How Much Power Oversubscription is Safe and Allowed in Data Centers?

Xing Fu, Xiaorui Wang
University of Tennessee, Knoxville, TN 37996
The Ohio State University, Columbus, OH 43210
{xfu1, xwang}@eecs.utk.edu

Charles Lefurgy
Power-Aware Systems Group
IBM Research, Austin, TX 78758
lefurgy@us.ibm.com

ABSTRACT

Data centers attempt to maximize return on investment by achieving high levels of utilization. This means deploying the maximum number of servers possible within existing power supply capabilities. Therefore, a key problem is determining how many servers can be safely accommodated. Recently, a variety of power capping solutions have been proposed to safely allow oversubscription of available power, but they conservatively assume that peak power should never exceed the rated power distribution equipment capacity. Hence, the open question is: how much power oversubscription is indeed safe? In this paper, we focus on data center branch circuits and systematically study the tripping characteristics of their circuit breakers (CBs). Our results on a physical testbed show that instantaneous violations of the rated CB power limit are not necessarily fatal because CBs are designed to sustain a certain amount of power overload. Whether a CB trips or not depends primarily on the transient behaviors of a power overload, such as the magnitude and duration time, as well as ambient temperature. We propose two adaptive power control strategies that utilize the CB tripping characteristics to aggressively optimize the system performance without causing the CB to trip. Our extensive hardware results with SPEC CPU2006, SPECJBB, and LINPACK benchmarks show that the proposed CB-aware power control solutions achieve 38%, 75%, and 68% better average performance, respectively, than a state-of-the-art baseline. A key contribution of our work is to provide a practical upper bound of the server power oversubscription allowed on branch circuits. As a result, our solutions allow a data center to host three times more servers than traditional static power provisioning schemes and 54% more servers than the current power capping practice.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: Design studies; C.5.5 [**Computer System Implementation**]: Servers

General Terms

Performance, Design, Experimentation

Keywords

Power capping, circuit breaker, power oversubscription, power provisioning, data center.

ICAC'11, June 14–18, 2011, Karlsruhe, Germany.

1. INTRODUCTION

In recent years, server power consumption has become a first-order concern for modern enterprise data centers. In order to amortize the non-recurring investments in the power supply facility of a data center, it is preferable to operate the facility as close as possible to its maximum capacity [11]. An additional pressure on facility operators is that upgrades in power delivery systems are extremely expensive and often lag behind required increases in hosted servers to support new business. Both of these reasons result in pressure to load as many servers as possible on the branch circuits that supply power to computer racks.

Traditionally, branch circuits are provisioned conservatively based on server nameplate power ratings, which results in significant waste of the branch circuit's power supply capacity. Recently, a promising solution is to oversubscribe the branch circuit. This involves placing more servers on it than it can support if all the servers use their maximum power consumption at the same time. To prevent overload of the circuit, *power capping* has been proposed to limit the aggregate server power to the branch circuit capacity. This provides better performance when power demand is below the branch circuit capacity and prevents undesired shutdowns by slowing down servers occasionally when the power demand is over the branch circuit capacity. Server manufacturers have responded by providing power capping as a standard feature to limit the power draw to a user-defined limit (power cap) [14][3]. Additionally, a variety of research studies have proposed power capping solutions at different levels, such as a single server [17], a rack enclosure [22][27], a data center [21][28], and a chip microprocessor [15][19].

An important issue for all power capping solutions is to select an appropriate power cap. In order to maximize the number of hosted servers in a data center, a common practice is to set the server power cap as the rated current limit of the branch circuit divided by the number of servers [14][13][11][22]. The main rationale of this practice is that *peak* power should never exceed the branch circuit capacity, otherwise the branch circuit's circuit breaker (CB) might trip and cause undesired server shutdowns, or even power outages. If the peak power becomes higher than the cap at runtime due to workload increases, immediate actions (such as processor throttling) are taken to maintain the power below the cap as soon as possible. Some studies even suggest having a safety margin below the cap to avoid any instantaneous power overloads [27].

In this paper, we argue that this common practice is too conservative, even though power capping is already a step ahead of traditional power provisioning based on nameplate power ratings. This conservativeness can result in an unnecessarily low system performance because even a small, short-lived power overload causes servers to slow down in spite of the fact that the circuit breakers will not trip. If harmless power overloads could be tolerated by power capping, then we can have higher performance, as well as

more hosted servers with the same circuit capacity. Therefore, we propose to address the open question: *how much power oversubscription is indeed safe and allowed by circuit breakers?* In other words, *what is the upper bound of power oversubscription*? To answer this question, we systematically study the tripping characteristics of a typical CB used in data centers. Our results on a physical testbed show that instantaneous violations of the rated CB power limit are not necessarily fatal because CBs are designed to sustain a certain amount of power overload. Whether a CB trips or not depends primarily on the transient behaviors of a power overload, such as the magnitude and time duration. The time interval for a CB to sustain a power overload is determined by the magnitude of the overload and normally, a higher magnitude leads to a shorter interval. Generally, a CB will trip only when the duration of an overload is longer than the allowed time interval. The allowed interval is also affected by the ambient temperature. Therefore, as long as the server power consumption is controlled to stay lower than the current overload magnitude within the allowed time interval, the CB is designed not to trip.

Based on those observations, we propose an adaptive power control strategy that utilizes the tripping characteristics of the equipped CB to aggressively optimize the system's performance without causing the CB to trip. When a power overload occurs, our solution first checks the magnitude and then the allowed time interval. Based on those transient behaviors, the proposed solution adaptively changes the parameters used in the power controller so that the settling time for power to return back to the cap is just marginally shorter than the allowed interval. When the magnitude is reduced and the allowed time interval increases, our solution continues adjusting the control parameters to fully utilize the increased interval for optimized system performance. The power controller is designed based on an advanced adaptive control theory for parameter tuning and to adapt to variations in ambient temperature. To explore the upper bound of power oversubscription, we also propose a proactive control solution, which uses available power margin in the power delivery system for a short-term turbo boost by running the server workload at a higher frequency. The proactive solution then runs the adaptive controller to safely lower the power consumption.

At least two important benefits can be achieved with the safe power oversubscription feature of our solutions. First, when servers are running under stringent power constraints (which can be expected in future data centers), our solutions can lead to higher workload performance. Second,our solutions allow a data center to host more servers than the current power capping practice. Specifically, this paper makes several major contributions:

- We present a systematic study to investigate the tripping characteristics of a typical CB used in many data centers. While previous solutions simply assume that power can never exceed the CB's capacity, to the best of our knowledge, our work is the first that utilizes transient CB tripping behaviors to optimize server performance or host additional servers. We also consider the impacts of ambient temperature on transient CB behaviors.

- In contrast to most existing power capping solutions that rely on simplistic heuristics, we use control theory to design an adaptive power controller that precisely controls the transient response of power overload to follow the designed CB trip curve. We also propose a proactive control solution to explore the practical upper bound of power oversubscription.

- Our extensive hardware results with the SPEC CPU2006, SPECJBB, and LINPACK benchmarks show that the proposed CB-aware power control solutions achieve 38% better performance, on average, than a state-of-the-art baseline that simply uses the CB capacity as the power cap without considering the CB's tripping characteristics.

- We conduct analyses to show that our adaptive power capping solutions allow a server rack to host three times more servers than traditional static power provisioning schemes and 54% more servers than the current power capping practice widely used in the industry.

The remainder of the paper is organized as follows. Section 2 provides the background about circuit breakers. Section 3 describes the design of our CB-aware power controllers. Section 4 discusses the potential applications of the proposed solutions. Section 5 provides the implementation details of the control solutions. Section 6 presents the results of our empirical experiments conducted on a physical testbed. Section 7 analyzes the impacts of our solutions on data center power provisioning. Section 8 discusses the related work. Section 9 concludes the paper.

2. BACKGROUND ON CIRCUIT BREAKER

In a data center, groups of servers (racks) are powered from branch circuits. The branch circuits connect back to a panel box that receives power from a Power Distribution Unit (PDU). Inside the panel box there is a circuit breaker for each branch circuit. The National Electric Code (NEC) [6], used in the United States, limits the long-term power load on a circuit breaker to be 80% of the circuit breaker rating. This 80% power load represents the cap of the current power capping controllers used in industry. Therefore, a data center can, at least, safely oversubscribe the circuit breaker by 25% without causing the CB to trip, according to the above NEC rule. This can be a signficant benefit for data centers. For example, the Environmental Protection Agency (EPA) estimated that the data center power consumption has an annual increase of 9% [26]. In that case, the 25% increase in power oversubscription from power capping with an increased power cap would allow new data center construction costs, ranging in the hundreds of millions USD, to be deferred for approximately three years. However, the power cap cannot be simply raised in that way because if the power draw is not well controlled, unexpected workload variations may lead to power spikes that could trip the CB. On the other hand, if we can properly control the power draw based on the tripping characteristics of the CB, a data center can even further oversubscribe the CB without causing undesired shutdowns. Such controlled power oversubscription is an efficient way for a data center to host additional servers without significantly upgrading the power supply infrastructure. Therefore, safe power oversubscription is practical, low-risk, and financially attractive for data centers.

Generally, the majority of circuit breakers have two types of trip time behaviors which are specified in the UL489 standard. First, short-circuits (for example, over 500% of the rated load) cause the CB to trip within a few milliseconds. Second, overload conditions for a less severe current draw can trip the circuit breaker on a time scale from milliseconds to hours or even weeks, depending on the severity of the overload. Only the overload condition is relevant in this paper since practical uses of power oversubscription do not reach load levels sufficient to cause a short-circuit trip condition. Also, note that other devices in the power infrastructure, such as transformers and Uninterruptible Power Supplies (UPS), are also designed to tolerate overloads since fluctuations are common in power systems. Therefore, as long as the CB does not trip, power oversubscription should be safe for data centers. We discuss the impacts of power oversubscription on other devices in Section 4. In

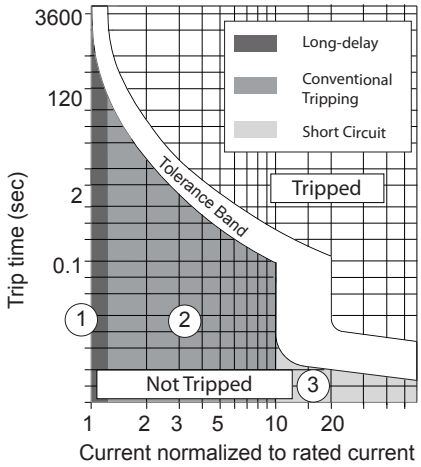


Figure 1: The trip curve of a typical circuit breaker.

an overload condition (*i.e.*, an oversubscription beyond 25% above the NEC rating), the overload must be resolved before the trip time in the CB specification. For example, circuit breakers based on UL489 available from the Rockwell Automation exhibit trip times of more than 2 minutes when overloaded to 125% of the rated load (oversubscription of 56%).

Table 1: Testbed circuit breaker at $40°C$

Current(A)	Measured trip time (sec)
slightly less than 1.35	> 7200
1.42	193
1.55	80
1.67	56
6.8	≈2
≈10	<1

Figure 1 shows the trip curve of the Rockwell Allen-Bradley 1489-A Industrial CB used in our experiments (at a temperature of $40°C$) [7]. Rockwell CBs are used in many data centers. Their trip curves follow the UL489 standard and are similar to Figure 1. This selected CB has a rated current of $I_n = 1A$. As shown in Figure 1, the trip curve of the CB is actually a band called the *tolerance band*. The area above the band is the *tripped* area, which means that the CB will trip if the duration of the CB current is longer than the specified trip time. The area below the band is the *not-tripped* area. This band represents the area where it is non-deterministic if the CB will trip. The lower and upper limits of the band are specified by the UL489 standard. The actual implementation is determined by the manufacturer [16]. The CB has three types of trip time behaviors that are shown as different regions below the tolerance band [7]. Region 1 is the long-delay tripping zone with the overload current as ($1I_n \leq I \leq 1.35I_n$). In this zone, the CB trip time is minutes to hours to even days. Region 2 is the conventional tripping zone ($1.35I_n < I \leq 10I_n$). Region 3 is the instantaneous tripping zone ($I > 10I_n$) that is designed to handle short-circuits. Table 1 shows that the actual measurements of CB trip time on our testbed for all three regions are consistent with the trip curve shown in Figure 1. While previous power capping solutions conservatively treat all the regions as the instantaneous tripping zone, a key contribution of our paper is to have different strategies for different regions. As a result, we can fully utilize the long-delay tripping zone to safely boost server performance and host additional servers.

In order to fully utilize the long-delay tripping zone without tripping the CB, one has to ensure that the overload current is reduced to I_n before the trip time specified by the lower bound of the tolerance band. Based on this observation, we choose to design a power controller based on feedback control theory because recent studies (*e.g.*, [17, 23, 27]) show that control theory can provide quantitative analyses and guarantees for system stability and settling time (*i.e.*, the time for the overload current to return to I_n). A key difference between our work and existing studies is the systematic analysis of the controller settling time. As shown in Figure 1, the allowed settling time increases whenever the overload current is reduced to a lower value. Therefore, to fully utilize the long trip time that continues to increase at each step, we propose to adopt adaptive control theory that can adjust the controller parameters based on the varying requirements of the settling time. Unlike previous power capping solutions that rely on a static power budget (*e.g.*, $0.8I_n$), our adaptive controller features a *dynamic power budget* that varies in every control period based on the overload current and its corresponding trip time. Ideally, the dynamic power budget can equal the lower bound of the tolerance band, which can be regarded as the *ideal upper bound* of safe power oversubscription. In other words, power oversubscription is safe as long as it is lower than the lower bound of the tolerance band. A major contribution of our work is that we identify this ideal upper bound and develop adaptive control solutions to explore a practical upper bound of safe power oversubscription. Note that the tripping behavior of the CB is also impacted by the ambient temperature. The relationship between the overload current and temperature can be modeled and handled in the proposed adaptive control framework, as discussed in detail in Section 3.2.

3. CB-AWARE ADAPTIVE POWER CONTROL

In this section, we first present the design and analysis of the proposed CB-Adaptive control solution. We then introduce a method to calibrate CB-Adaptive according to temperature fluctuations. Finally, we describe CB-Proactive to further improve performance.

3.1 CB-Adaptive Control

CB-Adaptive is more than just a standalone controller. It is a control methodology that adapts the parameters of existing power controllers to engineer their settling times according to the trip curves of circuit breakers. CB-Adaptive can be applied to controllers at different levels (*e.g.*, server, rack, and data center) and to different control techniques (*e.g.*, proportional-integral-derivative (PID), model predictive control (MPC)), though the detailed steps to tune parameters can be different. In this paper, as an example, we choose a state-of-the-art server-level power controller [17] as a baseline to demonstrate the design of CB-Adaptive.

As introduced in [17], the controlled variable of the server-level power controller is the power consumption of the server in the kth control period, i.e., $p(k)$. The manipulated variable is the level of the CPU Dynamic Voltage and Frequency Scaling (DVFS), i.e., CPU frequency $f(k)$. $d(k)$ is the difference between $f(k+1)$ and $f(k)$. Specifically $d(k) = f(k+1) - f(k)$. The power model used in [17] is:

$$p(k+1) = p(k) + Ad(k) \tag{1}$$

where A is a parameter determined by specific server configurations and the benchmark running on the server. Based on the power model, the controller designed in [17] in the Z-domain form is:

$$C(z) = \frac{1}{A} \tag{2}$$

In contrast to the original power controller which simply uses the rated current of the circuit breaker as its power budget, the design goal of CB-Adaptive is to enforce a dynamic power budget that

varies in every control period, based on the breaker trip curve, to guarantee that the circuit breaker does not trip if workloads vary. As a result, the server can run at its maximum performance level.

We design CB-Adaptive by adapting the controller parameter A in every control period according to the trip curve of the circuit breaker. Since the trip time is a non-linear function of the magnitude of the power overload, to reduce complexity, we use piecewise linear equations to approximate the trip curve. We then derive the desired settling time to equal the current CB trip time. The settling time is commonly defined as the time interval for power to be controlled within the set point with less than 2% errors. Finally, we derive the controller parameter A^* as a function of the desired settling time. The Z-domain form of our adaptive controller is:

$$C(z) = \frac{1}{A^*} \tag{3}$$

where

$$A^* = \frac{A}{1 - \sqrt[k]{0.02}} \tag{4}$$

where $k = \left\lfloor \frac{settling\ \ time(\mathrm{sec})}{\mathrm{T(sec)}} \right\rfloor > 0$. $settling\ time$ is set to the trip time of the circuit breaker when power is $p(k)$.

Example. Suppose $A = 76$ for a specific configuration of a server running LINPACK. In one control period, the measured current is $1.53A$. Since the current is greater than the rated current of our CB (1A), based on Figure 1, the trip time is about 80 seconds. The settling time of the proportional controller (2) is set to the trip time by adapting the control parameter according to (4). Thus $A^* = 350.38$. In the next control period, the measured current may be reduced to $1.42A$ due to DVFS throttling, the trip time becomes 190 seconds. Since the allowed settling time is now longer than before, we set $A^* = 776.89$. The key feature of CB-Adaptive is continuously adjusting the control parameter to fully utilize the allowed interval for optimized system performance.

We now consider the impacts of workload variations on the design of CB-Adaptive. In production data centers, the workload of a server may differ from the benchmark based on which we design the controller (3). To prevent the CB from tripping when the workload varies at runtime, it is necessary to analyze the impact of the different workloads on the controller using control theory. Our analysis shows that the controller parameter A in (2) needs to be changed by adding a *safety margin*. We outline the main steps of the analysis as follows.

1. We test a wide range of workloads to determine the range of the parameter A in (2) for the typical workload, such as SPECJBB, SPECCPU 2-cores in addition to LINPACK, by conducting system identification. The dynamic model of the real system is in the following format

$$p(k+1) = p(k) + gAd(k) \tag{5}$$

where the *system gain* g is used to model the variation between the real system model (5) and the nominal model (1). For example, $g = 1.5$ means that the actual change to the power consumption of the server is 1.5 times the estimated change in the event of DVFS.

2. Based on the real model (5) that models workload variations, we derive the controller parameter of CB-Adaptive as:

$$A^*_{real} = \frac{gA}{1 - \sqrt[k]{0.02}}. \tag{6}$$

The new transfer function of the adaptive controller is

$$\mathrm{C}_{real}(z) = g\mathrm{C(z)}. \tag{7}$$

3. The key difference between (6) and (4) is g. Based on step 2, we set the safety margin as $\max\{g\}$ for the various workloads we will run on the servers. As long as we run the workloads corresponding to the range of g, the safety margin guarantees that the settling time of the adaptive controller (8) is shorter than or equal to the trip time in spite of the workload variations and the circuit breaker will not trip. In case the running workload is not corresponded to the range of g, the DVFS level is decreased quickly to prevent the circuit breaker from tripping when the power consumption of the server is higher than the power budget.

$$\mathrm{C}_{real}(z) = \max\{g\}\,\mathrm{C(z)}. \tag{8}$$

In addition to the settling time, we also need to check whether the adaptive controller is stable. The derived stability range is $0 < g < 2$, which is much wider than the variation range of g observed in our extensive experiments with various workloads. Therefore, for typical workloads such as SPECJBB, SPEC CPU2006 and LINPACK, CB-Adaptive is guaranteed to be stable.

3.2 Temperature-aware CB-Adaptive

In Section 3.1, we assume that circuit breakers operate at their normal temperature ($40^\circ C$). However, in a production data center, servers and circuit breakers may run at different temperatures since the temperature distribution is not uniform. A typical raised-floor data center is divided into hot aisles and cold aisles to improve the data center cooling efficiency. Poor air recirculation at the ends of rows and top of racks often causes server inlet temperatures to vary widely (from $15^\circ C$ to $45^\circ C$) [8]. Since airflow in a data center is not ideal and the CB trip time depends on temperature, we calibrate the adaptive controller parameter as follows

$$A^*(T_{CB}) = \frac{A}{1 - \sqrt[k(T_{CB})]{0.02}} \tag{9}$$

To determine $k(T_{CB})$, we first calculate the rated current impacted by temperature. According to the CB manual [7], the relationship between the rated current and the ambient temperature is (10). We then calculate the normalized current with respect to the rated current specified for the measured temperature. Based on the piece-wise equations which approximate the trip curve, we calculate the trip time under the measured temperature as:

$$I_n^T = (C_1 T_{CB} + C_2)\, I_n \tag{10}$$

where I_n is the rated current at the nominal temperature (normally $40^\circ C$). C_1 and C_2 are constants and specified in CB data sheets. I_n^T is the rated current adjusted by temperature at $T^\circ C$. T_{CB} is the ambient temperature of the circuit breaker.

We assume that CB temperatures can be measured in real time. This is reasonable since deployments of sensor networks in data centers are already used in practice [4]. Section 5 discusses CB temperature monitoring in detail.

Example. For the circuit breaker we used in the experiments, $I_n = 1$. Suppose the ambient temperature of the CB is $10^\circ C$. According to the CB manual [7], $C_1 = -0.004167$ and $C_2 = 1.167$. Using (10), $I_n^T = 1.125$. When calculating $k(T_{CB})$ based on Figure 1, the measured current should be normalized with respect to I_n^T instead of I_n.

3.3 CB-Proactive Control

In this subsection, we provide a proactive control solution called *CB-Proactive* which further improves the performance of the server as compared to CB-Adaptive.

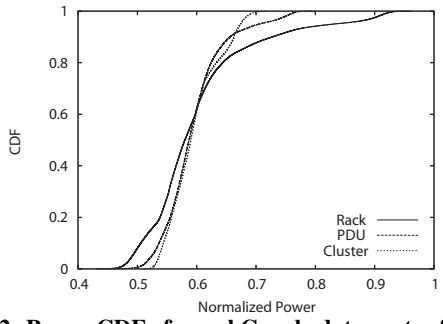


Figure 2: Power CDF of a real Google data center from [11].

The main design issue for CB-Proactive is when to increase the frequency and to what extent. As long as the circuit breaker does not trip, we want to increase the CPU DVFS level as frequently and aggressively as possible. It is necessary to consider the impact of a proactive DVFS level increase on CB-Adaptive to design CB-Proactive. When the frequency is changed proactively, the control analysis presented in Section 3.1 will not hold because the calculated frequency is overridden by the proactive frequency increase. It is not desirable to increase the frequency to increase the power consumption far beyond the power budget because, according to the trip curve, the high power consumption implies a short trip time and settling time. If the calculated frequency calculated by CB-Adaptive is overridden when the CB runs in the conventional tripping zone, the circuit breaker may trip. Based on our analysis, we design a heuristic solution which works as follows: (1) it detects whether the power consumption is lower than the lower limit of the conventional tripping zone of the circuit breaker (i.e., in the long-delay zone), and (2) if yes, it increases the DVFS level to the highest level.

4. DISCUSSION

The proposed CB-aware power control solutions can have many potential applications in data center power management. In this section, we discuss hosting additional servers in a data center. First, we present our method based on proposed CB-aware power control solutions. Then, we investigate whether it is safe to apply our method in a data center.

The allowed number of servers hosted within a rack is determined by the power consumption profile of the servers. Currently, the measured peak power consumption of the servers is equal or less than the 80% of rated power capacity of the circuit breaker according to the NEC requirement. In contrast, CB-Adaptive allows hosting additional servers by configuring the measured peak power consumption of the servers beyond 80% of the rated power capacity of the circuit breaker. As shown in Figure 2, Fan et al. [11] present the cumulative distribution functions (CDF) of the power consumption of servers running a wide range of data center workloads in a real Google data center. The highest power consumption of servers during most time is much lower than the measured peak power. Suppose we configure the allowed peak power consumption of servers as the 135% of rated power capacity of the circuit breaker. According to the CDF, during most time, the power consumption of the servers may be below 80% of rated power capacity of the circuit breaker. The time interval that violates the 80% of rated power capacity of the circuit breaker is only on a scale of minutes. Those short-term violations are allowed by NEC. Furthermore, CB-Adaptive can guarantee that the circuit breaker will not trip and boost the performance of the servers compared to the current conservative practice. Section 7 provides a detailed quantitative analysis.

From Figure 2, which shows the power load behavior for racks, PDUs, and clusters in a highly-optimized Google data center [11], we observe that among racks (40 servers), PDUs (800 servers) and clusters (5000 servers), only racks occasionally get close to 100% of the possible peak aggregate server power. Since branch circuits directly feed the rack-level, we assume that branch circuits will see similar load behavior. An important point is that load behavior varies considerably between racks (branch circuits) and this is a key reason that the PDU and cluster-level load behavior does not come close to the 100% peak power consumption possible [11]. In fact, at the cluster level, only about 70% of the peak possible server power is observed.

If a data center is attempting to utilize its power infrastructure by hosting as many servers as possible, it will likely experience overloads first at the branch circuit. For this reason, we apply the CB-Adaptive method on the circuit breaker for each branch circuit. The prior data suggests that different branch circuits will overload at different times but not together. This means that we can focus effort on controlling overloads at the branch circuit and that they will not transfer all the way to the root of the power distribution system. On average, some branch circuits will need additional overload capacity while others do not, so overloads will be rarely seen at the PDU or cluster levels. In this case, only the circuit breaker and branch circuit cable is relevant for determining the length of operation in overload.

In the unlikely case that a data center workload drives all CB-Adaptive branch circuits to operate beyond 100% capacity, overloads will be experienced by higher-levels in the facility and their overload times become relevant for consideration. In this rare case, all components of the power delivery system of a data center are relevant for determining the length of operation in overload. CB-Adaptive controllers would need to be informed by a higher-level controller to determine the overload duration time.

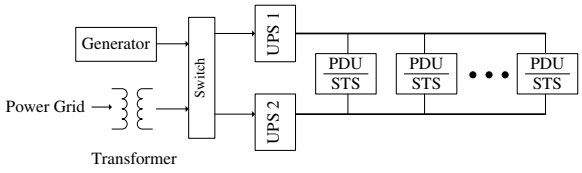


Figure 3: A typical power delivery system of a data center.

Figure 3 shows a typical power delivery system of a data center. All the components have an *overload capacity* in addition to their rated capacity. Although power overloads beyond the rated capacity in a very long term might damage a component, in many cases, tolerating short-term power overloads is necessary in practice [25]. We now summarize typical overload capacities of the components in Table 2. Note that all components can tolerate an overload higher than the value listed in the table, but components generally tolerate a higher overload for a shorter time interval.

Table 2: Overload capacities of power delivery components.

Components	Overload capacity normalized to the rating	Trip time (minutes)
Static Transfer Switch	125%	60
Various cables	125%	3.5 to 110
UPS	125%	0.5
Generator	110%	60
Transformer	150%	30

Table 2 shows the overload capacities of all components. Static transfer switch can tolerate large over currents. The limiting factor of the overload capacity is the heat dissipation [1]. If over currents

are too large, the heat generated by over currents cannot be dissipated. Cables can tolerate overloads for a short period of time [9] but overloading cables for long periods of time could damage their insulation. Generators comply with electrical standards which allow a 10% or more overload [5]. The overload capacities of transformers depend on ambient temperature, type of insulation, size of transformer and method of cooling [25]. UPS can also tolerate a short period of overload [2]. For example, certain models of data center level UPSs from APC can tolerate 125% overload for 30s. This fact implies that a single UPS basically cannot tolerate much overload. However, Figure 3 shows that in a normal state, each UPS only runs, at most, half of its capacity for fault tolerance. If one UPS is down, the power load of the UPS shut down will be transferred to the operating UPS. In the rare case where a UPS is down, it is not desirable to perform the proposed power oversubscription solutions any longer. The servers have to run at a lower power budget.

The most important contribution of our work is to provide a technically feasible solution that allows a data center to gain the maximized return on existing investments in their power supply facilities by safely accommodating the maximum number of servers. It is important to note that our technique is not limited to circuit breakers, because it can be applied to the component with the lowest tolerance level in the power delivery system. As a result, safe power oversubscription can be achieved. More importantly, our work offers insightful discussion on the technical part of the power oversubscription problem and explores a practical upper bound for power capping, revealing that a power overload is not necessarily fatal as commonly assumed.

5. IMPLEMENTATION

In this section, we introduce our physical testbed and benchmarks, as well as the implementation details of each component in the control loop.

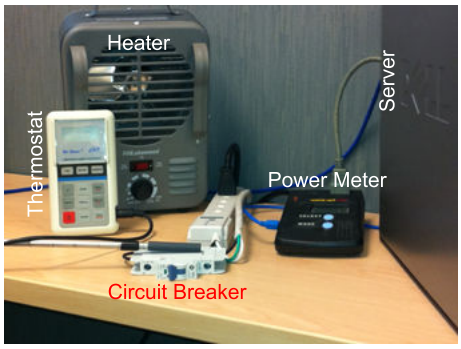


Figure 4: Hardware testbed.

Our testbed uses a single server to represent the load on the branch circuit. Note that CB-Adaptive can also be integrated with existing branch circuit level or data center level power controls. Recent proposals like SHIP [28] allow for control of branch circuit power by monitoring and controlling the aggregate power of many servers. A natural place for CB-Adaptive is within such a control system. Details of controlling multiple servers to realize an aggregate power are presented in the prior work [28].

Our testbed shown in Figure 4 consists of a Rockwell Automation circuit breaker, a heater to change the ambient temperature of the breaker, a thermostat, a power meter and a Dell OptiPlex desktop with an AMD Athlon(tm) 64 X2 Dual Core Processor 4400+ with a 2MB on-die L2 cache and 800 MHz FSB. The processor supports five DVFS levels: 2.3GHz, 2.2GHz, 2GHz, 1.8GHz, and 1GHz. The operating system is a Fedora Core 8 with a Linux kernel 2.6.23 with real-time patches. The circuit breaker model is a Rockwell Allen-Bradley 1489-A Industrial circuit Breaker with a rated current of 1A. Rockwell circuit breakers are widely used by data center operators.

We run the SPEC CPU2006 suite (V1.0), SPECJBB (2005), and High Performance Computing LINPACK Benchmark (HPL) (V1.0a) as our workloads. For the SPEC CPU2006, each performance measurement is the average of the four copies and is recorded as the performance ratio, i.e., the relative speed of the processor to complete each benchmark (compared to a reference Sun UltraSparc II machine at 296 MHz). The CPU2006 includes a collection of 29 benchmarks and is divided into CINT2006 and CFP2006, each of which consist of integer and floating-point benchmarks, respectively. SPECJBB reports throughputs of a sequence of measurements with an increasing number of warehouses. In our experiment, the number of warehouses is equal to or greater than 2. HPL is a software package that solves a (random) dense linear system in double precision (64 bits) arithmetic. The problem size of HPL is configured to be 10,000 $\times$ 10,000 and the block size is 64 in all experiments, unless otherwise noted.

The control loop consists of three components: temperature monitor and power meter (sensor), adaptive controller (controller), and CPU frequency modulator (actuator).

CB Temperature Monitor: Circuit breakers typically do not have built-in thermal sensors, however the industry is rapidly adding temperature measurement to data center products. For example, Arch Rock (Now Cisco)'s PhyNet Wireless Sensor Network is being integrated into IBM's Active Energy Manager [4]. These inexpensive and low-power sensors can easily be added to the circuit breaker panel to measure the temperature of a circuit breaker.

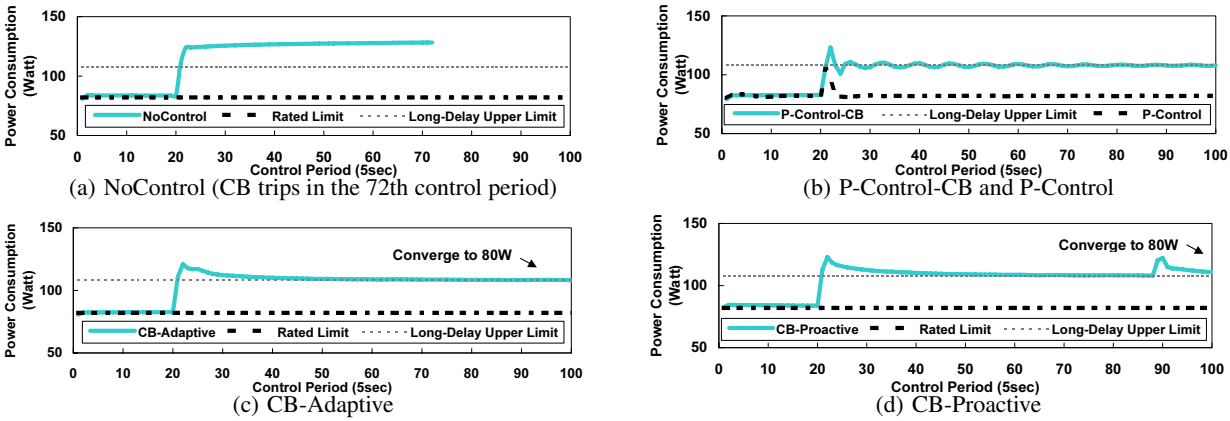
Power Meter: The power consumption of the server is measured with a WattsUp Pro power meter which has an accuracy of $\pm 1.5\%$ of the measured value. To access the power data, the data port of the power meter is connected to the USB port of the desktop. A device file is then generated for a power reading in the Linux system. The power meter samples the power data every 5 seconds and responds to requests by writing all new readings after the last request to the system file. The controller then reads the power data from the device file and conducts the control computation.

Adaptive Controller: The adaptive controller which implements CB-Adaptive or CB-Proactive runs at the highest priority (real-time priority) to guarantee fast response times. Otherwise, the controller process may be preempted by other processes with a higher priority which may cause the circuit breaker to trip due to improper control. The Linux system call *sched_setscheduler* sets both the scheduling policy and the associated parameters for the process identified by *PID*. A key advantage of CB-Adaptive and CB-Proactive is their small overheads in terms of time, space, and power consumption.

CPU Frequency Modulator: We use AMD's Cool'n'Quiet technology to enforce the new frequency. To change the CPU frequency, the cpufreq package contains command line tools to determine the current frequency levels and modify them. The root privilege is needed to write the new frequency levels into the system file /sys/devices/system/cpu/cpu[x]/cpufreq/scaling_setspeed. A BIOS routine periodically checks this file and resets the CPU frequency accordingly. The average overhead (i.e., transition latency) for the BIOS to change frequency in AMD Athlon processors is approximately 100 μs.

6. EVALUATION RESULTS

We first introduce the state-of-the-art baselines, then present our empirical results conducted on the physical testbed.



(a) NoControl (CB trips in the 72th control period)

(b) P-Control-CB and P-Control

(c) CB-Adaptive

(d) CB-Proactive

Figure 5: Power capping comparison between CB-Adaptive, CB-Proactive, and the baselines.

6.1 Baselines

Our first baseline is NoControl. NoControl estimates the peak power consumption of a server by measuring a high-power workload like SPECJBB over a few days. It assumes the real peak power consumption will never exceed the estimation. Although NoControl may run without any problems for weeks or months, it is risky because unexpected workloads or high input rates may drive even higher power consumption which causes the CB to trip. The second baseline, referred to as P-Control, is a state-of-the-art power provisioning algorithm widely deployed in IBM servers [17]. P-Control is briefly summarized as follows. 1) In each control period, the power meter on each server sends the server power consumption in the last control period to a controller through its power management infrastructure [30]. 2) The proportional controller calculates the CPU frequencies in the next control period. 3) The calculated frequencies are enforced using a first-order delta-sigma modulator. The third baseline is P-Control-CB. The only difference of P-Control-CB from P-Control is that its power budget is set according to the upper limit of the long-delay zone.

A fundamental difference between P-Control and our solutions is that P-Control assumes the power budget must stay below the rated power of circuit breaker as soon as possible, without considering the trip curve of the circuit breaker. Moreover, P-Control adopts classical proportional control without adapting the gain of the controller.

6.2 Power Capping Comparison

In this experiment set, we compare NoControl, P-Control, P-Control-CB, CB-Adaptive, and CB-Proactive under a power emergency in which the power consumption of the server increases abruptly. To emulate the power emergency, we launch a power hungry benchmark LINPACK in the middle of the experiment. Figure 5(a) shows that with NoControl, the power consumption increases from 83W to 125W after LINPACK is launched in the 20th control period. Since the server draws a much higher current than the upper-limit of the long-delay zone, the circuit breaker trips quickly, in approximately the 72th control period. Figure 5(b) shows that P-Control controls the power consumption without tripping the circuit breaker within 3 control periods to the set point which corresponds to the rated current of the circuit breaker. Similar to P-Control, Figure 5(b) also shows the P-Control-CB controls the power consumption within 3 control periods to the set point which corresponds to the upper-limit of the long-delay zone. In contrast, in Figure 5(c), it takes approximately 70 control periods for CB-Adaptive to control the power consumption to the upper limit of the long-delay zone. Within the time interval of the experiment, the power consumption is still higher than the rated limit. The reason is that CB-Adaptive changes the CPU frequencies according to the circuit breaker trip curve and the controller parameter is updated accordingly. From the 20th control period to the 90th control period, the circuit breaker runs in the conventional tripping zone. From the 90th control period on, the circuit breaker runs in the long-delay tripping zone. Since the trip time in this zone is on the scale of days, the controller decreases the frequency slowly. Thus, the decrease of the power consumption is not visible. As shown in Figure 5(d), CB-Proactive further increases the power consumption of the server by increasing the CPU DVFS level to the highest level proactively when the CB enters the long-delay zone. After the DVFS increases, an abrupt power increase is observed approximately in the 90th control period.

Note that P-Control-CB controls the power consumption to the upper limit of the long-delay zone, which is approximately 108W. Although the trip time corresponding to 108W is on the scale of days, it is not infinite. Thus, P-Control-CB cannot guarantee that the circuit breaker will not trip. Therefore, P-Control-CB is not safe and should not be used in practice. In contrast, although CB-Adaptive and CB-Proactive settle to 108W during the limited time interval of the experiment, they can guarantee that the circuit breaker will never trip by further reducing the power consumption according to the long-delay zone. This experiment set demonstrates that CB-Adaptive and CB-Proactive can safely oversubscribe the circuit breaker without tripping it during a power emergency. NoControl and P-Control-CB may cause the circuit breaker to trip during a power emergency. Although P-Control will not cause the circuit breaker to trip, as we will show in the next experiment set, the performance degradation of P-Control is large.

6.3 Performance Comparison

In this experiment set, we study the performance benefits of CB-Adaptive and CB-Proactive by comparing them to P-Control. Although P-Control-CB is not a safe power provisioning solution, we include it in the comparison to explain CB-Adaptive and CB-Proactive. We first test the solutions with LINPACK and SPECJBB, then run all the 29 benchmarks of SPEC CPU2006 to test the robustness of the solutions under a wide range of workloads.

Figure 6 compares the LINPACK and SPECJBB performance of P-Control, P-Control-CB, CB-Adaptive, and CB-Proactive. For LINPACK, the performance of P-Control is the lowest and is only 0.85 Gflops. P-Control-CB, CB-Adaptive, and CB-Proactive outperform P-Control by 66.00%, 69.06 %, and 70.12 %, respectively. The relationship between the performance and the power consump-

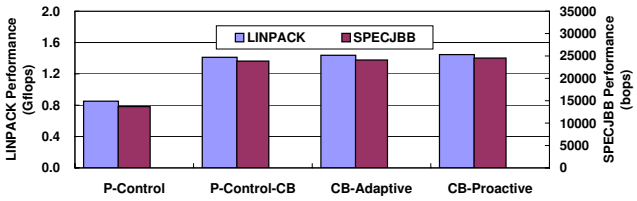


Figure 6: LINKPACK and SPECJBB performance comparison between proposed CB-aware solutions and the baselines.

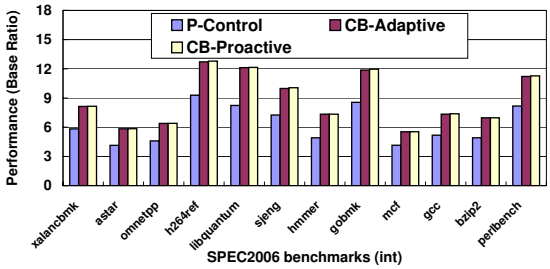


Figure 7: SPEC CPU2006 CINT performance comparison.

tion is approximately 0.02 Gflops per Watt. For SPECJBB, the performance improvement over P-Control is 74.21%, 75.93%, and 79.10% respectively. From the results, CB-Adaptive and CB-Proactive improve the performance significantly as compared to the state-of-the-art P-Control. It is demonstrated that the primary performance boost comes from the long-delay zone because within the time interval of the experiment the circuit breaker runs at the upper-limit of the long-delay zone, according to Figures 5(b), 5(c), and 5(d). Although the performance of P-Control-CB is comparable to CB-Adaptive and CB-Proactive, as shown in Section 6.2, it may trip the circuit breaker over the long-term and thus it is not safe. For CB-Adaptive and CB-Proactive, their performance is impacted significantly by the long-delay zone of a circuit breaker. For different models of circuit breakers from different manufacturers, the upper-limit of the long-delay zone may vary. Generally, the higher the upper-limit of the long-delay zone is, the better the performance can be. The slower the decreasing rate of the trip time with respect to the increasing overload current in the conventional tripping zone, the better the performance can be.

We also test the solutions by running the SPEC CPU2006 benchmarks and study their performance in terms of the base rate. We only test P-Control, CB-Adaptive and CB-Proactive in this experiment because previous experiments in Section 6.2 have shown that NoControl and P-Control-CB may trip the circuit breaker and thus should not be used in practice. Figures 7 and 8 compare the CPU2006 performance of P-Control and CB-Adaptive. As shown, CB-Adaptive achieves better performance than P-Control for all benchmarks since CB-Adaptive can provision the server at a higher power budget safely as compared to P-Control. The maximum performance improvement of CB-Adaptive is 49% over P-Control with the benchmark *hmmer*, while the minimum improvement is 29% with the benchmark *povray*. The average improvement of CB-Adaptive is 38%. This experiment set demonstrates that CB-Adaptive and CB-Proactive can boost performance significantly during an overload condition for SPEC CPU2006.

6.4 Impacts of Ambient Temperature

While the previous experiments are conducted at a normal room temperature, this experiment set studies the feasibility of incorporating the temperature into the design of CB-Adaptive. We use a fan heater to heat the circuit breaker to emulate non-uniform tem-

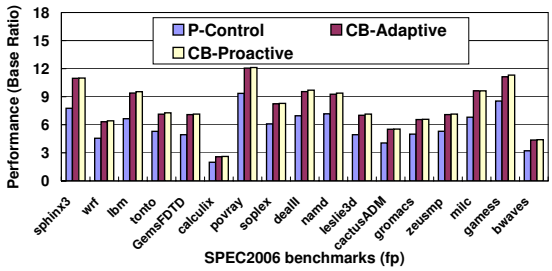


Figure 8: SPEC CPU2006 CFP performance comparison.

perature within a data center. For each experiment, we monitor the temperature of the circuit breaker to be approximately stable using a regular thermostat. We first study the impact of the temperature on the circuit breaker without any control. The actual rated current cannot be measured directly by the power meter. To examine the impact, we measure the trip time instead of the actual rated current by running the LINPACK benchmark on the two cores of the server. We vary the temperature from $21.7^{\circ}C$ to $34.8^{\circ}C$. The temperature range is a subset of the normal temperature range within a data center. We cannot lower the temperature below $21.7^{\circ}C$ without a cooler due to our room temperature limit. The first four bars of Figure 9 show that the temperature has a significant impact on the trip time of the circuit breaker. For example, the trip time of the circuit breaker is 490 seconds at $21.7^{\circ}C$ while the trip time at $34.8^{\circ}C$ is only 210 seconds. For each temperature, the trip time is an average of several repeated experiments and the deviation is negligibly small. The key feature of CB-Adaptive and CB-Proactive is their adaptive control designs based on the trip curve of the circuit breaker. Since the temperature impacts the circuit breaker, it will also impact CB-Adaptive and CB-Proactive.

We configure P-Control-CB, CB-Adaptive, and CB-Proactive based on the temperature of $10^{\circ}C$ because the low average operating temperature in some data centers is $12^{\circ}C$. Since the temperature distribution within a data center is non-uniform [8], we then test the circuit breaker at a maximum temperature of approximate $45^{\circ}C$. The last three bars of Figure 9 show that the circuit breaker still trips at $45^{\circ}C$ even though all solutions can safely provision power at $10^{\circ}C$. Since the controllers are configured to $10^{\circ}C$, they assume that the circuit breaker's rated current, as adjusted by the temperature, is 1.1 A. However, the actual rated current, adjusted by the temperature, at $45^{\circ}C$ is actually only 0.9 A. Since the controllers operate according to an over-optimistic trip curve, the temperature-blinded circuit breakers still trip. Because CB-Adaptive and CB-Proactive run at a higher power budget than P-Control-CB, they trip more quickly than P-Control-CB. This experiment demonstrates that it is necessary to adopt the temperature-aware CB-Adaptive in real data center operating environments.

6.5 Temperature Awareness

In this experiment set, we study the performance of the temperature-aware CB-Proactive presented in Section 3.2 under different temperatures and compare the performance of P-Control, P-Control-CB, CB-Adaptive, and CB-Proactive. The temperature-aware P-Control, CB-Proactive, and CB-Proactive guarantee the safety of power provisioning under different temperatures. We present the results of LINPACK because, as shown in Section 6.3, the results of SPEC CPU2006 and SPECJBB are very similar. Figure 10 shows that, as the temperature increases from $10^{\circ}C$ to $45^{\circ}C$ which is the normal temperature range of a production data center, the performance of the system decreases. The reason is that as the temperature increases, the rated current adjusted by the temperature

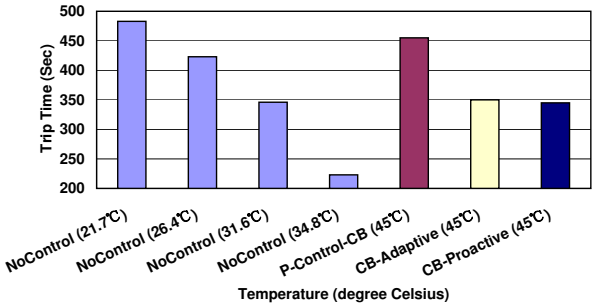


Figure 9: Impacts of ambient temperature on NoControl and three CB-aware solutions.

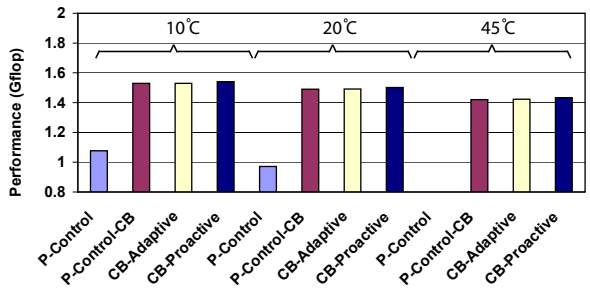


Figure 10: Impacts of temperature on LINPACK performance.

decreases. In other words, all the solutions suffer from the lower power budget, thus performance decreases. However, Figure 10 shows that the performance degradation is modest even when the temperature range is wide. For P-Control, when the temperature is $45°C$, the CPU DVFS is changed to its lowest level but the circuit breaker still trips, resulting in no performance reading. This experiment demonstrates that the temperature-aware CB-Proactive and CB-Proactive presented in Section 3.2 can successfully conduct power capping for a range of temperatures with only a modest performance degradation.

7. POWER PROVISIONING ANALYSIS

As discussed in Section 4, one of the potential applications of CB-Adaptive is hosting additional servers. In this section, we quantitatively compare NoControl, P-Control, and CB-Adaptive in terms of the maximum number of servers that can be hosted within a data center. It is shown that CB-Adaptive can host many more servers than NoControl and P-Control without a performance penalty for short-term overloads.

As shown by previous experiments, NoControl may trip a circuit breaker and is not safe. In order to make NoControl safe, instead of estimating the power consumption of a server running a fixed benchmark, we assume the power consumption is the server's nameplate power value. Thus, NoControl will never trip a circuit breaker since the power consumption will never exceed its nameplate power value. For NoControl, the number of servers within a rack is the branch circuit capacity multiplied by 80%, then divided by the nameplate power consumption. Since the nameplate power consumption of a server is very conservative, the *actual peak power consumption* of the server when running most power intensive benchmarks with 100% utilization is much smaller than its nameplate power value. For P-Control, the number of servers within a rack is the branch circuit capacity multiplied with 80%, then divided by the actual peak power consumption of the server running real data center workloads. For CB-Adaptive, the number of servers calculation is based on the branch circuit capacity multiplied by the *oversubscription ratio* (OSR).

OSR is related to the *violation interval* (i.e., the time during which the power consumption of a rack is higher than the 80% rated power capacity of the circuit breaker) and the cumulative distribution function (CDF) of a rack running real data center workloads [11]. We define OSR as:

$$OSR = \frac{0.8}{cdf^{-1}\left(1 - (violation\ time/24)\right)} \tag{11}$$

where cdf^{-1} is the inverse function of the CDF and the unit of violation time is 1 hour. A power CDF of a real Google data center [11] is shown in Figure 2. Note although Figure 2 is based on 6 months of measurements, since server activity correlates strongly with the hour of the day, we assume it is representative of a typical 24 hour period in our analysis.

Example. We now present an example to demonstrate the calculation of OSR. Suppose the violation time is 3 hours which conforms to the NEC requirement [6]. Then the ratio of the violation time to 24 hours is 12.5%. Given the percentage, we look at the y-axis of Figure 2 and find the point on the x-axis corresponding to the 87.5% (1-12.5%) on the y-axis. In this case, the value is 0.68. Thus the actual peak power consumption of the rack is $0.8/0.68 = 1.175$ which means that the actual peak power consumption of the rack is 111% of the rated capacity, i.e., $OSR = 1.175$. The number of servers within a rack is the branch circuit capacity multiplied with OSR, then divided by the actual peak power consumption of the server running real data center workloads.

For a data center configuration, the rated power capacity of each rack is about 2.5kW [11]. The nameplate power consumption of the hosted server is 251W. The actual power consumption of the server when running the most power intensive benchmark with a 100% utilization is just 145W.

Table 3: Power provisioning comparison.

Solutions	# of hosted servers
NoControl (80% branch circuit capacity)	7
NoControl	9
P-Control (80% branch circuit capacity)	13
P-Control	17
CB-Adaptive $OSR = 1$	17
CB-Adaptive $OSR = 1.1$	18
CB-Adaptive $OSR = 1.175$	20

Table 3 shows that CB-Adaptive with $OSR = 1.175$ can host 54% more servers than the state-of-the-art P-Control (conforming to NEC) in each branch circuit and about three times as many servers using NoControl (conforming to NEC).

8. RELATED WORK

Recently, the power management issue has attracted a large amount of attention from both academia and industry. For example, Meisner et al. [18] proposed a PowerNap scheme to reduce the server's idle power. Ahmad et al. [8] optimized the idle and cooling power in a data center. However, these studies focus primarily on power minimization instead of power provisioning.

Power provisioning is an important technique for data centers to avoid expensive upgrade costs and to maximize the power infrastructure utilization; thus, it becomes an important, practical issue in data center operation. Fan et al. [11] investigated the workload characteristics of the data center and demonstrate the existence of a great potential for oversubscription in the production data center. Lefurgy et al. [17] proposed a control-theoretic approach to power provisioning and showed the advantages of this method in terms of performance as compared with commercial ad hoc solutions. Pelley et al [20] proposed a novel power router to make the

flexible power budget usable. Femal et al [12] investigated how to improve throughoutput given a fixed power budget. Yet, each of these studies still does not answer the question of how much power can be safely over-subscribed. Govindan et al. [13] adopted statistical profiling-based techniques for power provisioning. They considered sustainable power budgets; however, they did not systematically investigate the CB tripping characteristics. In addition, their soft fuse method is essentially a heuristic-based approach.

The control-theoretic approach is a promising adaptation mechanism in power and thermal management. Donald et al.[10] proposed a PI-controller based solution for multicore thermal management. Skadron et al. [23] designed a PID controller approach for accurate and localized dynamic thermal management. Srikantaiah et al [24] adopted a reinforced oscillation resistant controller for shared cache management. Wang et al. [29] designed a model prediction controller to limit the peak power of a chip multiprocessor. Those studies focus on power and thermal management issues for individual computer systems. None of them consider the adaptation of control parameters since there is no design constraint on the settling times of controllers in those studies.

9. CONCLUSIONS

While a variety of power capping solutions have been recently proposed, a conservative assumption made by existing solutions is that peak power should never exceed the rated CB capacity. In this paper, we systematically study the tripping characteristics of a typical CB used in many data centers. We identify that the ideal upper bound of safe power oversubscription is the lower bound of the tolerance band in the trip curve of the circuit breaker. We then propose two adaptive power control strategies that utilize the tripping characteristics of the CB to aggressively optimize the system performance without causing the CB to trip. Furthermore, our control schemes can also adapt to the variation of ambient temperature that is known to affect the CB tripping behaviors. Empirical results on a physical testbed show that the proposed CB-aware power control solutions achieve 29% to 49% better SPEC CPU2006 performance than a state-of-the-art baseline, with an average of 38%. The proposed solutions also achieve 68% better LINPACK performance and 75% better SPECJBB performance than the baseline. In addition, our solutions allow a data center to host three times more servers than traditional static power provisioning schemes and 54% more servers than the current power capping practice.

10. ACKNOWLEDGEMENTS

This work is supported, in part, by NSF under CAREER Award CNS-0845390 and Grants CNS-0720663 and CCF-1017336. We thank Dr. Leon Tolbert at the University of Tennessee for his help with the devices in data center power delivery systems.

11. REFERENCES

[1] AEG Static Transfer Switch Technical Specifications.

[2] Experts speak on UPS output Watt, VA, and Power Factor ratings. http://www.ptsdcs.com/whitepapers/12.pdf.

[3] IBM PowerExecutive Installation and User's Guide. ftp://ftp.software.ibm.com/systems/support/system_x_pdf/pwx2.10_docs_user.pdf.

[4] IBM to Support Arch Rock's PhyNet Wireless Sensor Network in Active Energy Manager. http://www.businesswire.com/news/home/20100111005596/en/IBM-Support-Arch-Rocks-PhyNet-Wireless-Sensor.

[5] Mitsubishi diesel generator technical specifications. http://powercare.com.au/catalog/i32.html.

[6] *National Fire Protection Association National Electrical Code*. 2008.

[7] Rockwell Automation Inc Bulletin 1489 Circuit Breakers Selection Guide. http://literature.rockwellautomation.com/idc/groups/literature/documents/sg/1489-sg001_-en-p.pdf, 2010.

[8] F. Ahmad and T. N. Vijaykumar. Joint optimization of idle and cooling power in data centers while maintaining response time. In *ASPLOS*, 2010.

[9] R. Atkinson and H. W. Fisher. Current rating of electrical cables. *Transactions of the American Institute of Electrical Engineers*, Feb. 1913.

[10] J. Donald and M. Martonosi. Power efficiency for variation-tolerant multicore processors. In *ISLPED*, 2006.

[11] X. Fan, W.-D. Weber, and L. A. Barroso. Power provisioning for a warehouse-sized computer. In *ISCA*, 2007.

[12] M. Femal and V. Freeh. Safe overprovisioning: Using power limits to increase aggregate throughput. In *PACS*, 2004.

[13] S. Govindan et al. Statistical profiling-based techniques for effective power provisioning in data centers. In *EuroSys*, 2009.

[14] HP. Dynamic Power Capping TCO and Best Practices White Paper. http://h20195.www2.hp.com/v2/GetPDF.aspx/4AA2-3107ENW.pdf.

[15] C. Isci, A. Buyuktosunoglu, C.-Y. Cher, P. Bose, and M. Martonosi. An analysis of efficient multi-core global power management policies: Maximizing performance for a given power budget. In *MICRO*, 2006.

[16] H. Joshi. *Residential, Commercial and Industrial Electrical Systems: Equipment and selection*. Tata McGraw-Hill, 2008.

[17] C. Lefurgy, X. Wang, and M. Ware. Server-level power control. In *ICAC*, 2007.

[18] D. Meisner, B. T. Gold, and T. F. Wenisch. Powernap: Eliminating server idle power. In *ASPLOS*, 2009.

[19] K. Meng, R. Joseph, R. P. Dick, and L. Shang. Multi-optimization power management for chip multiprocessors. In *PACT*, 2008.

[20] S. Pelley, D. Meisner, P. Zandevakili, T. F. Wenisch, and J. Underwood. Power routing: Dynamic power provisioning in the data center. In *ASPLOS*, 2010.

[21] R. Raghavendra, P. Ranganathan, V. Talwar, Z. Wang, and X. Zhu. No power struggles: Coordinated multi-level power management for the data center. In *ASPLOS*, 2008.

[22] P. Ranganathan et al. Ensemble-level power management for dense blade servers. In *ISCA*, 2006.

[23] K. Skadron, T. Abdelzaher, and M. R. Stan. Control-theoretic techniques and thermal-rc modeling for accurate and localized dynamic thermal management. In *HPCA*, 2002.

[24] S. Srikantaiah, M. Kandemir, and Q. Wang. Sharp control: Controlled shared cache management in chip multiprocessors. In *MICRO*, 2009.

[25] S. Tenbohlen et al. Assessment of overload capacity of power transformers by on-line monitoring systems. In *Power Engineering Society Winter Meeting*, 2001.

[26] United States Environmental Protection Agency. Report to congress on server and data center energy efficiency, 2007.

[27] X. Wang and M. Chen. Cluster-level feedback power control for performance optimization. In *HPCA*, 2008.

[28] X. Wang, M. Chen, C. Lefurgy, and T. W. Keller. SHIP: Scalable Hierarchical Power Control for Large-Scale Data Centers. In *PACT*, 2009.

[29] Y. Wang, K. Ma, and X. Wang. Temperature-constrained power control for chip multiprocessors with online model estimation. In *ISCA*, 2009.

[30] M. Ware et al. Architecting for power management: The IBM POWER7 approach. In *HPCA*, 2008.