

Power Provisioning for a Warehouse-sized Computer

Xiaobo Fan

Wolf-Dietrich Weber

Luiz André Barroso

Google Inc.
1600 Amphitheatre Pkwy
Mountain View, CA 94043
{xiaobo,wolf,luiz}@google.com

ABSTRACT

Large-scale Internet services require a computing infrastructure that can be appropriately described as a warehouse-sized computing system. The cost of building datacenter facilities capable of delivering a given power capacity to such a computer can rival the recurring energy consumption costs themselves. Therefore, there are strong economic incentives to operate facilities as close as possible to maximum capacity, so that the non-recurring facility costs can be best amortized. That is difficult to achieve in practice because of uncertainties in equipment power ratings and because power consumption tends to vary significantly with the actual computing activity. Effective power provisioning strategies are needed to determine how much computing equipment can be safely and efficiently hosted within a given power budget.

In this paper we present the aggregate power usage characteristics of large collections of servers (up to 15 thousand) for different classes of applications over a period of approximately six months. Those observations allow us to evaluate opportunities for maximizing the use of the deployed power capacity of datacenters, and assess the risks of over-subscribing it. We find that even in well-tuned applications there is a noticeable gap (7 - 16%) between achieved and theoretical aggregate peak power usage at the cluster level (thousands of servers). The gap grows to almost 40% in whole datacenters. This headroom can be used to deploy additional compute equipment within the same power budget with minimal risk of exceeding it. We use our modeling framework to estimate the potential of power management schemes to reduce peak power and energy usage. We find that the opportunities for power and energy savings are significant, but greater at the cluster-level (thousands of servers) than at the rack-level (tens). Finally we argue that systems need to be power efficient across the activity range, and not only at peak performance levels.

Categories and Subject Descriptors: C.0 [Computer Systems Organization]: General - System architectures; C.4 [Computer Systems Organization]: Performance of Systems - Design studies, Measurement techniques, Modeling techniques.

General Terms: Measurement, Experimentation.

Keywords: Power modeling, power provisioning, energy efficiency.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ISCA'07, June 9–13, 2007, San Diego, California, USA.

Copyright 2007 ACM 978-1-59593-706-3/07/0006 ...\$5.00.

1. INTRODUCTION

With the onset of large-scale Internet services and the massively parallel computing infrastructure that is required to support them, the job of a computer architect has expanded to include the design of warehouse-sized computing systems, made up of thousands of computing nodes, their associated storage hierarchy and interconnection infrastructure [3]. Power and energy are first-order concerns in the design of these new *computers*, as the cost of powering server systems has been steadily rising with higher performing systems, while the cost of hardware has remained relatively stable. Barroso [2] argued that if these trends were to continue the cost of the energy consumed by a server during its lifetime could surpass the cost of the equipment itself. By comparison, another energy-related cost factor has yet to receive significant attention: the cost of building a datacenter facility capable of providing power to a group of servers.

Typical datacenter building costs fall between \$10 and \$20 per deployed Watt of peak critical power (power for computing equipment only, excluding cooling and other ancillary loads) [25], while electricity costs in the U.S. are approximately \$0.80/Watt-year (less than that in areas where large datacenters tend to be deployed). Unlike energy costs that vary with actual usage, the cost of building a datacenter is fixed for a given peak power delivery capacity. Consequently, the more under-utilized a facility, the more expensive it becomes as a fraction of the total cost of ownership. For example, if a facility operates at 85% of its peak capacity on average, the cost of building the facility will still be higher than all electricity expenses for ten years of operation¹. Maximizing usage of the available power budget is also important for existing facilities, since it can allow the computing infrastructure to grow or to enable upgrades without requiring the acquisition of new datacenter capacity, which can take years if it involves new construction.

The incentive to fully utilize the power budget of a datacenter is offset by the business risk of exceeding its maximum capacity, which could result in outages or costly violations of service agreements.

Determining the right deployment and power management strategies requires understanding the simultaneous power usage characteristics of groups of hundreds or thousands of machines, over time. This is complicated by three important factors: the rated maximum power (or *nameplate* value) of computing equipment is usually overly conservative and therefore of limited usefulness; actual consumed power of servers varies significantly with the amount of activity, making it hard to predict; different applications exercise large-scale systems differently. Consequently only the monitoring

¹Assumes typical Tier-2 [25] datacenter costs of \$11/Watt of critical power and a 50% energy overhead for cooling and conversion losses.

of real large-scale workloads can yield insight into the aggregate load at the datacenter level.

In this paper we present the power usage characteristics of three large-scale workloads as well as a workload mix from an actual datacenter, each using up to several thousand servers, over a period of about six months. We focus on critical power, and examine how power usage varies over time and over different aggregation levels (from individual racks to an entire cluster). We use a light-weight yet accurate power estimation methodology that is based on real-time activity information and the baseline server hardware configuration. The model lets us also estimate the potential power and energy savings of power management techniques, such as power capping and CPU voltage/frequency scaling.

To our knowledge, this is the first power usage study of very large scale systems running real live workloads, and the first reported use of power modeling for power provisioning. Some of our other key findings and contributions are:

- The gap between the maximum power actually used by large groups of machines and their aggregate theoretical peak usage can be as large as 40% in datacenters, suggesting a significant opportunity to host additional machines under the same power budget. This gap is smaller but still significant when well-tuned large workloads are considered.
- Power capping using dynamic power management can enable additional machines to be hosted, but is more useful as a safety mechanism to prevent overload situations.
- We observe time intervals when large groups of machines are operating near peak power levels, suggesting that power gaps and power management techniques might be more easily exploited at the datacenter-level than at the rack-level.
- CPU voltage/frequency scaling, a technique targeted at energy management, has the potential to be moderately effective at reducing peak power consumption once large groups of machines are considered.
- We evaluate the benefits of building systems that are power-efficient across the activity range, instead of simply at peak power or performance levels.

2. DATACENTER POWER PROVISIONING

It is useful to present a typical datacenter power distribution hierarchy since our analysis uses some of those concepts (even though the exact power distribution architecture can vary significantly from site to site).

Figure 1 shows a typical Tier-2 [25] datacenter facility with a total capacity of 1 MW. The rough capacity of the different components is shown on the left side. A medium voltage feed (top) from a substation is first transformed down to 480 V. It is common to have an uninterruptible power supply (UPS) and generator combination to provide back-up power should the main power fail. The UPS is responsible for conditioning power and providing short-term back-up, while the generator provides longer-term back-up. An automatic transfer switch (ATS) switches between the generator and the mains, and supplies the rest of the hierarchy. From here, power is supplied via two independent routes in order to assure a degree of fault tolerance. Each side has its own UPS that supplies a series of power distribution units (PDUs). Each PDU is paired with a static transfer switch (STS) to route power from both sides and assure an uninterrupted supply should one side fail. The PDUs are rated on the order of 75 - 200 kW each. They further transform the voltage (to 110 V in the US) and provide additional conditioning and monitoring equipment, as well as distribution panels from

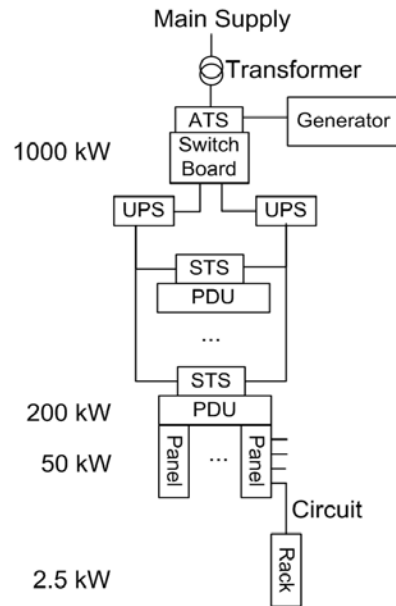


Figure 1: Simplified datacenter power distribution hierarchy.

which individual circuits emerge. Circuits power a rack’s worth of computing equipment (or a fraction of a rack).

Power deployment decisions are generally made at three levels: rack, PDU, and facility or datacenter. Here we consider a rack as a collection of computing equipment that is housed in a standard 19" wide, and 7' tall enclosure. Depending on the types of servers, a rack can contain between 10 and 80 computing nodes, and is fed by a small number of circuits. Between 20 and 60 racks are aggregated into a PDU.

Enforcement of power limits can be physical or contractual in nature. Physical enforcement means that overloading of electrical circuits will cause circuit breakers to trip, and result in outages. Contractual enforcement is in the form of economic penalties for exceeding the negotiated load (power and/or energy). Physical limits are generally used at the lower levels of the power distribution system, while contractual limits show up at the higher levels. At the circuit level, breakers protect individual circuits, and this limits the power that can be drawn out of that circuit². Enforcement at the circuit level is straightforward, because circuits are typically not shared between users. As we move higher up in the power distribution system, larger power units are more likely to be shared between multiple different users. The datacenter operator must provide the maximum rated load for each branch circuit up to the contractual limits and assure that the higher levels of the power distribution system can sustain that load. Violating one of these contracts can have steep penalties because the user may be liable for the outage of another user sharing the power distribution infrastructure. Since the operator typically does not know about the characteristics of the load and the user does not know the details of the power distribution infrastructure, both tend to be very conservative in assuring that the load stays far below the actual circuit breaker limits. If the operator and the user are the same entity, the margin between expected load and actual power capacity can be reduced, because load and infrastructure can be matched to one another.

²In fact the National Electrical Code Article 645.5(A) [9] limits the load to 80% of the ampacity of the branch circuit.

2.1 Inefficient use of the power budget

The power budget available at a given aggregation level is often underutilized in practice, sometimes by large amounts. Some of the important contributing factors to underutilization are:

- *Staged deployment* - A facility is rarely fully populated upon initial commissioning, but tends to be sized to accommodate business demand growth. Therefore the gap between deployed and used power tends to be larger in new facilities.
- *Fragmentation* - Power usage can be left stranded simply because the addition of one more unit (a server, rack or PDU) might exceed that level's limit. For example, a 2.5kW circuit may support only four 520W servers, which would guarantee a 17% underutilization of that circuit. If a datacenter is designed such that the PDU-level peak capacity exactly matches the sum of the peak capacities of all of its circuits, such underutilization percolates up the power delivery chain and become truly wasted at the datacenter level.
- *Conservative equipment ratings* - Nameplate ratings in computing equipment datasheets often reflect the maximum rating of the power supply instead of the actual peak power draw of the specific equipment. As a result, nameplate values tend to drastically overestimate achievable power draw.
- *Variable load* - Typical server systems consume variable levels of power depending on their activity. For example, a typical low-end server system consumes less than half its actual peak power when it is idle, even in the absence of any sophisticated power management techniques. Such variability transforms the power provisioning problem into an activity prediction problem.
- *Statistical effects* - It is increasingly unlikely that large groups of systems will be at their peak activity (therefore power) levels simultaneously as the size of the group increases.

Load variation and statistical effects are the main dynamic sources of inefficiency in power deployment, and therefore we will focus on those effects for the remainder of this paper.

2.2 Other consumers of power

Our paper focuses on critical power, and therefore does not directly account for datacenter-level power conversion losses and the power used for the cooling infrastructure. However in modern, well designed facilities, both conversion losses and cooling overheads can be approximately modeled as a fixed tax over the critical power. Less modern facilities might have a relatively flat cooling power usage that does not react to changes in the heat load. In either case, the variations in the critical load will accurately capture the dynamic power effects in the facility, and with the aid of some calibration can be used to estimate the total power draw.

3. POWER ESTIMATION

One of the difficulties of studying power provisioning strategies is the lack of power usage data from large-scale deployments. In particular, most facilities lack on-line power monitoring and data collection systems that are needed for such studies. We circumvent this problem by deploying an indirect power estimation framework that is flexible, low-overhead and yet accurate in predicting power usage at moderate time intervals. In this section we describe our framework, and present some validation data supporting its accuracy. We begin by looking at the power usage profile of a typical server and how nameplate ratings relate to the actual power draw of machines.

| Component | Peak Power | Count | Total |
|----------------|------------|-------|-------|
| CPU [16] | 40 W | 2 | 80 W |
| Memory [18] | 9 W | 4 | 36 W |
| Disk [24] | 12 W | 1 | 12 W |
| PCI slots [22] | 25 W | 2 | 50 W |
| Motherboard | 25W | 1 | 25 W |
| Fan | 10 W | 1 | 10 W |
| System Total | | | 213 W |

Table 1: Component peak power breakdown for a typical server

3.1 Nameplate vs. actual peak power

A server is typically tagged with a nameplate rating that is meant to indicate the maximum power draw of that machine. The main purpose of this label is to inform the user of the power infrastructure required to safely supply power to the machine. As such, it is a conservative number that is guaranteed not to be reached. It is typically estimated by the equipment manufacturer simply by adding up the worst case power draw of all components in a fully configured system [19].

Table 1 shows the power draw breakdown for a server built out of a motherboard with 2 x86 CPUs, an IDE disk drive, 4 slots of DDR1 DRAM, and 2 PCI expansion slots. Using the maximum power draw taken from the component datasheets we arrive at a total DC draw of 213 W. Assuming a power supply efficiency of 85% we arrive at a total nameplate power of 251 W.

When we actually measure the power consumption of this server using our most power intensive benchmarks we instead only reach a maximum of 145W, which is less than 60% of the nameplate value. We refer to this measured rating as the *actual peak power*. As this example illustrates, actual peak power is a much more accurate estimate of a system's peak consumption, therefore we choose to use it instead of nameplate ratings in our subsequent analysis.

The breakdown shown in Table 1 does nevertheless reflect the power consumption breakdown in a typical server. CPUs and memory dominate total power, with disk power becoming significant only in systems with several disk drives. Miscellaneous items such as fans and the motherboard components round out the picture.

3.2 Estimating Server Power Usage

Our power model uses CPU utilization as the main signal of machine-level activity. For each family of machines with similar hardware configuration, we run a suite of benchmarks that includes some of our most representative workloads as well as a few micro-benchmarks, under variable loads. We measure total system power against CPU utilization and try to find a curve that approximates the aggregate behavior. Figure 2 shows our measurements alongside a linear model and an empirical non-linear model that more closely fits our observations. The horizontal axis shows the CPU utilization reported by the OS as an average across all CPUs (u). A calibration parameter r that minimizes the squared error is chosen (a value of 1.4 in this case). For each class of machines deployed, one set of calibration experiments is needed to produce the corresponding model; an approach similar to Mantis [10].

The error bars in Figure 2 give a visual indication that such models can be reasonably accurate in estimating total power usage of individual machines. Of greater interest to this study, however, is the accuracy of this methodology in estimating the dynamic power usage of groups of machines. Figure 3 shows how the model compares to the actual measured power drawn at the PDU level (a few

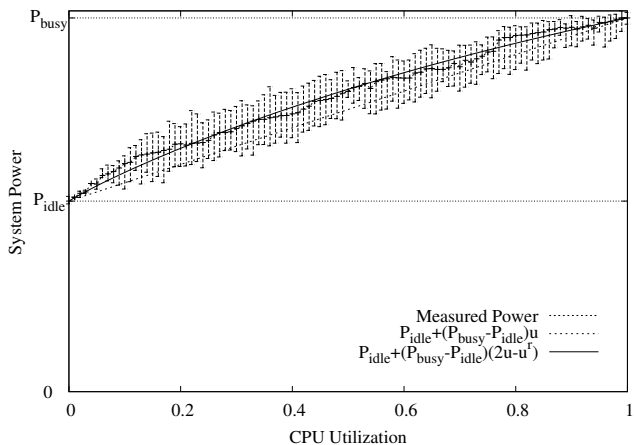


Figure 2: Model fitting at the machine level

hundred servers) in one of our production facilities. Note that except for a fixed offset, the model tracks the dynamic power usage behavior extremely well. In fact, once the offset is removed, the error stays below 1% across the usage spectrum and over a large number of PDU-level validation experiments.

The fixed offset is due to other loads connected to the PDUs that are not captured by our model, most notably network switching equipment. We have found that networking switches operate on a very narrow dynamic range³, therefore a simple inventory of such equipment, or a facility-level calibration step is sufficient for power estimation.

We were rather surprised to find that this single activity level signal (CPU utilization) produces very accurate results, especially when larger numbers of machines are considered. The observation can be explained by noting that CPU and memory are in fact the main contributors to the dynamic power, and other components either have very small dynamic range⁴ or their activity levels correlate well with CPU activity. Therefore, we found it unnecessary so far to use more complex models and additional activity signals (such as hardware performance counters).

This modeling methodology has proved very useful in informing our own power provisioning plans.

3.3 The Data Collection Infrastructure

In order to gather machine utilization information from thousands of servers, we use a distributed collection infrastructure as shown in Figure 4. At the bottom layer, collector jobs gather periodic data on CPU utilization from all our servers. The collectors write the raw data into a central data repository. In the analysis layer, different jobs combine CPU activity with the appropriate models for each machine class, derive the corresponding power estimates and store them in a data repository in time series format. Analysis programs are typically built using Google’s Mapreduce [8] framework.

4. POWER USAGE CHARACTERIZATION

Here we present a baseline characterization of the power usage of three large scale workloads and an actual whole datacenter, based on six months of power monitoring observations.

³Measurements show that Ethernet switch power consumption can vary by less than 2% across the activity spectrum.

⁴Our component measurements show that the dynamic power range is less than 30% for disks, and negligible for motherboards.

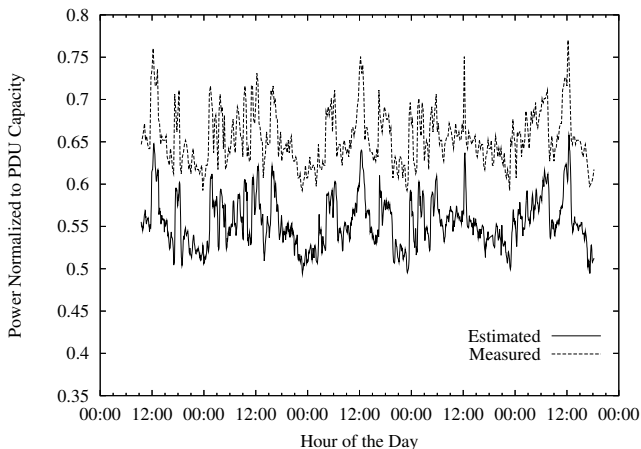


Figure 3: Modeled vs. Measured Power at the PDU Level

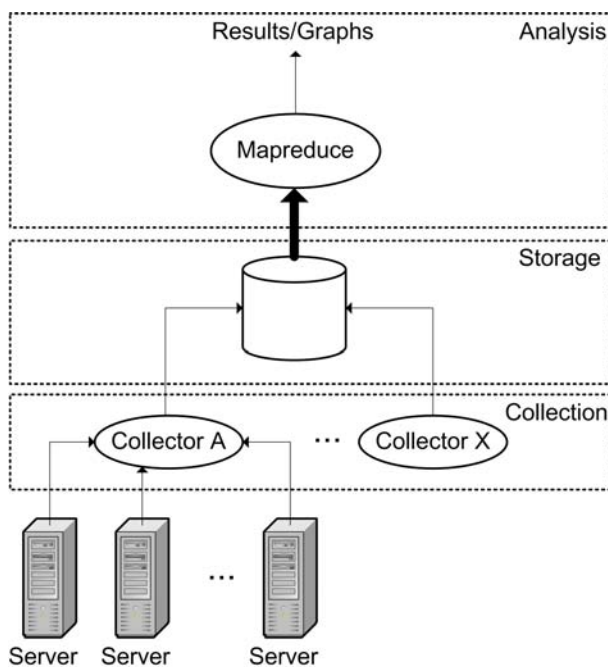


Figure 4: Collection, storage, and analysis architecture.

4.1 Workloads

We have selected three workloads that are representative of different types of large-scale services. Below we briefly describe the characteristics of these workloads that are relevant to this study:

Websearch: This represents a service with high request throughput and a very large data processing requirements for each request. We measure machines that are deployed in Google’s Web search services. Overall activity level is generally strongly correlated with time of day, given the online nature of the system.

Webmail: This represents a more disk I/O intensive Internet service. We measure servers running GMail, a web-based email product with sophisticated searching functionality. Machines in this service tend to be configured with a larger number of disk drives, and each request involves a relatively small number of servers. Like Websearch, activity level is correlated with time of day.

Mapreduce: This is a cluster that is mostly dedicated to running

large offline batch jobs, of the kind that are amenable to the mapreduce [8] style of computation. The cluster is shared by several users, and jobs typically involve processing terabytes of data, using hundreds or thousands of machines. Since this is not an online service, usage patterns are more varied and less correlated with time of day.

4.2 Datacenter setup

For the results in this section, we picked a sample of approximately five thousand servers running each of the workloads above. In each case, the sets of servers selected are running well-tuned workloads and typically at high activity levels. Therefore we believe they are representative of the more efficient datacenter-level workloads, in terms of usage of the available power budget.

The main results are shown as cumulative distribution functions (CDFs) of the time that a group of machines spends at or below a given fraction of their aggregate peak power (see for example Figure 5). For each machine, we derive the average power over 10 minute intervals using the power model described earlier. The aggregate power for each group of 40 machines during an interval makes up a rack power value, which is normalized to their actual peak (i.e., the sum of the maximum achievable peak power consumption of all machines in the group). The cumulative distribution of these rack power values is the curve labeled "Rack" in the graph. The "PDU" curve represents a similar aggregation, but now grouping sets of 20 racks (or about 800 machines). Finally, the "Cluster" curve shows the CDF for all machines (approximately 5000 machines).

4.3 CDF power results

Let's take a closer look at the power CDF for Websearch (Figure 5). The Rack CDF starts at around 0.45 of normalized power, indicating that at no time does any one rack consume less than 45% of its actual peak. This is likely close to the idle power of the machines in the rack. The curve rises steeply, with the largest fraction of the CDF (i.e. the most time) spent in the 60 - 80% range of actual peak power. The curve intercepts the top of the graph at 98% of the peak power, indicating that there are some time intervals where all 40 machines in a given rack are operating very close to their actual peak power. The right graph of Figure 5 zooms in on the upper part of the CDF, to make the intercepts with the top of the graph clearer. Looking at the PDU and Cluster curves, we see that they tend to have progressively higher minimum power and lower maximum power. The larger the group of machines is, the less likely it is that all of them are simultaneously operating near the extreme minimum or maximum of power draw. For Websearch, some racks are reaching 98% of actual peak power for some time interval, whereas the entire cluster never goes over 93%. It is striking to see that groups of many hundreds of machines (PDU-level) can spend nearly 10% of the time within 10% of their aggregate peak power.

The corresponding CDFs for Webmail are shown in Figure 6. The shape of these is similar to that of Websearch, with two notable differences: the dynamic range of the power draw is much narrower, and the maximum power draw is lower. Webmail machines tend to have more disks per machine, and disk power draw does not vary significantly with changes in activity levels. Hence a larger fraction of the power draw of these machines is fixed and the dynamic range is reduced. The max power draw is also lower. Interestingly, we see a maximum of about 92% of peak actual power at the rack level, and 86% at the cluster level; an even higher gap than Websearch.

The curves for Mapreduce (Figure 7) show a larger difference

between the Rack, PDU, and Cluster graphs than both Websearch and Webmail. This indicates that the power draw across different racks is much less uniform; likely a result of its less time-dependent activity characteristics. This behavior leads to a much more noticeable averaging effect at the cluster level. While the racks top out at very close to 100% of peak actual power, the cluster never goes above about 90%.

These results are significant for machine deployment planning. If we use the maximum power draw of individual machines to provision the datacenter, we will be stranding some capacity. For Websearch, about 7.5% more machines could be safely deployed within the same power budget. The corresponding numbers for Webmail and Mapreduce are even higher, at 16% and 11%.

The impact of diversity - Figure 8 presents the power CDF when all the machines running the three workloads are deployed in a hypothetical combined cluster. This might be representative of a datacenter-level behavior where multiple high-activity services are hosted. Note that the dynamic range of the mix is narrower than that of any individual workload, and that the highest power value achieved (85% of actual peak) is also lower than even that of the lowest individual workload (Webmail at 86%). This is caused by the fact that power consumption peaks are less correlated across workloads than within them. It is an important argument for mixing diverse workloads at a datacenter, in order to smooth out the peaks that individual workloads might present. Using the highest power of the mix to drive deployment would allow 17% more machines to be deployed to this datacenter.

An actual datacenter - So far, we have looked only at large, well-tuned workloads in a fully deployed environment. In a real datacenter there will be additional workloads that are less well-tuned, still in development, or simply not highly loaded. For example, machines can be assigned to a service that is not yet fully deployed, or might be in various stages of being repaired or upgraded, etc. Figure 9 shows the power CDF for one such datacenter. We note the same trends as seen in the workload mix, only much more pronounced. Overall dynamic range is very narrow (52 - 72%) and the highest power consumption is only 72% of actual peak power. Using this number to guide deployment would present the opportunity to host a sizable 39% more machines at this datacenter.

4.4 Value of Power Capping

One of the features that stands out in the power CDF curves presented in the previous section is that the CDF curve intercepts the 100% line at a relatively flat slope, indicating that there are few time intervals in which close to the highest power is drawn by the machines. If we could somehow remove those few intervals we might be able to further increase the number of machines hosted within a given power budget. *Power capping* techniques accomplish that by setting a value below the actual peak power and preventing that number from being exceeded through some type of control loop. There are numerous ways to implement this, but they generally consist of a power monitoring system (possibly such as ours or one based on direct power sensing) and a power throttling mechanism. Power throttling generally works best when there is a set of jobs with loose service level guarantees or low priority that can be forced to reduce consumption when the datacenter is approaching the power cap value. Power consumption can be reduced simply by descheduling tasks or by using any available component-level power management knobs, such as CPU voltage/frequency scaling.

Note that the power sensing/throttling mechanisms needed for power capping are likely needed anyway even if we do not intend to cap power, but simply want to take advantage of the power usage gaps shown in the CDF graphs. In those cases it is required to

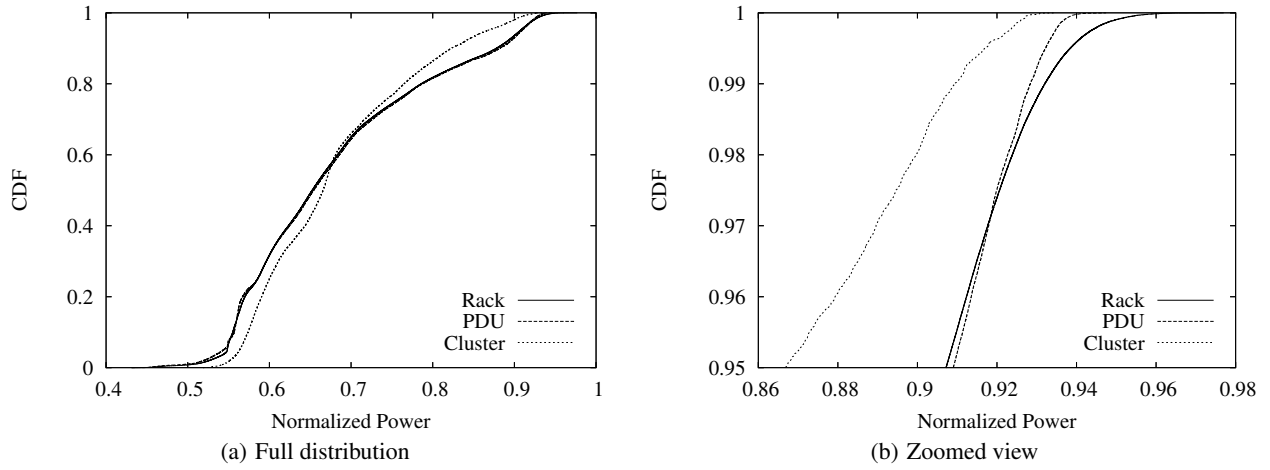


Figure 5: Websearch - CDF of power usage normalized to actual peak

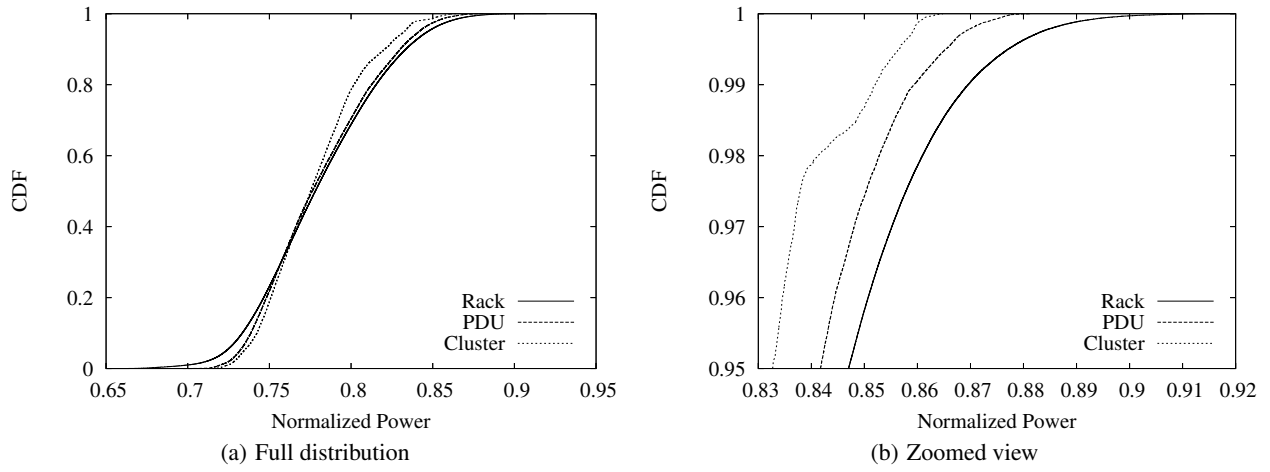


Figure 6: Webmail - CDF of power usage normalized to actual peak

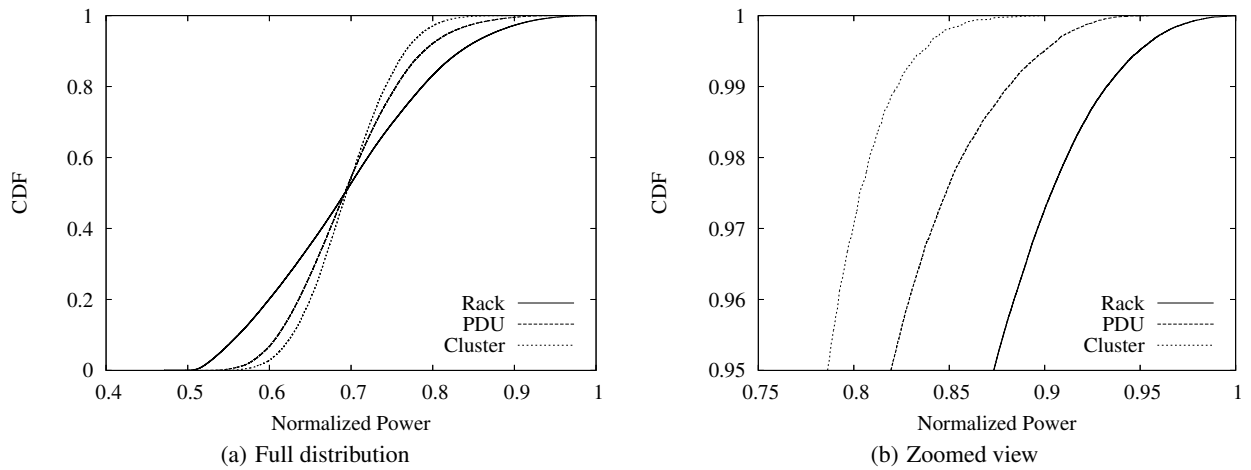


Figure 7: Mapreduce - CDF of power usage normalized to actual peak

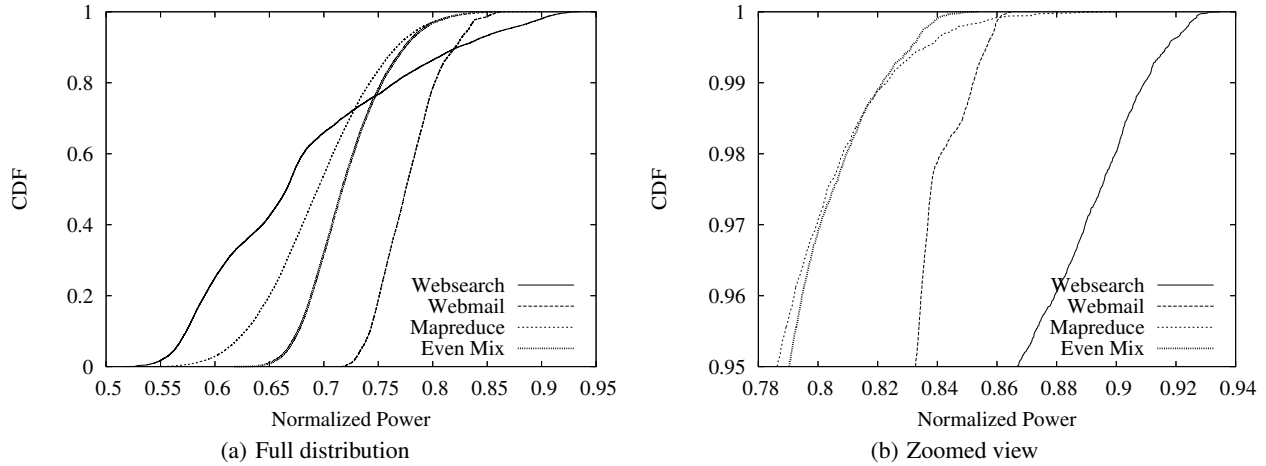


Figure 8: CDF of Websearch, Webmail, Mapreduce and the mixture of all at the cluster level

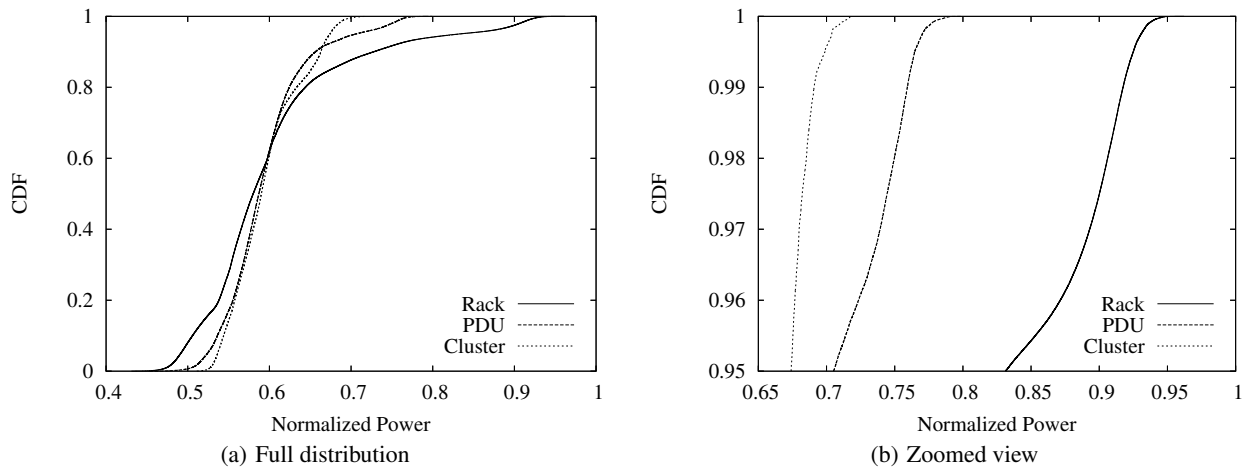


Figure 9: CDF of a Real Datacenter

insure against poorly-characterized workloads or unexpected load spikes.

Table 2 presents the gains that could be achieved with such a scheme. For each workload, we show the potential for increased machine deployment, given an allowance of 1 or 2% of time spent in power-capping mode. We also include the no power capping numbers for comparison. We have excluded Websearch and Webmail (by themselves) from power capping, because given their on-line nature they might not have much opportunity for power reduction at peak load.

Overall, the additional gains in machine deployment are noticeable but relatively modest. Generally, 1% captures most of the benefits with only little additional gains for 2% of capping time. The best case is Mapreduce, which shows an increase from 11% in potential increased machine deployment without power capping, to 24% with capping 2% of the time. Notably, mixing the workloads diminishes the relative gains, because the different workloads are already decreasing the likelihood of a simultaneous power spike in all machines.

The table also shows the number and length of power-capping intervals that would be incurred for each workload. This information gives some insight into how often the power capping system would be triggered, which in turn is useful for deciding on what

kind of mechanism to use. Fewer, longer intervals are probably more desirable, because there is always some loss upon entering and leaving the power capping interval.

Perhaps the biggest advantage dynamic power capping is that it can relax the requirement to accurately characterize workloads prior to deployment, and provide a safety valve for cases where workload behavior changes unexpectedly.

4.5 Average versus peak power

Another interesting observation that can be derived from our data is the difference between the average and observed peak power draw of a workload or mix of workloads. While peak power draw is the most important quantity for guiding the deployment of machines to a datacenter, average power is what determines the power bill. We mentioned load-dependent power variations as one of the factors leading to inefficient use of the power budget in an earlier section, we are now able to quantify it.

Table 3 shows the ratio of average power to observed peak power (over the half-year interval) for the different workloads and mixes of workloads. The ratios reflect the different dynamic ranges for the different workloads: Websearch has the highest dynamic range and lowest average to peak ratio at 73%. Mapreduce is somewhat higher, and Webmail has the highest ratio at close to 90%. The two

| Workload | Percentage of Time in Power-Capping Mode | Increase in Machine Deployment | N. of Intervals per Month | Median Interval (min) | Avg Interval (min) |
|-----------------|--|--------------------------------|---------------------------|-----------------------|--------------------|
| Websearch | 0% | 7.0% | - | - | - |
| Webmail | 0% | 15.6% | - | - | - |
| Mapreduce | 0% | 11.0% | - | - | - |
| | 1% | 21.5% | 21.0 | 10.0 | 20.5 |
| | 2% | 23.8% | 38.8 | 20.0 | 22.2 |
| Mix | 0% | 17.1% | - | - | - |
| | 1% | 21.7% | 12.2 | 20.0 | 35.3 |
| | 2% | 23.5% | 23.1 | 20.0 | 37.3 |
| Real Datacenter | 0% | 39.1% | - | - | - |
| | 1% | 44.7% | 9.0 | 20.0 | 47.9 |
| | 2% | 46.0% | 12.5 | 40.0 | 69.3 |

Table 2: Impact of Power Capping

| Workload | Average Power | Observed Peak Power | Average / Observed |
|-----------|---------------|---------------------|--------------------|
| Websearch | 68.0 % | 93.5 % | 72.7 % |
| Webmail | 77.8 % | 86.5 % | 89.9 % |
| Mapreduce | 69.6 % | 90.1 % | 77.2 % |
| Mix | 72.1 % | 85.4 % | 84.4 % |
| Real DC | 59.5 % | 71.9 % | 82.8 % |

Table 3: Average and observed peak power (normalized to actual peak) at the cluster level

mixed workloads also show higher ratios, with 84% for the mix of the three tuned workloads, and 83% for the real datacenter.

We see that a mix of diverse workloads generally reduces the difference between average and peak power, another argument in favor of this type of deployment. Note that even for this best case, on the order of 15% of the power budget remains stranded simply because of the difference between average and peak power, which further increases the relative weight of power provisioning costs over the cost of energy.

5. TWO POWER SAVINGS APPROACHES

In the previous section we used the power modeling infrastructure to analyze actual power consumption of various workloads. Here we take the same activity data from our machines over the six month time period and use the model to simulate the potential for power and energy saving of two schemes.

5.1 CPU Voltage/Frequency Scaling

CPU voltage and frequency scaling (DVS for short) is a useful technique for managing energy consumption that has recently been made available to server-class processors. Here we ask our power model to predict how much energy savings and peak power reductions could have been achieved had we used power management techniques based on DVS in the workloads analyzed in the previous section.

For simplicity and for the purpose of exploring the limits of the benefit, we use an oracle-style policy. For each machine and each data collection interval, if the CPU utilization is below a certain threshold, we simulate DVS activation by halving⁵ the CPU com-

⁵There are various CPUs in the market today that are capable of such power reductions through DVS [17, 1]

ponent of the total power, while leaving the power consumption of the remaining components unchanged.

Without detailed application characterization we cannot determine how system performance might be affected by DVS, therefore we simulate three CPU utilization thresholds for triggering DVS: 5%, 20%, 50%. We pick 5% as a conservative threshold to examine how much benefit can be achieved with almost no performance impact. We use 50% as a very aggressive threshold for the scenario where performance can be degraded significantly or the application has substantial amount of performance slack.

Figure 10 shows the impact of CPU DVS at the cluster level on our three workloads and on the real datacenter. DVS has a more significant potential impact on energy than peak power, with savings of over 20% when using the more aggressive threshold in two out of four cases. We expected this, since in periods of cluster-wide peak activity it is unlikely that many servers will be below the DVS trigger threshold. It is still surprising that there are cases where DVS can deliver a moderate but noticeable reduction in maximum observed power. This is particularly the case for the real datacenter, where the workload mix enables peak power reductions between 11-18%.

Among the three workloads, Websearch has the highest reduction in both peak power and energy. Websearch is the most compute intensive workload, therefore the CPU consumes a larger percentage of the total machine power, allowing DVS to produce larger reductions relative to total power. DVS achieves the least energy savings for Webmail, which has the narrowest dynamic power range and relatively high average energy usage. Webmail is generally deployed on machines with more disks, and therefore the CPU is a smaller contributor to the total power, resulting in a correspondingly smaller impact of DVS. Mapreduce shows the least reduction in peak power, since it also tends to use machines with more disks while achieving even higher peak power usage than Webmail. These two factors create the most difficult scenario for DVS.

It is also worth noting that due to our somewhat coarse data collection interval (10 min) the DVS upside is somewhat underestimated here. The switching time of the current DVS technology can accommodate a sub-second interval, so bigger savings might be possible using finer-grained triggers.

5.2 Improving Non-Peak Power Efficiency

Power efficiency of computing equipment is almost invariably measured when running the system under maximum load. Generally when "performance per Watt" is presented as a rating, it is

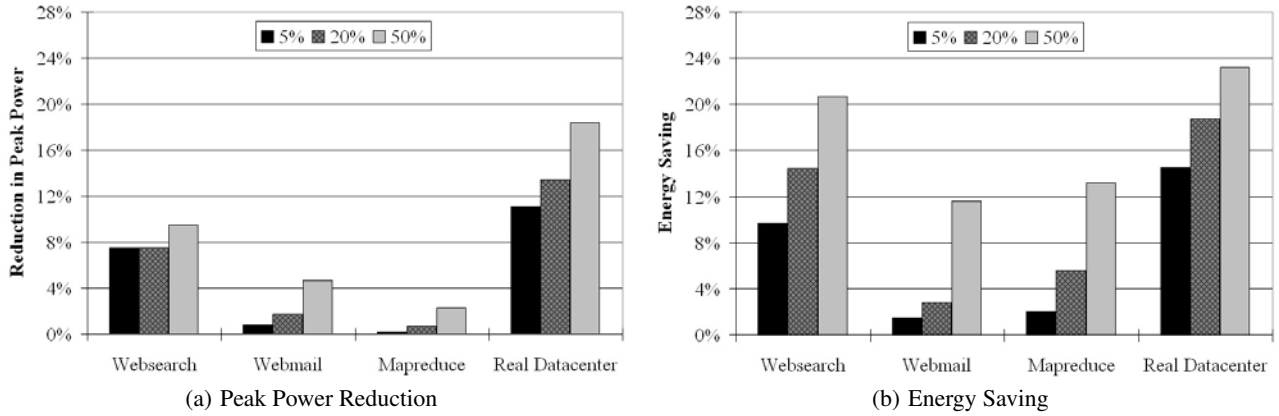


Figure 10: Impact of CPU DVS at Datacenter Level

implicitly understood that the system was exercised to maximum performance, and upon reaching that the power consumption was measured. However, as the analysis in the previous section showed, the reality is that machines operate away from peak activity a good fraction of the time. Therefore it is important to conserve power across the activity spectrum, and not just at peak activity.

Figure 11 shows the power consumption at idle (no activity) as a fraction of peak power from five of the server configurations we deploy. Idle power is significantly lower than the actual peak power, but generally never below 50% of peak. Ideally we would like our systems to consume no power when idle, and for power to increase roughly proportionally with increased activity; a behavior similar to the curves in Figure 2 but where P_{idle} is near zero. Arguably, systems with this behavior would be equally power efficient regardless of activity level. To assess the benefits of such behavioral change, we altered our model so that idle power for every machine was set to 10% of the actual peak power. All other model parameters, including actual peak power, remained the same as before.

The results, shown in Figure 12, reveal that the gains can be quite substantial. The maximum cluster-level peak power was reduced between 6-20% for our three workloads, with corresponding energy savings of 35-40%. In a real datacenter, however, the observed maximum power consumption dropped over 30%, while less than half the energy was used. The fact that such dramatic gains are possible without any changes to peak power consumption strongly suggest that system and component designers should strive to achieve such behavior in real servers.

It is important to note that the machines in our study, especially the ones running the three workloads, were rarely fully idle. Therefore, inactive power modes (such as sleep or standby modes) are unlikely to achieve the same level of savings.

6. POWER PROVISIONING STRATEGIES

From the results in the previous sections, we can draw some conclusions about strategies for maximizing the amount of compute equipment that can be deployed at a datacenter with a given power capacity.

First of all, it is important to understand the actual power draw of the machines to be deployed. Nameplate power figures are so conservative as to be useless for the deployment process. Accurate power measurements of the machines are needed in the actual configurations to be deployed and running benchmarks that maximize overall power draw.

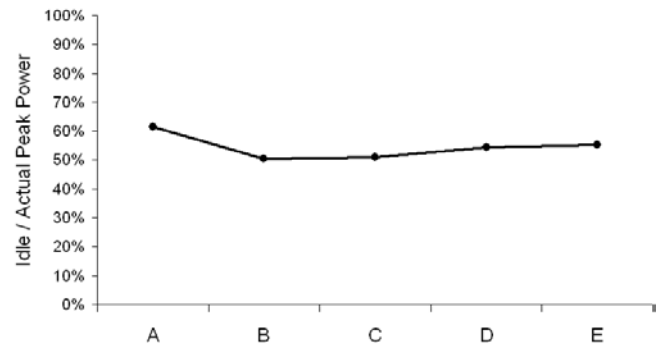


Figure 11: Idle power as fraction of peak power in 5 server configurations

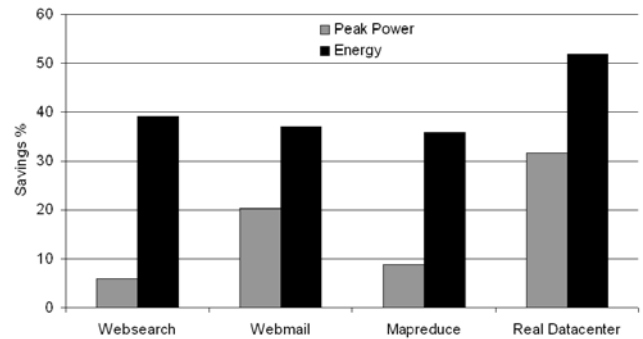


Figure 12: Power and energy savings achievable by reducing idle power consumption to 10% of peak

The characterization of application power draw at different levels of deployment granularity allows us to judge the potential for safely over-subscribing pieces of the power distribution hierarchy. Over-subscription at the rack level is not safe. In both Websearch and Mapreduce, individual racks approach very close to peak actual power during some time intervals. Webmail has a little room for over-subscription at the rack level, at 92%. At the PDU level, more potential for over-subscription exists. At the cluster level, there is

a noticeable difference between observed and actual peak power, allowing for the deployment of between 7 - 16% more machines for individual applications. The headroom increases when applications are mixed together, indicating that it is desirable to do so. Mixing also leads to a narrowing of average to peak power, which is desirable from a utilization of infrastructure standpoint. Finally, we have shown that in a real cluster the deployment of less well-tuned applications and other conditions leading to poorly-utilized machines can drive the headroom close to 40%. Once again, this is using peak actual power to guide deployment. The more common practice of using nameplate power further inflates these numbers, leading to headroom for 80 - 130% more machines to be deployed.

A dynamic power management scheme to cap the peak power draw at some pre-determined value has two advantages. First of all, it can act as a safety valve, protecting the power distribution hierarchy against overdraw. It thus allows for aggressive deployment of machines, even in the face of poorly characterized applications or unexpected load spikes. Secondly, it enables additional over-subscription of the available power. Capping power for even a small fraction of overall time can deliver noticeable additional gains in machine deployment.

While dynamic voltage/frequency scaling may not produce much reduction of peak power draw at the rack level, there is a noticeable reduction at the cluster level. Depending on application, peak power reductions of up to 10% are seen for aggressive schemes, growing up to 18% for the real datacenter workload mix. Even the least aggressive scheme netted an 11% reduction in peak power for the real datacenter mix.

7. RELATED WORK

To overcome the conservativeness of nameplate power, the industry starts to provide coarse-grained power calculators based on customized component and activity level selections [15]. Models that estimate power usage based on activity metrics have been studied by a number of researchers. Contreras and Martonosi [7] use hardware event counters to derive power estimates that are accurate at sub-second time intervals. Our approach is more similar to that of Economou et al [10], which is based on coarser activity metrics such as CPU load and I/O activity. An even coarser modeling scheme is presented by Bohrer et al [4], in their study of energy management techniques. Our results further validate the usefulness of relatively simple low-overhead power modeling techniques.

There is growing number of studies of power management techniques and power/energy-aware scheduling policies. At the single system level, Felter et al [11] study *power shifting*, a technique to reduce peak power with minimal performance impact that is based on dynamically re-allocating power to the most performance critical components. Carrera et al [5] instead focus on the disk subsystem, proposing energy management strategies that include the use of multi-speed disks and combinations of server-class and laptop drives.

At the cluster or datacenter-level, Chase et al. [6] treat energy as a resource to be scheduled by a hosting center's management infrastructure, and propose a scheme that can reduce energy usage by 29% while still meeting a specified level of service. Moore et al. [20] present a framework for measurement and analysis of datacenter-level workloads, with a measurement infrastructure that has some similarities to ours. They use a synthetic workload to evaluate their framework running on two moderate-sized clusters. Our paper is not concerned with any specific power management scheme. We believe that a wide range of techniques could be effective at the datacenter level, given the large time constants involved.

The studies of Femal and Freeh [12, 13], and of Ranganathan et

al [23] are probably the most closely related to ours. Ranganathan et al note that more efficient power management solutions can be reached by managing power at the rack (or *ensemble*) level than at individual blades. While we agree with that assertion, our results seem to contradict their observation that synchronized power usage spikes never happen in practice. Our data indicates that such spikes at the rack level do happen, suggesting that the kind of power management solutions they proposed might be more appropriate for much larger groups of machines. We speculate that the much larger scale of our workloads, and how well they are tuned, are partly responsible for this discrepancy in observed behavior. Femal and Freeh [12, 13] deal directly with the issue of power over-subscription in small clusters (tens of servers), and propose a dynamic control scheme based on dynamic CPU DVS to reduce peak consumption. Although our estimates appear to be more modest than theirs, we agree that CPU DVS can have an impact on cluster-level peak power savings.

Finally, some researchers focus on the cooling infrastructure, and temperature management [21, 14]. As previously stated, our paper does not deal with the energy or power used for the cooling infrastructure. These are important research areas that are complementary to our work.

8. CONCLUSIONS

Some of the most interesting computing systems being built today look more like a warehouse than a refrigerator. Power provisioning decisions for such systems can have a dramatic economic impact as the cost of building large datacenters could surpass the cost of energy for the lifetime of the facility. Since new datacenter construction can take tens of months, intelligent power provisioning also has a large strategic impact as it may allow an existing facility to accommodate the business growth within a given power budget.

In this paper we study how power usage varies over time, and as the number of machines increases from individual racks to clusters of up to five thousand servers. By using multiple production workloads, we are also able to quantify how power usage patterns are affected by workload choice. The understanding of power usage dynamics can inform the choice of power management and provisioning policies, as well as quantify the potential impact of power and energy reduction opportunities. To our knowledge this is the first power usage study at the scale of datacenter workloads, and the first reported use of model-based power monitoring techniques for power provisioning in real production systems.

We echo commonly held beliefs that nameplate ratings are of little use in power provisioning as they tend to grossly overestimate actual maximum usage. Using a more realistic peak power definition, we were able to quantify the gaps between maximum achieved and maximum theoretical power consumption of groups of machines. These gaps would allow hosting between 7% and 16% more computing equipment for individual (well-tuned) applications, and as much as 39% in a real datacenter running a mix of applications, through careful over-subscription of the datacenter power budget. We find that power capping mechanisms can enable us to capitalize on those opportunities by acting as a safety net against the risks of over-subscription, and are themselves able to provide additional albeit modest power savings. We note, however, that over-subscribing power at the rack level is quite risky, given that large Internet services are capable of driving hundreds of servers to high-activity levels simultaneously. The more easily exploitable over-subscription opportunities lie at the facility level (thousands of servers).

We also find that CPU dynamic voltage/frequency scaling might

yield moderate energy savings (up to 23%). Although it has a more limited peak power savings potential, it is still surprising that a technique usually dismissed for peak power management can have a noticeable impact at the datacenter level.

Finally, we argue that component and system designers should consider power efficiency not simply at peak performance levels but across the activity range, as even machines used in well tuned large scale workloads will spend a significant fraction of their operational lives below peak activity levels. We show that peak power consumption at the datacenter level could be reduced by up to 30% and energy usage could be halved if systems were designed so that lower activity levels meant correspondingly lower power usage profiles.

9. ACKNOWLEDGMENTS

This study is based on the computing infrastructure built by our Google colleagues in the Platforms Engineering and Hardware Operations teams. We are specially grateful to Selver Corhodzic, Bill Hambrun, Urs Hölzle, Justin Moore, Rob Pike, Eduardo Pinheiro, Sean Quinlan, Partha Ranganathan, and the anonymous reviewers for their feedback on earlier drafts. The support of Walt Drummond and Urs Hölzle made this study possible.

10. REFERENCES

- [1] Advanced Micro Devices, Inc. AMD Opteron Processor with Direct Connect Architecture. http://enterprise.amd.com/downloads/2P_Power_PID_41497.pdf, September 2006.
- [2] L. A. Barroso. The price of performance: An economic case for chip multiprocessing. *ACM Queue*, 3(7), September 2005.
- [3] L. A. Barroso, J. Dean, and U. Hölzle. Web search for a planet: The google cluster architecture. *IEEE Micro*, 23(2):22–28, March/April 2003.
- [4] P. Bohrer, E. Elnozahy, T. Keller, M. Kistler, C. Lefurgy, C. McDowell, and R. Rajamony. The case for power management in web servers. *Power Aware Computing*, 2002.
- [5] E. V. Carrera, E. Pinheiro, and R. Bianchini. Conserving disk energy in network servers. In *ICS '03: Proceedings of the 17th annual international conference on Supercomputing*, pages 86–97, 2003.
- [6] J. S. Chase, D. C. Anderson, P. N. Thakar, A. M. Vahdat, and R. P. Doyle. Managing energy and server resources in hosting centers. In *SOSP '01: Proceedings of the eighteenth ACM symposium on Operating systems principles*, pages 103–116, 2001.
- [7] G. Contreras and M. Martonosi. Power prediction for intel xscale processors using performance monitoring unit events. In *ISLPED '05: Proceedings of the 2005 international symposium on Low power electronics and design*, pages 221–226, 2005.
- [8] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. In *Proceedings of the 6th USENIX Symposium on Operating Systems Design and Implementation (OSDI'04)*, pages 137 – 150, December 2004.
- [9] M. Early, J. Sargent, J. Sheehan, and J. Caloggero. *NEC 2005 Handbook*. National Fire Protection Association, Quincy, Massachusetts, 2005.
- [10] D. Economou, S. Rivoire, C. Kozyrakis, and P. Ranganathan. Full-system power analysis and modeling for server environments. In *Workshop on Modeling Benchmarking and Simulation (MOBS)*, June 2006.
- [11] W. Felter, K. Rajamani, T. Keller, and C. Rusu. A performance-conserving approach for reducing peak power consumption in server systems. In *ICS '05: Proceedings of the 19th annual international conference on Supercomputing*, pages 293–302, 2005.
- [12] M. E. Femal and V. W. Freeh. Safe overprovisioning: Using power limits to increase aggregate throughput. In *4th International Workshop on Power Aware Computer Systems (PACS 2004)*, pages 150 – 164, December 2004.
- [13] M. E. Femal and V. W. Freeh. Boosting data center performance through non-uniform power allocation. In *Second International Conference on Autonomic Computing (ICAC)*, pages 250 – 262, June 2005.
- [14] T. Heath, A. P. Centeno, P. George, L. Ramos, Y. Jaluria, and R. Bianchini. Mercury and freon: Temperature emulation and management for server systems. In *International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, October 2006.
- [15] Hewlett-Packard Development Company, L.P. HP Enterprise Configurator. <http://h30099.www3.hp.com/configurator/powercalcs.asp>, 2006.
- [16] Intel Corporation. Intel Xeon Processor with 512KB L2 Cache at 1.80 GHz to 3 GHz Datasheet. <http://download.intel.com/design/Xeon/datashts/29864206.pdf>, March 2003.
- [17] Intel Corporation. Dual-Core Intel Xeon Processor LV and ULV Datasheet. <http://download.intel.com/design/intarch/datashts/31139101.pdf>, September 2006.
- [18] Micron Technology, Inc. Calculating Memory System Power for DDR. <http://download.micron.com/pdf/technotes/ddr/TN4603.pdf>, 2001.
- [19] J. Mitchell-Jackson, J. G. Koomey, B. Nordman, and M. Blazek. Data center power requirements: measurements from silicon valley. *Energy (Energy) ISSN 0360-5442*, 28(4):837 – 850, 2003.
- [20] J. Moore, J. Chase, K. Farkas, and P. Ranganathan. Data center workload monitoring, analysis, and emulation. In *Eighth Workshop on Computer Architecture Evaluation using Commercial Workloads*, February 2005.
- [21] J. Moore, J. Chase, P. Ranganathan, and R. Sharma. Making scheduling "cool": Temperature-aware workload placement in data centers. In *USENIX Annual Technical Conference*, pages 61 – 75, April 2005.
- [22] PCI-SIG. Conventional PCI 2.3 - An Evolution of the Conventional PCI Local Bus Specification. http://www.pcisig.com/members/downloads/specifications/conventional/conventional_pci_2_3.pdf.
- [23] P. Ranganathan, P. Leech, D. Irwin, and J. Chase. Ensemble-level power management for dense blade servers. In *ISCA '06: Proceedings of the 33rd annual international symposium on Computer Architecture*, pages 66–77, 2006.
- [24] Seagate Technology LLC. Product manual Barracuda 7200.7. <http://www.seagate.com/support/disc/manuals/ata/cuda7200pm.pdf>, September 2005.
- [25] W. P. Turner IV, J. H. Seader, and K. G. Brill. Tier classifications define site infrastructure performance. *The Uptime Institute, White Paper*, 2006.