce the power of a running system. The dynamic power of a system scales linearly with frequency and quadratically with voltage. If we can drop the voltage and frequency, then we can expect a drop in power. In 1994 we saw 25–65% savings up through 2009, when work saw a 30% energy savings with a meager 4% performance loss. Unfortunately, there is also a static power component that includes leakage current, memory refresh power, hard drive motors, etc. Thus, running more slowly can potentially degrade the overall power.

The authors consider the mcf and gzip benchmarks and observe that mcf gets the largest benefit from DVFS because it is memory bound rather than CPU bound. Looking across three different Opteron CPUs, they found that the older two had a power-optimal point around 1.6GHz but that, with a more modern processor, DVFS was ineffective at saving power. One shortcoming of this technique is that it assumes that the computer is not consuming any power after the benchmark completes. To consider the idle power following the experiment run, Etienne described their padding methodology to measure the idle power up to the time of the longest running DVFS-scaled benchmark. Using this technique, we start to see some reduction in energy, but we only see improvements in energy delay if running on all system cores.

Moving forward, we can expect that shrinking feature sizes and increasing cache size and memory bandwidth will make improvements by scaling down via DVFS even less likely. Surprisingly, however, features such as Intel's Turbo-Boost can actually reduce total power by scaling up the clock frequency and racing to idle effectively.

An audience member asked how DVFS can impact embedded platforms. Etienne observed that the CPU power can be very low relative to the total system power and that DVFS won't be able to impact total power. Another audience member observed that AMD has introduced DVFS on the cache and memory controllers.

A Case for Opportunistic Embedded Sensing in Presence of Hardware Power Variability

Lucas Wanner, Charwak Apte, Rahul Balani, Puneet Gupta, and Mani Srivastava, University of California, Los Angeles

Puneet Gupta demonstrated how shrinking feature sizes leads to immense variability in the physical manifestation of hardware designs. For example, in an experimental 80-core processor the performance spread across cores on a single die was 80%. The degree of variability is not tied to the design alone, as the same design sourced from different manufacturers can get different degrees of variability. Furthermore, aging can cause wires to slow and reduce performance by 20–30%. Today, variability is masked by guard bands. Processor manufacturers typically bin processors by functional speed, with space for any aging effects. Unfortunately, scaling with the guard band is much worse than the nominally achievable results.

Puneet advocates that exposing aspects of this variability to software can allow improved functionality. Such exposure could happen via active measurement or prior testing, but it can have a significant impact. For instance, in sensing applications the sleep power is dominant and can affect the amount of data that can be acquired on a given power budget. The authors found that for 10 off-the-shelf Cortex M3 processors, the active power varied by 10%, but the sleep power varied by 80%. Furthermore, the sleep and active power vary with temperature. Using a power model calibrated to temperature and sensors for power, Puneet demonstrated that they can achieve more effective sensing by energy-aware duty cycling. Thus, a node with lower sleep power can sample 1.8x more data than it would have if all nodes were timed to the worst sleep power. Moving forward, one challenge is to discover the right interface for exposing the variability and sense data to software.

An audience member observed that this data is already integrated in modern CPU power management units for managing the frequency within temperature and power bounds. Puneet responded that it is important to actually expose this information to the software layer, which most of these techniques fail to do. Another audience member asked about the overhead of these techniques. Puneet observed that the information is already being collected for quality control but it is not exposed to software. Finally, an audience member asked about the difficulty managing other system components. Puneet said that the complexity would depend on the abstraction.

Invited Talk

Summarized by John McCullough (jmccullo@cs.ucsd.edu)

Datacenter Power Efficiency: Separating Fact from Fiction

Kushagra Vaid, Microsoft Corporation

Kushagra Vaid posed the question of how to maximize power efficiency at a large scale. A data center consists of a power substation, chillers, batteries, an ops room, generators, fuel, and computing equipment. The efficiency of a facility is often measured by PUE, which is the facility power divided by the IT equipment power. Common PUE values range around 1.5–2.0, but good deployments reach close to ideal at 1.05.

Kushagra observed that the cost of power is fairly low relative to the overall costs of the facility. Amortizing using a three-year replacement model, power consumption consists of 16% of the cost. Thus, reducing dynamic power and energy-proportional computing has limited benefits. The capital expense of servers accounts for the largest portion of the total cost of ownership.

To minimize costs in provisioning, Microsoft is moving towards modular data centers. A video showing the modules is available at http://www.microsoft.com/showcase/en/us/details/84f44749-1343-4467-8012-9c70ef77981c. The modules function using adiabatic cooling with outside air and only using top-of-rack fans to adjust for cooling/heating as appropriate. To reduce the cost of power, there are techniques for eliminating conversion steps. Surprisingly, running AC to the servers is not significantly different from DC in total conversion efficiency. Right-sizing for density suggests a sweet spot, with the lowest-voltage processor getting a surprising performance/watt/cost benefit. Right-sizing storage can be achieved by in-depth storage trace analysis to understand workload patterns and dynamic range to pack better.

Overall, researchers need to be sure that they take a holistic view. Current research areas are in optimal provisioning for high dynamic range workloads, addressing energy proportionality via system architecture innovations, power-aware task scheduling on large clusters, and energy-conscious programming using controlled approximation.

One audience member asked why their efficiency approaches don’t work in all data centers. Kushagra responded that getting all of the layers of UPSes and voltage conversion requires control of the entire data center, which few have the scale to accomplish. What are the implications of low-power CPUs in data centers? For Bing workloads, Xeon systems are 2.3x better in performance/watt/cost. Someone asked why they don’t turn off servers. Kushagra replied that it can lead to response time spikes, that turn-on time can be long, and that if you can turn off servers, you brought too many online or you are doing poorly at task scheduling.

Data Center I

Summarized by John McCullough (jmccullo@cs.ucsd.edu)

Analyzing Performance Asymmetric Multicore Processors for Latency Sensitive Datacenter Applications

Vishal Gupta, Georgia Institute of Technology; Ripal Nathuji, Microsoft Research

Asymmetric multicore processors (AMPs) are being explored in multiple dimensions, where cores are combined with different performance characteristics and deployed with different functional characteristics. Vishal Gupta presented their technique for understanding the impact that AMPs can have on data centers with respect to power and throughput.

Vishal described two use cases: energy scaling and parallel speedup. Energy scaling involves execution on a combination of the small and large cores to achieve the same computation within a deadline at lower power. Parallel speedup occurs when a larger core on an AMP can execute serial sections faster. To ascertain the effects of these use cases, Vishal treats each processor as an M/M/1 queue, which models processing times as an exponential distribution where the processing time is parameterized in proportion to the chip area. Overall completion time is parametrized by the paralellizable fraction of the code. Using this model, Vishal finds that for a higher fraction of parallelizable work, power savings increase with AMP use. While there are practical considerations, AMPs offer more potential for parallel speedup than for energy speedup.

Energy Conservation in Multi-Tenant Networks through Power Virtualization

Srini Seetharaman, Deutsche Telekom R&D Lab, Los Altos

Networks are typically power oblivious and it is hard for network users to ascertain the impact. By packing flows into fewer devices and turning off unused devices, the system can turn off individual ports and even entire switches. Given a multi-tenant data center, how can tenants be influenced to reduce their system usage? Switching from a flat rate for networking to a power-based price can incentivize power savings.

Srini Seetharaman proposed the idea of virtual power. Because network power is not proportional usage, the most intuitive definition for virtual power is to split the power consumption of a component over all sharing tenants. This has the effect of penalizing a tenant for being the only occupant and encourages reuse of pre-paid/pre-powered-on elements. An implementation is in progress but there are no results yet. The specific methods of billing and pricing can influence the outcome; for instance, auctions might introduce different behavior than allocations that degrade over time. In the future, the question is how we can achieve good performance while conserving power.

In the Q&A, Srini clarified that it makes more sense to conserve in a reactive mode, turning on devices as necessary. One audience member asked whether rate-based power differences can affect power. Srini replied that the power differences are typically small. The same person also asked whether the placement of tenants in the data center could penalize them in terms of the available pricing. Srini replied that this was a concern.

Data Center II

Summarized by Etienne Le Sueur (elesueur@cse.unsw.edu)

Energy Savings in Privacy-Preserving Computation Offloading with Protection by Homomorphic Encryption

Jibang Liu and Yung-Hsiang Lu, Purdue University

Yung-Hsiang Lu presented this paper, which discusses the issues that arise when compute-intensive tasks are offloaded from mobile devices to a centralized server. The main issue their work addresses is that of privacy, when sensitive data needs to be transferred off the mobile device, using a public network, to a server which may not be trusted.

Their work mitigates this privacy issue by protecting the data using a homomorphic encryption algorithm. Homomorphic encryption is unlike traditional encryption techniques in that computations can be done on the cipher-text itself rather than being decrypted first. This way, the server operating on the data need not actually know what information the data contains.

The use-case they describe in the paper deals with image-matching—for example, when a user of a mobile phone takes a photo and wants a remote server to do some analysis to try and determine what objects the photo contains. They used an iPad portable device and a server with a 2GHz CPU for processing the images, with evaluation based on how many positive matches were made. Using their modified Gabor filter, they were able to get a correct match approximately 80% of the time when the analysis was performed on the cipher-text.

The work seems promising, and a future direction was clearly given which will address the issues with noise in the encryption system.

An audience member asked whether there were other classes of functions where it makes sense to offload computation. They haven't reached a point where there's a general rule for when to apply the technique. How strong is the encryption and can it easily be broken? The strength of the encryption depends on the key length. The discussion was continued offline.

Green Server Design: Beyond Operational Energy to Sustainability

Jichuan Chang, Justin Meza, Parthasarathy Ranganathan, Cullen Bash, and Amip Shah, Hewlett Packard Labs

Justin Meza presented this paper, which discusses a way to quantify sustainability when designing data centers. The usual motivations for the work were given: e.g., reduction of carbon footprint, secondary costs (power and cooling), and government regulation.

To measure sustainability, several costs need to be addressed: extraction of materials, manufacture of systems, operation and infrastructure, and recycling once end-of-life is reached.

They use a metric called "exergy," which basically constitutes TCO (total cost of ownership) plus the energy required to manufacture devices. Objects such as servers accumulate "exergy" by breaking them down into components, such as hard drives and processors, and trying to determine the cost of the raw materials that go into making them.

Supply-chain information is included in the "exergy" calculations, which include cost of transportation. For an example server, they find that the cost of manufacturing the server was roughly 20%, 27% was infrastructure-based cost, and 53% was operational costs like power and cooling.

The next part of the talk discussed how using certain techniques to alter energy-efficiency affected "exergy." On one hand, they looked at consolidating servers (reducing total idleness) and on the other hand they looked at energy-proportional computing techniques, such as reducing CPU frequency when there is idleness. They found that if reducing total "exergy" is the goal, then consolidation is the best approach, but if reducing operational "exergy" is the goal, then energy-proportional computing techniques give better results.

GreenHDFS: Towards an Energy-Conserving, Storage-Efficient, Hybrid Hadoop Compute Cluster

Rini T. Kaushik, The University of Illinois at Urbana-Champaign and Yahoo! Inc.; Milind Bhandarkar, Yahoo! Inc.

Rini T. Kaushik presented the authors' attempt to leverage the distributed nature of the Hadoop file system. The basic premise is that the cluster of servers is divided into two sections: a hot section, which contains recently used data, and a cold section, which contains data that has been untouched for some time.

Initially, they did an analysis of the evolution of files in an HDFS (Hadoop Distributed File System) cluster, by looking at three months' worth of traces from Yahoo. They found that 90% of data is accessed within two days after a file is first created. Additionally, they found that 40% of data lies dormant for more than 20 days before it is deleted. This essentially means that data is "hot" for a short period after creation, and then "cold" for a much longer period.

Computation follows the 'hot" zones, and the "cold" zones see significant idleness and can be scaled down or turned off. They determined the best division was 70% hot and 30% cold. They claim that these techniques can save 26% of the energy used in the cluster they were testing.

An audience member asked the presenter to clarify how they were able to save 26% energy. A large number of the servers in the "cold" zone were in fact never turned on during the three-month simulation.

Myths, Modeling, and Measurement

Summarized by Lucas Wanner (wanner@ucla.edu)

Demystifying 802.11n Power Consumption

Daniel Halperin, University of Washington; Ben Greenstein and Anmol Sheth, Intel Labs Seattle; David Wetherall, University of Washington and Intel Labs Seattle

Anmol Sheth started by observing that WiFi is becoming ubiquitous and that the increasing bandwidth demands of networked applications are leading to the widespread adoption of 802.11n, the latest version of the standard. In battery-operated devices such as mobile phones, radio interfaces can account for a significant portion of total power budget. There is little data available to help designers operate 802.11n devices in an energy-efficient way. This work presents measurements of 802.11n in various configurations and modes of operation.

802.11n devices may use multiple active RF chains. The characterization study in this work found that power increases sub-linearly with additional antennas, and increase in signal processing power is negligible. Power consumption is hence not a multiple of active RF chains and is asymmetric for transmission and reception. Nevertheless, the use of wider RF channels is more power-efficient than multiple spatial streams.

Another source of energy efficiency for 802.11n communication is "racing to sleep," i.e., transmitting data in high-rate bursts and subsequently putting the card in sleep mode. Because sleep mode may use approximately 10x less power than simply leaving the card in reception mode all the time, this may lead to significant savings.

The first question addressed the issue of energy-efficient operation for bandwidth-intensive applications, such as streaming video. In these cases, racing-to-sleep may be impossible. A second question was about backward compatibility between n and g devices in the same network. This can potentially decrease energy efficiency, as devices must operate at the lowest common denominator. Finally, the test environment was discussed: in the tests conducted in the study, the hosts were in close proximity. Further work is required to evaluate lower-quality links typical of homes and offices.

Chaotic Attractor Prediction for Server Run-time Energy Consumption

Adam Lewis, Jim Simon, and Nian-Feng Tzeng, University of Louisiana

Full system power models are used to estimate energy consumption in servers. This is accomplished by looking at complete system energy and trying to approximate the energy used by various components. Linear methods are simple and have fairly average median error but potentially very high maximum errors. Power traces for a series of benchmarks suggest both correlation and chaotic behavior between samples.

This work showed that models constructed from autoregressive methods demonstrate behavior that makes them problematic for predicting server energy consumption. This work proposed a Chaotic Attractor Predictor which overcomes the limitations of the previous linear regression-based methods by addressing the non-linear aspects of energy consumption in time and captures the underlying chaotic behavior of the system.

During the questions session, it was pointed out that only single-threaded applications were used to predict power consumption in the work. This may be one of the reasons for the non-linear behavior found in power consumption, as the power manager can put the CPU to sleep when it is idle.

Automatic Server to Circuit Mapping with the Red Pills

Jie Liu, Microsoft Research

The objective of this work is to map servers in a data center to the circuit that powers them. Due to complex wiring setups, it's hard to identify which circuit powers each server. Having this information would be beneficial for failover analysis, accounting, power provisioning, and balancing.

The basic idea in this work is to manipulate a server's workload to generate a power signature that can be identified by circuit power measurements, using the power line as a communication channel. A pure, single-frequency signal would be ideal for this identification, but is hard to generate by manipulating CPU utilization. Periodic square wave signals are easier to generate. In the Red Pill system, a manager requests an identification signal from the server through the network and detects this through the power measurement connected to the manager. IDs with 64 samples can be detected with high probability, even with fairly low amplitude signals. Each signature with 64 samples would take about 16 minutes for detection. As the number of servers increases beyond 20, detection likelihood decreases.

Is it hard to identify server-to-power circuit mapping in general, or is this only a problem with "incorrect deployments? Because of the dynamic nature of datacenter deployments, this is a fairly common problem. How can correctness be verified in the system? In the tests conducted for the paper, the mapping (ground truth) was known. In deployment systems, metrics of confidence could be included, and repeated measurements could be used to increase accuracy.