A Market Approach for Handling Power Emergencies in Multi-Tenant Data Center*

Mohammad A. Islam Xiaoqi Ren† Shaolei Ren Adam Wierman† Xiaorui Wang‡
University of California, Riverside †California Institute of Technology
‡The Ohio State University

ABSTRACT

Power oversubscription in data centers may occasionally trigger an emergency when the aggregate power demand exceeds the capacity. Handling such an emergency requires a graceful power capping solution that minimizes the performance loss. In this paper, we study power capping in a multi-tenant data center where the operator supplies power to multiple tenants that manage their own servers. Unlike owner-operated data centers, the operator lacks control over tenants' servers. To address this challenge, we propose a novel market mechanism based on supply function bidding, called COOP, to financially incentivize and coordinate tenants' power reduction for minimizing total performance loss (quantified in performance cost) while satisfying multiple power capping constraints. We build a prototype to show that COOP is efficient in terms of minimizing the total performance cost, even compared to the ideal but infeasible case that assumes the operator has full control over tenants' servers. We also demonstrate that COOP is "win-win", increasing the operator's profit (through oversubscription) and reducing tenants' cost (through financial compensation for their power reduction during emergencies).

1. INTRODUCTION

The emergence of Internet and cloud services has significantly fueled demand for data centers worldwide, resulting in an aggregate power demand of 38GW as of 2012 (a growth of 63% compared to 2011) [1]. Accommodating the accelerated demand, however, is costly. It can be a multi-million or even multi-billion dollar project to construct a new data center or expand an existing data center's capacity (typically measured in IT critical power). For example, power infrastructure, including back-up generation and uninterrupted power supplies (UPS), is sized based on the critical power budget and estimated at U.S.$10-25 per watt [2]. The capital expense (CapEx) in power and cooling infrastructure even exceeds 1.5 times of the total energy cost of operating a data center over a 15-year lifespan [2–4]. Moreover, other limitations, such as space and grid capacity, may also prohibit the expansion of data center capacity.

In view of the high CapEx and practical constraints for building new capacity, data center operators aggressively oversubscribe the existing infrastructure throughout the power hierarchy (e.g., UPS level and PDU level) by deploying more servers than the power budget/capacity allows. This is equivalent to under-provisioning the capacity to reduce CapEx for new data center construction: to deploy the same number of servers, the data center capacity can be downsized to save CapEx. The rationale underlying oversubscription is that, in most cases, not all servers simultaneously run at their peak powers and thus, the servers' aggregate power usage remains well below the power budget with a very high probability, as illustrated by measurements in [3, 5].

A dangerous consequence of oversubscription is the emergence of power *emergencies* that bring significant challenges for data center uptime. Although uncommon, when loads on many servers peak simultaneously, the aggregate power demand will exceed the capacity (e.g., overloading UPS), thus compromising the desired power availability and even leading to unplanned downtime incidents [4, 6]. Such power emergencies have become a major cause of unplanned data center outages, which may take several hours or even days to fully recover and incur significant economic losses (estimated average of $901,560 per incident) [7, 8].

Naive techniques to handle emergencies, e.g., arbitrarily putting involved servers into low power states or switching them off, are not appealing [3, 6, 9], because they may result in significant performance degradation and even business disruption. Instead, a graceful power capping solution is required to coordinate servers' power usage at a minimum performance loss. Towards this end, prior research has studied various techniques, e.g., judiciously scaling down CPU frequency [6], admission control, and workload migration (to public clouds and/or other servers not subject to power emergency) [5, 10, 11]. These studies, although promising, are only applicable for owner-operated data centers (e.g., Google), where data center operators have control over the physical servers and hence can easily coordinate the servers to minimize performance impact.

In sharp contrast, we study power capping in multi-tenant data centers, an under-explored but even more common type of data center. In a multi-tenant data center, multiple individual tenants house and manage their own physical servers, while the data center operator is responsible for power and cooling infrastructure support. Like owner-operated data centers, multi-tenant data center operators also aggressively oversubscribe capacity to gain more revenue and/or save CapEx by selling the capacity to more tenants than it allows.

*This work was supported in part by the National Science Foundation under grants CNS-1143607 (CAREER), CNS-1319820, CNS-1421452, CNS-1551661 (CAREER), and CNS-1565474.

978-1-4673-9211-2/16/$31.00 ©2016 IEEE

To handle a power emergency due to oversubscription, however, multi-tenant data center operators cannot directly apply existing power capping techniques (e.g., through server and workload management [5, 6]), because of lack of control over tenants' servers. Thus, a common practice today for a multi-tenant data center operator is to simply take the risk of capacity overloading when many tenants' power demands peak simultaneously. In other words, whether or not an outage will occur depends largely on the robustness of infrastructure. Consequently, according to a 2014 Uptime Institute survey, 25% of the tenants have experienced at least one power outage (for which capacity overloading is a major cause) over the past year [7, 8, 12].

To address the lack of coordination among tenants to shed power during a power emergency, we propose a novel CO-Ordinated Power management solution, called COOP, that leverages a market mechanism called supply function bidding [13] commonly used in electricity markets in order to incentivize and coordinate individual tenants' power demand reduction. The challenge in designing such a mechanism is that the mechanism must not bring much overhead and the overall impact of power reduction on tenants' application performance needs to be as little as possible (similar to the design objective of power capping techniques, e.g., [5,6], for owner-operated data centers). COOP achieves both. It only solicits one bidding parameter from each participating tenant which, when plugged into the supply function, specifies the amount of power reduction and corresponding reward the tenant is willing to accept. More importantly, the overall performance impact across all the participating tenants is small: the total performance cost incurred by the tenants is very close to the ideal case where the data center operator is assumed to have full control over tenants' servers.

Contribution. The novelty of this study is that COOP is the first market-based solution for handling an emergency caused by capacity oversubscription in a multi-tenant data center, an important yet rarely-studied type of data center.

Concretely, this paper makes the following contributions. First, we introduce and formulate the problem of multi-level power capping in a multi-tenant data center. Second, we propose a supply function bidding based mechanism, motivated by the literature on electricity markets [13, 14], to incentivize and coordinate tenants' power reduction during a power emergency, capping the aggregate power demand while minimizing the total performance cost. Third, we validate COOP using realistic settings on a testbed. Our results show that COOP is efficient in terms of minimizing total performance cost and that COOP is "win-win", increasing the operator's profit and reducing tenants' cost (through financial compensation).

2. OPPORTUNITIES AND CHALLENGES

Multi-tenant data centers are common in practice. There are over 1,400 multi-tenant data centers in the U.S. alone [15]. As a quickly growing data center segment, it consumes as much as **five times** energy of those Google-type owner-operated data centers combined together (37.3% v.s. 7.8%, in percentage relative to all data center energy usage, excluding tiny server closets) [16]. It provides a cost-effective and scalable data center solution to many industry sectors, including major websites (e.g., Twitter), banking, content delivery provider (e.g., Akamai) [17], and even IT giants (e.g., Microsoft) that leverage third-party data centers to complement their own facilities [18].

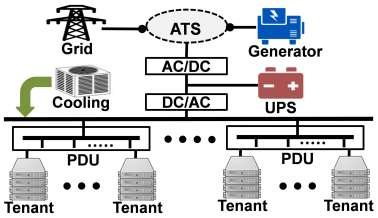


Figure 1: Data center infrastructure.

Despite their importance, multi-tenant data centers have been less investigated than the more visible owner-operated data centers (like Google). They present new challenges due to the operator's lack of central control over servers. This means that standard approaches for handling power emergencies do not apply to multi-tenant data centers [5, 6, 11].

2.1 Power in Multi-Tenant Data Centers

While different designs (e.g., using fuel cell as the main power source [19]) are emerging, most data centers, including new constructions, still heavily rely on diesel generators, uninterrupted power supplies (UPS), and power distribution units (PDU) for achieving high power availability. Fig. 1 illustrates the infrastructure commonly found in today's multi-tenant data centers: electricity first enters data center through a utility substation; next, through AC/DC and DC/AC double conversions, power goes to PDUs, which then distribute power to individual tenants' server racks. By default, the automatic transfer switch (ATS) takes power from the utility and, in the event of a grid failure, switches to the back-up generator. As the generator cannot be instantly turned on, a UPS will be discharged to supply continuous power until the diesel generator is fully activated.

The power hierarchy. In a multi-tenant data center, the power hierarchy often has a tree-type structure. At the top level sits the centralized UPS, which supports multiple cluster-level PDUs at a lower level. Each cluster-level PDU typically has a capacity of 200-300kW, supporting around 50 racks which then distributes power to servers at the lowest level. Individual tenants may have highly diverse power demands, ranging from a few kW (often in a retail multi-tenant data center) to hundreds of kW or even larger, depending on their needs. Each PDU or even rack may also have its own dedicated UPS (e.g., lead-acid battery, not shown in Fig. 1), which complements or even fully substitutes the centralized UPS while enhancing power availability at a lower cost [4].

The data center operator also provides reliable cooling. Among various designs, multi-tenant data center usually uses mechanical chiller or direct-expansion air conditioning as the cooling mechanism, depending on the data center size.

Tenants' power usage. A crucial motivation for power oversubscription is the heterogeneity of tenant power usage. We present in Fig. 2 the temporal analysis of power measurement in a commercial multi-tenant data center collected from May to July, 2015. The data includes the server power usage of 10 tenants, subscribing approximately 500kW in total and ranging from sectors of utility, education, media, content distribution and public clouds. Fig. 2 shows the cu-

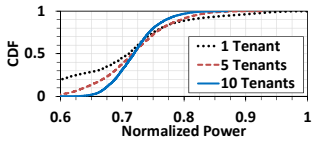


Figure 2: CDF of measured power usage.

mulative density function (CDF) of the power consumption by different groups of tenants, from 1 tenant to 10 tenants. The x-axis is normalized with respect to the sum of the maximum power usage of all the servers within that group of tenants.

We see that with more tenants (i.e., power hierarchy moves up from rack to clusters), statistical multiplexing effects of power demands become more significant, and it is even rarer for all tenants' servers to peak simultaneously. For example, for one tenant, the probability that the normalized power exceeds 80% of its peak is roughly 13%, whereas this number reduces to less than 3% for 10 tenants. Such observation has also been reported for owner-operated data centers like Google [9], whose sever clusters are equivalent to tenants'. While the specific CDF of power usage varies with different data centers, the qualitative insights hold widely: power oversubscription is safe in *most* cases.

2.2 Opportunities for Oversubscription

Leasing data center capacity with power and cooling (typically $150/kW/month in the U.S. [20]) is the most significant revenue source for a multi-tenant data center operator. Naturally, through oversubscription, the operator can earn extra revenue by serving more tenants without upgrading the power/cooling infrastructure.[1] Except for increased risk of downtime (due to capacity overloading), there is almost no additional operating expense resulting from the operator's oversubscription, because in many large (especially wholesale) multi-tenant data centers, the energy cost will be split across tenants depending on their actual usage.

In Table 1, we show the potential economic benefit of oversubscription for multi-tenant operators, based on a leasing cost of $150/kW/month [20]. For each kW capacity, with x% oversubscription, the operator earns an extra revenue of $\$150 \times 12 \times x\%$ per year. Overloading probability is obtained based on measured power usage of the 10-tenant cluster shown in Fig. 2: with x% oversubscription, overloading occurs if the aggregate power demand exceeds $100/(100+x)$ of its maximum.

We see from Table 1 that there is a great economic opportunity for oversubscription. The last row in Table 1 shows the maximum reward rate for power reduction that can be offered to tenants without decreasing the operator's profit (assuming that during each power emergency, the tenants' server power demand reaches the peak, i.e., rated capacity plus the oversubscribed amount). If the operator is not too aggressive and oversubscribes its capacity by less than 20%, it can offer a reward rate at more than 200 times of the market electricity price without losing profit.

[1]A tenant may also oversubscribe its reserved capacity to reduce leasing cost, but it must handle resulting emergencies by itself. Thus, this is addressed by prior research [5, 6].

Table 1: Analysis of Capacity Oversubscription.

Oversubscription	10%	15%	20%	25%
Extra Revenue ($/kW/year)	180	270	360	450
Probability of Overloading (%)	1%	1%	2%	3%
Est. Overloading Time (hours/year)	88	88	175	262
Max. Reward for Power Reduction ($/kW/hour)	22.60	23.63	12.32	8.56

2.3 Challenges for Oversubscription

The economic benefit of oversubscription is significant, but the danger of creating power emergencies cannot be ignored. While a power emergency may not necessarily lead to a downtime given infrastructure redundancy (e.g., "2N" duplicating all power/cooling units), ignoring it without proper attention is not a good practice, as IT critical loads exceeding the design capacity will lose the desired redundancy protection and increases outage risk [4, 5]. In fact, according to a recent survey [12], despite redundancy, 25% of the tenants have experienced at least one unplanned power-related downtime over the past year (for which IT loads' exceeding the design capacity is a major cause [7, 8]). Therefore, regardless of redundancy protection, it is very critical to handle power emergencies by gracefully capping the servers' power demand below the design capacity.

One approach for handling power emergencies is to temporarily "boost" power supply by discharging an energy storage device (ESD, e.g., battery in UPS) [4, 21]. However, a potential risk when leveraging ESD is that the cooling capacity (typically sized based on the IT critical load due to high CapEx) may still be exceeded [22, 23], because the servers' actual aggregate power consumption (i.e., cooling load) is not reduced to the designed level. As a result, discharging ESD can safely handle power capacity overloading, but *not* necessarily cooling capacity overloading, which can quickly lead to server overheating and is another major cause for unplanned outages [8]. Moreover, inappropriate/frequent discharging may drain the ESD sooner and compromise data center reliability (e.g., recent Google power failure incident, for which Google cited "extended or repeated battery drain" as a root cause [24]).

In this paper, we propose to handle power emergencies via IT power reduction from the tenants. In practice, however, the operator lacks control over tenants' servers and hence cannot enforce tenants' power reduction during an emergency, which is due to the operator's *fault* of oversubscription. Even assuming that the operator can somehow force tenants to cut power, which tenants should reduce power and by how much still needs to be decided so as to minimize tenants' performance degradation. This requires the knowledge of tenants' workloads and business values, which is private information and unknown to the operator. Thus, despite the huge economic benefit of power oversubscription, gracefully capping tenants' power to handle the resulting emergencies with a minimum performance impact on tenants presents significant challenges for multi-tenant data center operators.

3. THE DESIGN OF COOP

COOP is a market-based approach for extracting power

reduction from tenants when faced with a power emergency. The design takes inspiration from and also extends literature [13, 14] in electricity markets. In an electricity market, the market operator typically solicits bids from individual generators to reveal their planned generation amounts and at what prices, through a process called *supply function bidding*. Supply function bidding allows simple bidding that does not reveal private cost information. It also has strong theoretical support: prior work has proven that it is cost efficient, even compared to the ideal case of centralized management [13, 14].

3.1 Problem Formulation

A data center typically oversubscribes capacity at multiple interdependent power hierarchies (e.g., data center UPS-level, cluster PDU-level, and even rack-level), each having its own capacity below which the involved tenants' aggregate power demand should be capped at all times [5, 25]. COOP is not restricted to any particular levels. Like prior research [5], we consider the most typical two-level power oversubscription, i.e., cluster PDU-level and data center UPS-level, referred to as low and high levels, respectively.

Model. Consider a power emergency that involves a centralized UPS supporting M cluster PDUs and a total of N tenants denoted by a set $\mathcal{N}_0 = \{1, 2, \cdots, N\}$. The i-th PDU supplies power to a subset of tenants $\mathcal{N}_i \subseteq \mathcal{N}_0$, and the tenants served by two different PDUs are non-overlapping (i.e., $\cup_{i=1}^{M} \mathcal{N}_i = \mathcal{N}_0$ and $\mathcal{N}_i \cap \mathcal{N}_j = \varnothing$ if $i \neq j$). The high-level UPS capacity is exceeded by $D_0 \geq 0$, while the i-th low-level PDU capacity is exceeded by $D_i \geq 0$. Suppose that tenant i cuts power by s_i and incurs a cost of $c_i(s_i)$ that is increasing in s_i. Cutting power may result in service quality or performance degradation, and the cost $c_i(s_i)$ can therefore be interpreted as the performance cost, which converts the performance degradation into a monetary value. The function $c_i(s_i)$ is decided at the tenant's sole discretion as its *private* information that is unknown to the operator.

Objective. Like power capping for owner-operated data centers [5,6], we consider an *equivalent* objective: minimizing tenants' overall performance loss, formalized below.

$$\min_{s_i \geq 0, i=1,2,\cdots,N} \sum_{i=1}^{N} c_i(s_i) \tag{1}$$
$$s.t., \quad \sum_{i \in \mathcal{N}_j} s_i \geq D_j, \text{ for } j = 0, 1, 2, \cdots, M,$$

where the objective of (1) is a scalar measure of overall performance loss and impact on tenants, and the constraint specifies power capping requirements at the high (D_0) and low (D_j for $j = 1, \cdots, M$) levels, respectively.

Tenants typically test power-performance profiles before production deployment, since power is a major cost for tenant's leasing [5, 20]. Thus, given its own traffic load, tenant knows how much power can be shed and at what cost. If they are uncertain at runtime (due to, e.g., changes in power profiles), tenants can evaluate costs conservatively (see Section 5.6); hence, repeated profiling of $c_i(s_i)$ at runtime is not necessary, and the overhead for participating tenant is small.

The ideal case is when the operator can directly minimize the cost in (1), with full control over tenants' servers as in an owner-operated data center. We refer to the outcome of this idealized, but not feasible in practice, case as OPT. The choice of objective in (1) may seem counterintuitive, so let us discuss it briefly. One might expect to have the objective be operator profit. However, the operator has a priority of minimizing the impact of power capping on tenants' operation during an emergency since it is the operator's *fault* (due to oversubscription) the emergency occurred. This objective is consistent with prior power capping research on owner-operated data centers [5,6] and, further, in our context, if the operator still attempts to make profits during an emergency, power outage risk may increase, which is unacceptable since downtime incidents will significantly damage the operator's business image as well as its long-term profit. Additionally, note that the operator will not lose profit during emergency events since it can always set an upper bound (according to Table 1) to ensure that it will not lose profit due to oversubscription while minimizing tenant impact.

3.2 A Market-Based Solution

We propose a market mechanism COOP, based on parameterized supply function bidding [13], to incentivize and coordinate tenants' power reduction for power capping.

Like gathering cost information from generators to decide cost-effective generation in an electricity market, COOP asks tenants to report a pre-determined form of a *supply function* to the operator, indicating how much power they can reduce and at what prices. In our context, the operator has a demand of power reduction, while the tenants (i.e., suppliers) bid to fulfill the demand and receive rewards. The key to solving (1) is to decide the amount of supply provided by each tenant (i.e., power reduction), which is through a supply function explained as follows.

We consider a parameterized supply function $s_i(b_i, r) = \left[\delta_i - \frac{b_i}{r}\right]^+$, where δ_i with a unit of kW indicates tenant i's maximum possible power reduction, b_i is its bid (with a unit of \$) and r is the reward (\$ per kW, also called "price" in mechanism design) offered by operator to all tenants. The sign "+" in the supply function indicates that tenant cannot supply negative power (i.e., increase power). Similar forms of supply functions have also been considered in prior literature for power markets [13, 14].

The supply function $s_i(b_i, r) = \left[\delta_i - \frac{b_i}{r}\right]^+$ indicates tenant i's willingness to reduce its power by $s_i(b_i, r)$ if the operator offers r for each kW reduction. The actual power reduction is jointly determined by the following sequence.

Step 1: *Operator decides δ_i.* The data center operator decides δ_i and announces the form of supply function $s_i(b_i, r) = \left[\delta_i - \frac{b_i}{r}\right]^+$ to tenants by signalling to tenants' server control interfaces. Tenant i's current power usage can be set as its maximum possible power reduction δ_i (i.e., power reduction if tenant i shuts down all its servers).

Step 2: *Tenant decides b_i.* With the price r as an unknown variable, tenant i individually chooses and submits a bid b_i to the operator. Essentially, tenant i reports to the operator its power reduction flexibility: *if* offered a price of r, then it will cut power by $s_i(b_i, r)$. In other words, given b_i, the actual power reduction is still a function of the variable r.

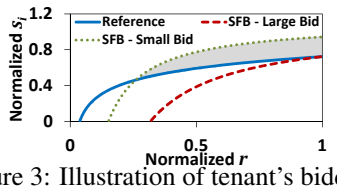


Figure 3: Illustration of tenant's bidding.

(We will discuss how to choose b_i later.)

Step 3: *Operator decides r.* Once the operator receives tenants' bids, it needs to decide r (called market clearing price), which is plugged into $s_i(b_i,r) = \left[\delta_i - \frac{b_i}{r}\right]^+$ to determine tenant i's power reduction.

How to choose bid b_i? Tenants have the full discretion to decide their own bids. We first illustrate the impact of bid on tenant's power reduction in Fig. 3. As a *reference*, we also plot the tenant's maximum power reduction without losing profit: the maximum power reduction s_i such that tenant i's net profit, i.e., operator's payment minus tenant's private performance cost, is non-negative. Given a price r, reducing more power than this reference value will incur a profit loss for tenants. We see from both $s_i(b_i,r) = \left[\delta_i - \frac{b_i}{r}\right]^+$ and Fig. 3 that a larger b_i means that tenant i is less willing to cut power given the same price r. We also notice that a too small bid may result in a profit loss when tenants are offered higher prices (i.e., shaded area in Fig. 3).

An expected outcome is the *equilibrium* point, at which each tenant i maximizes its net profit "$r \cdot s_i - c_i(s_i)$," thus having no incentives to choose arbitrarily high bids and representing a stabilized outcome. A brief explanation of equilibrium is provided in the appendix, and readers may refer to [13] for more details.

Setting too large a bid deviates from an equilibrium point, since tenant will be *priced out* or only asked to reduce a small amount of power when other participating tenants can reduce power at lower prices. For example, if $b_i \to \infty$, tenant i will be excluded from the mechanism without being asked to reduce any power, i.e., $s_i(b_i,r) = \left[\delta_i - \frac{b_i}{r}\right]^+ = 0$.

Tenants have the discretion to decide their bids, but the final price is set by the operator (which determines the actual power reduction for each tenant) and rational tenants will bid reasonably based on their private costs $c_i(s_i)$. One bidding strategy is that b_i is just large enough to avoid net profit loss over a price range (i.e., as illustrated in dashed line in Fig. 3).

To guide tenants' bidding towards the equilibrium, the operator can tell the tenants its expected price range (e.g., market price r will only be within $[r_{\min}, r_{\max}]$), such that tenants can bid to avoid profit loss by considering this restricted price range rather than the entire range.

How to decide price r? Given tenants' bids, the operator's goal is to set price r as low as possible, while satisfying *all* the power capping constraints. It is clear that, to ensure $\sum_{i\in\mathcal{N}_j} s_i \geq D_j$, the price r needs to satisfy $\sum_{i\in\mathcal{N}_j} s_i(b_i,r) = \sum_{i\in\mathcal{N}_j}\left[\delta_i - \frac{b_i}{r}\right]^+ \geq D_j$. Thus, the market price r can be decided as $r = \min_{r'}\{r' \in [r_{\min}, r_{\max}] \,|\, \sum_{i\in\mathcal{N}_j} s_i(b_i,r') = \sum_{i\in\mathcal{N}_j}[\delta_i - \frac{b_i}{r'}]^+ \geq D_j,\ \text{for } j = 0,1,\cdots,M\}$, i.e., the minimum price that satisfies all the power capping constraints and is within the range $[r_{\min}, r_{\max}]$. If no such price exists, the operator needs to activate the *failover* mode (see Section 3.4).

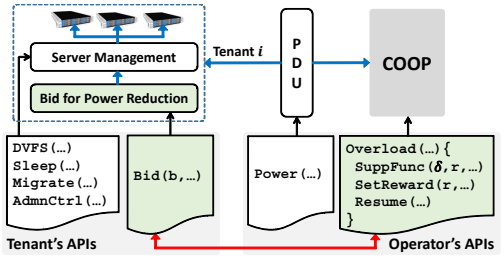


Figure 4: API diagram for COOP.

Scalability. COOP is highly scalable, as determination of the bid is performed by individual tenants in parallel and the market price is decided based on a simple rule $r = \min_{r'}\{r' \in [r_{\min}, r_{\max}] \,|\, \sum_{i\in\mathcal{N}_j}\left[\delta_i - \frac{b_i}{r'}\right]^+ \geq D_j,\ \text{for } j = 0,1,\cdots,M\}$. In practice, there are at most a few tens of tenants in wholesale data centers, and typically no more than a few hundreds of tenants in retail data centers. In either case, the complexity of COOP is reasonably low (further shown in Section 5).

Theoretical support. COOP adapts supply function bidding to the setting of power capping in multi-tenant data centers. Prior studies on supply function bidding in electricity market contexts have theoretically proved the cost efficiency of supply function bidding [13, 14] by showing that the equilibrium outcome never deviates much from the centralized optimum. While these results are often based on different contexts than we consider here, the core of the results can be adapted to this context to provide a theoretical support and justification for both our choice of a supply function bidding mechanism and the form of supply function we adopt. Further, we show through practical experiments in Section 5 that COOP can efficiently allocate power reduction among tenants for power capping, yielding a total performance cost close to that of the ideal case OPT.

3.3 Implementation

To implement COOP, we introduce a set of new APIs, for both the operator and tenants, as illustrated in Fig. 4, where new APIs are inside shaded boxes. The system flow is also described in Algorithm 1, where we consider a general two-level capping. Note that, if only a few low-level PDUs are overloaded without exceeding the high-level UPS capacity shared with other non-overloaded PDUs, then the operator will only notify tenants served by these overloaded PDUs to participate in COOP.

• **Detecting power emergency.** The operator monitors power by accessing power meter API `Power( )` at runtime, which is already in place in multi-tenant data centers. While short-duration load spikes (e.g., a few seconds) can be tolerated by the infrastructure itself [6, 8], a *sustained* power emergency of capacity overloading should invoke the power capping mode and execute COOP. The time threshold, i.e., T_w in Algorithm 1, for deciding power emergency depends on how much the aggregate demand exceeds the capacity: if not too much, a larger T_w (e.g., a few tens of seconds) is used; and vice versa.

• **Executing market mechanism.** Upon a power emergency, the market mechanism is executed following the steps described in Section 3.2 using new APIs. Specifically, the

Algorithm 1 COOP: Coordinated Power Management

1: Input: UPS and PDU capacities P_i^{cap} for $i = 0, 1, \cdots, M$
2: Monitor UPS and PDU power $P_i(t)$ continuously.
3: **if** $P_i(t) > P_i^{cap}$ for any $i = 0, 1, \cdots, M$ **then**
4: Start waiting timer T_w
5: **end if**
6: **while** T_w has not expired **do**
7: **if** $P_i(t) \leq P_i^{cap}$ for all $i = 0, 1, \cdots, M$ **then**
8: Go back to Line 2
9: **end if**
10: **end while**
11: ▷ *Entering "power capping" mode*
12: **if** $P_i(t) > P_i^{cap}$ for any $i = 0, 1, \cdots, M$ **then**
13: Set $D_i \leftarrow \left[P_i(t) - P_i^{cap}\right]^+$
14: Announce $s_i(b_i, r) = [\delta_i - \frac{b_i}{r}]^+$ to tenant i
15: Tenant i decides its bid b_i
16: Set price $r = \min_{r'}\{r' \in [r_{\min}, r_{\max}] \,|\, \sum_{i \in \mathcal{N}_j} s_i(b_i, r') \geq D_j, \text{ for } j = 0, 1, \cdots, M\}$
17: Each tenant i reduces $s_i(b_i, r)$ power
18: **end if**
19: ▷ *Leaving "power capping" mode*
20: **wait** until $P_i(t) \leq P_i^{cap} - D_i$ for all $i = 0, 1, \cdots, M$
21: Start capping timer T_c and **wait** until T_c expires or $P_i(t) > P_i^{cap} - D_i$ for any $i = 0, 1, \cdots, M$
22: **if** $P_i(t) > P_i^{cap} - D_i$ for any $i = 0, 1, \cdots, M$ **then**
23: Go back to Line 20
24: **end if**
25: **if** T_c expires **then**
26: Notify tenants to resume normal operation
27: Calculate the power capping duration T_o
28: Provide tenant i with a reward of $z_i = T_o \cdot r \cdot s_i$
29: Go back to Line 2
30: **end if**

operator communicates the supply function to tenants through `SuppFunc`$(\delta, r, \cdots)$ where the price r is a parameter to be decided, and tenant decides its bid and submits it to the operator through `Bid`$(b, \cdots)$. Then, the operator sets the price r using `SetReward`$(r, \cdots)$ and announces it to tenants. Note that, to guide the outcome towards equilibrium, the operator can tell tenants its anticipated price range $[r_{\min}, r_{\max}]$, such that tenants can set bids to avoid a profit loss by only considering this restricted price range instead of all possible prices.

• **Reducing power demand.** After the execution of the market mechanism, each participating tenant i cuts its power by $s_i(b_i, r)$. It is at each tenant's discretion to decide the actual power reduction techniques, using a combination of resource management APIs illustrated in Fig. 4 and/or its existing built-in power capping solutions [5, 6, 11]. Note that each knob has a different settling time for power reduction (e.g., DVFS is faster than load migration) and, depending on how much the power demand exceeds the capacity, the operator can also specify a timing constraint to guide tenants' selection of power reduction techniques.

• **Resuming normal operation.** When the tenants' aggregate power demand *without* power capping becomes lower than the capacity for a duration exceeding threshold T_c, the operator signals tenants to resume their normal operation using `Resume`$(\cdots)$. Tenants are compensated based on the power capping duration and price r.

3.4 Applicability of COOP

COOP applies to tenants who are interested in exchanging a temporary performance loss (due to power reduction) for financial compensation. It does not target tenants that have no tolerance on temporary performance loss (e.g., those running highly mission-critical workloads). These tenants will be served as premium clients on separated infrastructure *without* oversubscription.

In practice, a large portion of the operator's revenue (over 50%) comes from tenants running non-mission-critical workloads (e.g., R&D, lab computing, internal services, and recently, Bitcoin) that exhibit a great scheduling flexibility for temporarily reducing power [26]. Tenants also typically provision their servers based on the peak need, thus often having a slackness for reducing server power [27, 28]. Further, increasingly mature power capping techniques [6] and emerging techniques (e.g., approximate computing [29] that trades service quality for resource/power saving), have been constantly lowering the barrier for using COOP.

As shown in Table 1, the operator can offer more than $20/hour for each kW reduction, which is nearly 200 times of the market electricity price. If all tenants choose to neglect the operator's rewards and an unplanned downtime occurred, tenants would experience a costly business interruption but receive much less reward (around $3 per kW for each hour of downtime [20, 30]). Thus, it is also in the tenants' own interest to reduce power for handling emergencies.

In practice, tenants have no knowledge of whom they are sharing the PDU with. Further, if some tenants' power exceeds their own capacities, they will be penalized and may face involuntary power cut. Thus, in practice, it is very difficult and risky for (some) tenants to collude and create an artificial power emergency for rewards.

Finally, whenever tenants' aggregate power demand is not capped below the capacity by using COOP for any reasons (e.g., communication failure, or insufficient financial compensation for incentiving enough power reduction), the operator may resort to other complementary power capping techniques, e.g., discharging ESD [4] that avoids power capacity overloading (albeit not applicable for handling cooling capacity overloading) [22]. In any event, using COOP will *not* increase the risk of outages compared to the case in which COOP is not used.

Combining all these factors, we have a good reason to believe that COOP is appealing for reducing risks of outages when power emergencies arise in a multi-tenant data center.

3.5 Comparison with Other Market Designs

Conceptually, our formulation in (1) can be viewed as a multi-resource allocation problem where the resources are "power reduction D_i for $i = 0, 1, \cdots, M$" [31, 32]. It is challenging because: first, "resources" in our context are interdependent (e.g., high-level power capacity overlaps with low-level capacity), whereas the resources to allocate are mostly orthogonal in prior research (e.g., CPU and memory in clusters [31]); and second, tenants have private cost information $c_i(s_i)$ and manage their own servers without being controlled

Table 2: Testbed configuration.

Tenant	Type	No. of Servers	Tenant's Max. Power	Location	Cluster's Max. Power
#1	Web search	2	200 W	Cluster#A	740 W
#2	KVS	2	310 W		
#3	Hadoop	2	230 W		
#4	Web search	3	300 W	Cluster#B	530 W
#5	Hadoop	2	230 W		

by the operator.

While there are market-based studies (e.g., Nash bargaining) for multi-resource allocation [31, 32], their focus is on encouraging resource sharing (for improving utilization) and balancing efficiency versus fairness, whereas we aim at minimizing tenants' performance cost using a different mechanism — supply function bidding.

Market-based power management in (multi-tenant) data centers has recently received attention but differs from our work in problem formulation (due to our multiple *interdependent* power capping constraints) [32–39]. Further, most of the prior studies have considered pricing-based or Vickrey-Clarke-Groves (VCG)-based mechanisms, which are not suitable for our problem due to the following limitations.

Pricing-based mechanisms. Under a pricing-based mechanism, the operator offers a reward (also called"price") to incentivize tenants' power reduction [34, 35, 38]. The challenge of such designs is the determination of the price. In order to properly set prices such that tenants reduce a desired amount of power, the operator needs to know *a priori* how much power tenants would reduce in response to the offered price, and prior literature [34] has shown that inaccurate prediction can lead to undesired outcomes (e.g., power capping violation in our context). Further, power emergencies often occur unexpectedly and thus, estimating tenants' responses is inherently highly noisy during such periods.

VCG-based mechanisms. Another commonly-studied approach to solving (1) is the VCG auction mechanism [40, 41], i.e., the data center operator treats the power reduction quota as a *resource* and auctions it to tenants. Such designs require that tenants submit complex bids disclosing their full cost functions $c_i(\cdot)$, which are private information. Further, under such designs the payments made to tenants may be unbounded and reward rates for different tenants' power reduction are significantly different (creating unfairness issues). Thus, VCG auction mechanisms are rarely used in real large-scale systems (see [13] for a longer discussion).

Supply function mechanisms. In contrast, COOP adapts a variant of supply function bidding widely used in power markets that, besides its cost efficiency [13], has compelling advantages. First, through a supply function, the operator *proactively* solicits information from tenants as to how much power they would like to reduce if offered a certain price, while such information needs to be predicted by pricing-based mechanisms [34, 35]. Second, it uses the parameterized supply function as a proxy, thus avoiding tenants' disclosure of their private cost functions. Finally, it allows easy communication of the supply function through a single bidding parameter b_i from each tenant.

4. EVALUATION METHODOLOGY

We now describe our methodology for evaluating the efficiency of COOP in realistic scenarios. We first describe our prototype for a multi-tenant data center, and then formalize tenants' cost and performance models.

Following prior power capping research [5, 6], we build a scale-down testbed with two clusters (labelled as #A and #B, with six and five Dell PowerEdge R720 servers, respectively) in view of the practical difficulty in accessing commercial systems. The servers each have one 6-core Intel Xeon E-2620 Processor and 32GB memory. They are virtualized to create multiple nodes. All servers are powered through CloudPOWER meters to measure power at runtime. Our testbed configuration is presented in Table 2, which has five tenants on the two clusters: two tenants (#1 and #4) process web search workloads, another two (#3 and #5) process Hadoop jobs and the remaining tenant (#2) processes key-value store (KVS) workloads.

According to tenants' maximum power, the total subscribed power at Cluster#A is 740W and at Cluster#B is 530W, which we use as a baseline to determine the cluster-level power oversubscription. For example, if the capacity of Cluster#A is 672W, then 740W power subscription represents a 10% oversubscription. As illustrated in Fig. 4, we implement the APIs for the operator on a separate desktop server, and APIs for tenants as a separate process on their own servers.

4.1 Workloads

In the following, we describe our implementation of the web search, key-value store (KVS) and Hadoop workloads. While COOP is not restricted to these workloads, we choose our setting for two reasons: (1) it resembles the common setting in commercial data centers serving a diverse set of tenants, including CDN, web services and data analytics; and (2) our choice of workload is consistent with prior studies (e.g., [5]) that investigate power capping for owner-operated clusters (which can be viewed as "tenant" in our context).

Web search: We use web search benchmark from CloudSuite [42]. It benchmarks indexing process using Nutch search engine. We implement it for tenants #1 and #4. Tenant #1 has one Nutch front end and five index serving nodes, while tenant #4 has one Nutch front end and eight index serving nodes.

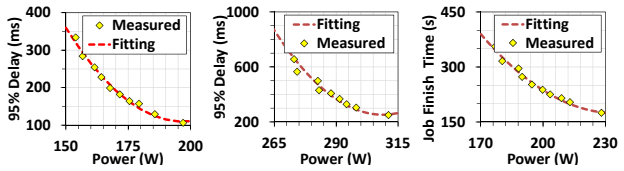
Key-value store (KVS): KVS resembles multi-tiered applications such as social networking. Tenant#2 has one load balancer VM, three Memcached VMs, three database VMs and nine application VMs.

Hadoop: Our Hadoop implementation for tenants #3 and #5 each consists of one master node and eleven worker nodes, using VMs hosted on two physical servers. We perform the *sort* benchmark on randomly generated files.

4.2 Performance and Cost Models

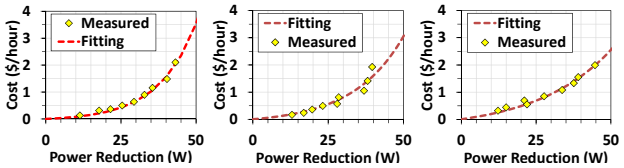
To participate in COOP, tenants need to employ power management (widely existing in today's systems [5, 27]) and evaluate their *costs* due to power reduction to decide bids.

Power and performance. Power reduction is normally accompanied by a performance degradation [6]. For the web search and KVS tenants, we use 95% response time as the performance metric (which is a key performance indicator for web services), while job completion time is used as the performance metric for the Hadoop tenants (due to the delay-tolerant nature). In our study, we consider that the tenants



(a) T#1- Web search (b) T#2- KVS (c) T#3- Hadoop

Figure 5: Power and performance models.



(a) T#1- Web search (b) T#2- KVS (c) T#3- Hadoop

Figure 6: Cost models.

reduce their power using dynamic voltage frequency scaling (DVFS) supported by most modern CPUs [43]. As a commonly-used knob for power capping [5, 6], DVFS enables almost instantaneous power reduction. The Intel Xeon CPUs in our testbed servers have 10 discrete DVFS levels with processing speeds ranging from 1.2GHz to 2.0Ghz.

We model the tenants' performance and power at different DVFS levels, and show the results for the three different types of workloads in Fig. 5. For the convenience of clarity, we set the same speed for all servers of a tenant, and only show results under a certain traffic load: tenant #1's delay performance is measured for 80 simultaneous search sessions, tenant #2's performance is measured for 30 requests/second. These are their maximum processing capacities under their subscribed power. For tenant #3 serving Hadoop, the file size is 3GB. Fig. 5 shows the non-linear relation between delay performance and power consumption, indicating a natural result that tenants suffer from a greater performance loss when they run their servers in lower power modes. We do not show tenants #4 and #5, which have similar configurations as tenant #1 and #3, respectively.

Cost model. In principle, tenants have the full discretion to decide their own cost models, considering one or more factors such as performance loss, risk attitude, among others. COOP applies to a large family of cost models in practice, although the theoretical efficiency guarantee only holds under a simplified setting with convex costs [13].

For evaluation purpose, we consider a cost model in terms of delay performance and model the performance cost for web search and KVS tenants using a piece-wise cost function adopted by [44] as follows:

$$c_{tenant} = \begin{cases} a \cdot d, & \text{if } d \le d_{th}, \\ a \cdot d + b \cdot (d - d_{th})^2, & \text{if } d > d_{th}, \end{cases} \quad (2)$$

where c_{tenant} is cost per job, a and b are tenants' own modeling parameters, d is 95% delay of interest and d_{th} is the delay threshold below which the performance cost only increases linearly (since end users can barely perceive the delay increase if it is already small). When the delay exceeds the threshold, however, performance cost will increase quadratically to account for degradation in user experiences.

For the Hadoop tenants, we use a linear cost model that increases with job completion time $c_{tenant} = \rho \cdot T_{job}$, where ρ is a modeling parameter and T_{job} is the job completion time of the Hadoop system.

Using the above cost models, we determine tenants' costs corresponding to different levels of power reduction (by setting $d_{th} = 100ms$ for web search and $d_{th} = 300ms$ for KVS). Fig. 6 shows the cost of power reduction of the tenants, under the same traffic setting as in Fig. 5. We subtract the tenants' original costs (without power reduction) from their cost models to ensure "zero cost" for zero power reduction. Setting cost model parameters is the task of individual tenants.

For *evaluation* purpose, we set the cost parameter such that the tenants' cost for power reduction is comparable to the extra revenue the data center operator gets from oversubscribing the capacity. Cost function is tenant's private information, and COOP uses supply function as a proxy to avoid the disclosure of tenant's cost information.

While the cost values can be arbitrarily set by tenants, our choice in this evaluation is logical: if there are mission-critical tenants which have a very high cost of power reduction, the operator will offer these tenants a premium service and not oversubscribe the capacity serving them.

Importantly, our results are not particularly sensitive to the details of the cost model described above, provided that costs are not arbitrarily high (otherwise, those tenants are considered as "premium" and served without oversubscription). We highlight this in Section 5 by varying the cost models.

4.3 Capacity Overloading

We apply COOP to handle a two-level power emergency involving five tenants in two low-level clusters sharing one high-level UPS, for the following levels of oversubscriptions.

• *Aggressive.* Cluster#A capacity is 643W and Cluster#B capacity is 460W (15% oversubscription), while the high-level capacity is 1050W (5% oversubscription, i.e., $1050 * 1.05 = 643 + 460$).

• *Moderate.* Cluster#A capacity is 672W and Cluster#B capacity is 481W (10% oversubscription), while the high-level capacity is 1098W (5% oversubscription, i.e., $1098 * 1.05 = 672 + 481$).

• *Conservative.* Cluster#A capacity is 704W and Cluster#B capacity is 504W (5% oversubscription), while the high-level capacity is 1150W (5% oversubscription, i.e., $1150 * 1.05 = 672 + 481$).

Note that, the three oversubscription cases described above are equivalent to a combined oversubscription at the high level of approximately 20%, 15% and 10%, respectively. We consider this combined 20% oversubscription as an "aggressive" strategy for two reasons. First, real-world data center power measurement demonstrates that the average power demand is roughly 70-80% of the peak [9, 45]: if the operator oversubscribes the capacity by more than 20% (equivalently, provisioning a capacity less than 83% of the peak demand), then the provisioned capacity may be quite close to or even below the servers' average power demand. Second, as shown in Table 1, if oversubscription is too large and exceeds 20%, the probability of overloading also increases and hence the reward rate that can be offered to tenants without decreasing the operator's profit actually decreases.

Power emergency. We create a power emergency by in-

creasing tenants' traffic load simultaneously. The top envelope in Fig. 8(a) illustrates the capacity overloading event: at around the 130th second, there is a spike in aggregate power demand, which begins to decrease by itself at around the 300th second when we decrease tenants' traffic (due to the completion of Hadoop jobs).

5. EVALUATION RESULTS

In this section, we evaluate COOP on the testbed described above. By assessing the efficiency of COOP in terms of total performance cost, we show that COOP is very close to OPT. Moreover, we demonstrate that COOP provides economic benefits to both the data center operator (through extra profit) and tenants (by reducing leasing costs).

5.1 Baseline and Metric

Baseline. We use OPT as the baseline, an ideal case where the operator minimizes the performance cost formulated in (1) and then dictates tenants' power reduction accordingly as if in an owner-operated data center.

Except for COOP, we are not aware of any alternative market mechanisms applied to handle a multi-level power capping in a multi-tenant data center. Furthermore, as shown later, COOP is very close to OPT in terms of the total performance cost (our key efficiency metric detailed below). Thus, we do not compare COOP with other market mechanisms which have yet to be introduced to multi-tenant data centers.

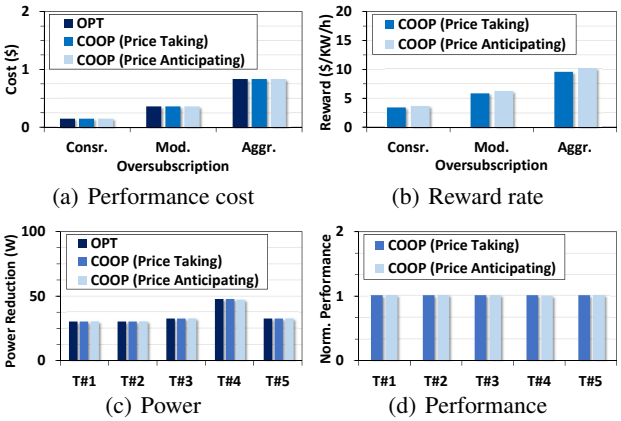
Metric. The key metric to assess COOP is *total performance cost* of the tenants, which, as formulated in (1) and quantified in monetary value, is a scalar measure of overall performance impact on tenants. We also evaluate the tenants' performance: 95-percentile delay for web search (tenant #1 and #4) and KVS (tenant #2), and throughput (job processing rate) for Hadoop tenants (#3 and #5).

Normalized performance. Tenants' power reduction results in performance degradation during an emergency [5,6]. Thus, we normalize tenants' performance under COOP with respect to that under OPT (our idealized baseline) to show how gracefully COOP can handle an emergency compared to OPT. Thus, the normalized performances are defined as: the ratio of OPT's 95% delay to COOP's 95% delay, and the ratio of COOP's throughput to OPT's throughput.

Tenants can be price-taking or price-anticipating. Price-taking means that tenants simply bid in a myopic way without predicting the impact of their bidding decisions on the market price. Price-anticipating means that tenants can predict how the operator sets price and more intelligently decide their bids to maximize their profits "$r \cdot s_i - c_i(s_i)$". See [13] for a detailed discussion of their different impacts on the equilibrium. For completeness, we show results for both cases under their respective equilibrium points, at which tenants maximize their own profits and have no incentives to deviate.

5.2 Efficiency

We first assess the efficiency of COOP in terms of the total performance cost. The results are shown in Fig. 7(a), where the absolute values are small due to the scale of our testbed. Under all the considered oversubscription levels, COOP is close to OPT, both when tenants are price-taking



(a) Performance cost (b) Reward rate

(c) Power (d) Performance

Figure 7: Comparison of different algorithms.

and when they are price-anticipating, which demonstrates that COOP is efficient in minimizing performance cost in practical settings that extend the theoretical study [13].

Fig. 7(b) shows the price/reward ($/kW/Hour) paid to tenants. There is no price in OPT, because it assumes the operator's full control over tenants' servers as in an owner-operated data center. As expected, when tenants are more "clever", i.e., price-anticipating, they explicitly predict the way to set market price and then bid accordingly, thereby driving up the price.

Next, with a moderate oversubscription, we show in Fig. 7(c) the breakdown of tenants' power reduction. Under both COOP and OPT, tenants' power reductions are almost identical, further confirming that COOP is close to OPT. Fig. 7(d) shows COOP's performance normalized with respect to OPT's performance: COOP is almost identical to OPT in terms of the performance impact on tenants.

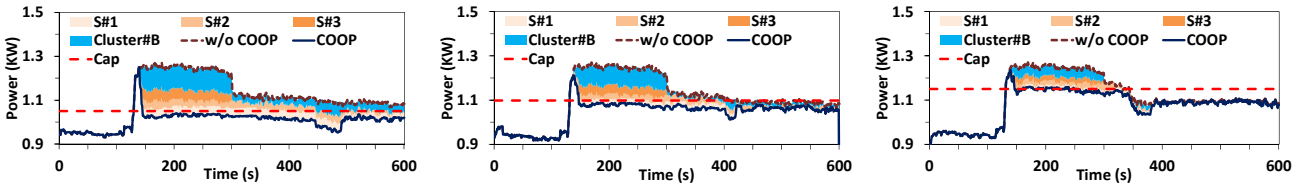
Settling time. There is a time lag, i.e., settling time, between the detection of power emergency and tenants' actual power reduction. First, the supply function bidding mechanism in COOP needs to be executed as described in Section 3.2. Each tenant needs to calculate its bidding parameter b_i based on its current traffic, which takes less than 50ms per computation and is performed in parallel; the operator clears the market price according to $r = \min_{r'}\{r' \in [r_{\min}, r_{\max}] \,|\, \sum_{i \in \mathcal{N}_j} s_i(b_i, r') \geq D_j$, for $j = 0, 1, \cdots, M\}$, taking very little time. The messaging delay between the operator and tenants is in the order of tens of milliseconds, as only the supply function and bid/price parameters need to be communicated and the number of involved tenants is typically small (a few tens). Thus, the total time for executing COOP is less than 0.5 second.

The next step is for tenants to reduce power as decided by COOP. Here, we use DVFS as it is a widely-adopted technique and can switch between different speeds very quickly to cut enough power (even for 20% oversubscription).

The overall settling time for COOP in our study is less than one second, which is quickly enough to handle a power emergency and consistent with recent power capping studies for owner-operated data centers [5].

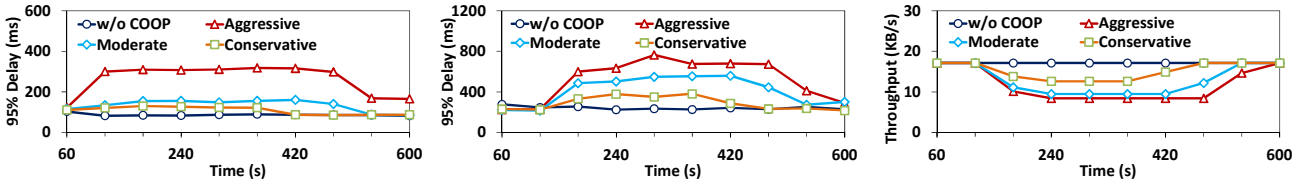
5.3 Execution

Fig. 7 shows that the total performance cost and tenants' power reduction are very similar, under both COOP and OPT. Thus, we only show the results of COOP (with price-



(a) Agressive oversubscription (b) Moderate oversubscription (c) Conservative oversubscription

Figure 8: Power traces under different oversubscription configurations.



(a) Tenant #1 (Web search) (b) Tenant #2 (Key-value store) (c) Tenant #3 (Hadoop)

Figure 9: Delay performance traces of the tenants under different oversubscription levels.

anticipating tenants) below.

5.3.1 *Power demand*

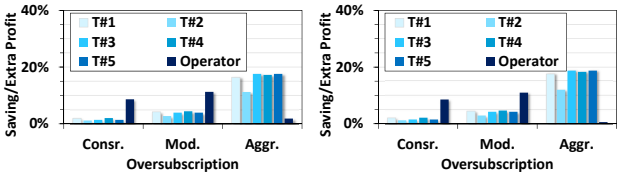
Aggressive oversubscription. Fig. 8(a) shows the power trace for aggressive oversubscription. The top envelop represents the tenants' aggregate power demand without any power reduction, while the bottom envelop is the reduced power demand when applying COOP. The shaded areas represent the individual contributions in power reduction. We combine the contribution from tenants #4 and #5 connected to Cluster#B as a whole for better clarity. We set the timer for initiating power capping as $T_w = 15$s. After power capping is applied at around time 145s, the aggregate power demand goes below (but close to) the provisioned capacity. Then, at around time 490s, there is a change in the aggregate power demand (lower envelope), because tenant #5 finishes its job and COOP is re-applied to decide power reductions for participating tenants.

Moderate and conservative oversubscription. Fig. 8(b) and Fig. 8(c) show the power traces under moderate and conservative oversubscription, respectively. We make similar observations as in Fig. 8(a), except for that tenants can resume normal operation sooner (since Hadoop tenants finish jobs sooner with a less aggressive oversubscription).

5.3.2 *Performance*

We show Cluster#A tenants' performance measured over a 60-second window in Fig. 9 to save space (which include all the three considered workload types) while tenants in Cluster#B also have similar results. Fig. 9(a) and 9(b) show the 95% delay performance of tenant #1 and tenant #2, respectively. Fig. 9(c) shows the Hadoop tenant's performance (measured in the throughput, which is the inverse of job completion time given a fixed file size). As expected, we see the worst performance when the capacity is most aggressively oversubscribed (15% at the low level and 5% at the high level in our study).

While performance degradation is often unavoidable to handle power emergencies [5, 6], by using COOP, tenants' performance loss is minimum, as compared to OPT in terms of total performance cost and shown in Fig. 7(a).



(a) Price taking (b) Price anticipating

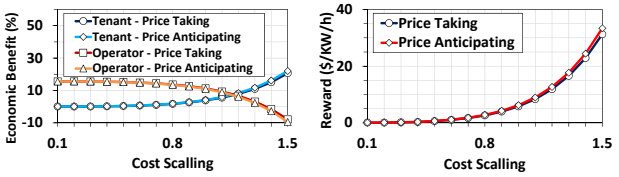
Figure 10: Economic benefit.

5.4 Economic Benefit

Fig. 10 shows economic benefits under different oversubscription levels: tenants save leasing cost through financial compensation for temporary power reduction, while the operator earns extra profit through oversubscription. Tenants' total reward is determined based on the reward rate and the probability of capacity overloading over a year (based on Fig. 2). Tenants' cost saving is calculated as the ratio of their total rewards to their total leasing costs based on the average market price of 150$/kW/month. We exclude tenants' performance *cost* which is a quantitative measure of tenants' performance consideration, and this is also the standard practice when assessing the cost saving benefit [4, 5]. The data center operator's extra profit is determined by subtracting the total payment to tenants from its additional revenue due to oversubscription. Fig. 10(a) and Fig. 10(b) show the economic benefits when tenants are price-taking and price-anticipating, respectively. In both cases, we see that tenants' cost saving goes up, as the level of oversubscription is increased. However, the operator has the highest extra profit under moderate oversubscription, because with aggressive oversubscription (20% combined oversubscription), the operator needs to pay a high price due to tenants' increasing reluctance in cutting more power (Fig. 6).

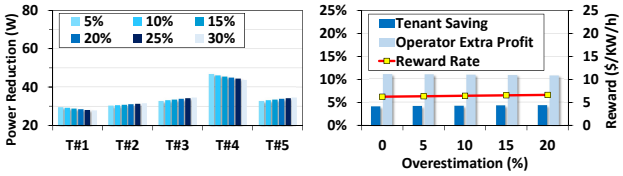
5.5 Tenant Costs

Tenant cost functions play a vital role in bidding decisions and hence the outcome of COOP. To illustrate the sensitivity of COOP to tenant cost functions we consider settings with costs scaled by a factor ranging from 0.1 to 1.5, and show the result under moderate oversubscription in Fig. 11.



(a) Economic Benefit (b) Reward rate

Figure 11: Impact of tenants' cost.



(a) Power Reduction (b) Price Taking

Figure 12: Impact of tenant cost overestimation.

We see that regardless of price-taking and price-anticipating behaviors, tenants' saving, averaged over the three tenants, increases with their scaling of performance cost, while the operator's extra profit goes down and even becomes negative when the scaling factor is more than 1.3. Fig. 11(b) shows the corresponding reward rates, which are going up as tenants' cost increases. This confirms that to earn extra profit through oversubscription, the operator should target those tenants that do *not* run highly mission-critical workloads and have a low cost for power reduction. We have also evaluated other cost models, and similar results hold.

5.6 Bidder Uncertainty

A main task for tenants in COOP is determination of the bidding strategy. One might expect that tenants have some uncertainty in this regard, and that this uncertainty, combined with risk aversion, may lead tenants to overestimate their costs when submitting bids. To illustrate the impact of this, we consider a setting where the web search tenants (#1 and #4) overestimate their costs by up to 30%. We see from Fig. 12(a) that power reduction decreases for the two web-search tenants, while the other tenants' power reduction increases to meet power capping constraints. However, the impact is not significant. As shown in Fig. 12(b), cost overestimation slightly drives up the reward rate and has a very little impact on savings (for both operator and tenants). This is because the impact of tenants with overestimated costs is mitigated by the other tenants. Similar results hold for price-anticipating tenants. If tenants bid arbitrarily high for any reason, they will be excluded from COOP (equivalent to *premium* tenants served without oversubscription) and lose cost saving benefits provided by COOP. In fact, it is at tenants' interests to bid reasonably (as discussed in Section 3.2) to reach an equilibrium, at which all participating tenants maximize their own net profits.

We also run a larger-scale simulation to evaluate COOP with more tenants. Our simulation shows that COOP still applies and mutually benefits the data center operator and participating tenants. These results are omitted for brevity.

6. RELATED WORK

There is a large and rich literature on power capping in owner-operated data centers. Various techniques have been proposed for minimizing performance loss, such as reducing CPU power [6, 46], admission control [47], virtualizing power allocation [5, 11], and load migration [5, 47]. These can be leveraged as power capping techniques by *individual tenants*, but they are not applicable for handling emergencies resulting from operator's oversubscription due to lack of control over tenants' servers. Recent studies [4, 10, 48] have explored discharging ESD (e.g., battery) to temporarily boost power supply for handling a emergency. These techniques can be viewed as "*supply*-side" solutions and are complementary to our "*demand*-side" power reduction. Further, discharging ESD might still overload the cooling capacity, which, typically sized based on the IT power, may increase overheating risk that is a major reason for downtimes [8, 23]. Recent work [49] proposes to place phase changing materials inside servers to avoid cooling capacity overloading, but tenants' servers may not have such advanced materials. COOP still works if cooling capacity is over-provisioned and/or phase changing materials are available, and in such cases, these techniques can be combined with COOP to enable more power oversubscription.

Our research is relevant to multi-resource allocation [31, 32] and data center demand response (broadly interpreted as reshaping the power demand towards a desired goal) [33–39, 50]. In addition to problem differences, our formulation and proposed mechanism are also different from those prior studies. Specifically, the prior studies on data center demand response [34–39,50] have all been focused on cutting power on a best-effort basis at the data center level, whereas we propose supply function bidding to address *multi-level* power capping. A detailed comparison is provided in Section 3.5.

7. CONCLUSION

This paper proposes COOP, a market-based approach for incentivizing and coordinating tenants' power reductions in the event of a power emergency in a multi-tenant data center. COOP uses a supply function bidding mechanism motivated by literature in electricity markets. We demonstrate the effectiveness of COOP by building a prototype and illustrating that COOP is efficient in minimizing the total performance cost, even compared to the ideal case OPT. We also demonstrate that COOP is "win-win", increasing the data center operator's profit and reducing tenants' cost by providing financial compensation for power reductions.

8. APPENDIX

To facilitate readers' understanding, we briefly explain the equilibrium point in COOP.

An equilibrium, denoted by $\mathbf{b}^* = (b_1^*, b_2^*, \cdots, b_N^*)$, represents a stabilized outcome at which all participating tenants maximize their own payoffs (i.e., reward minus performance cost). The resulting equilibrium depends on whether tenants are *price-taking* or *price-anticipating*. Given the parameterized supply function $s_i(b_i, r) = \left[\delta_i - \frac{b_i}{r}\right]^+$ and the price r set according to Line 16 in Algorithm 1, a price-taking tenant i decides its bid b_i to maximize its profit $u_i^t(b_i; r) = r \cdot s_i(b_i, r) - c_i(s_i(b_i, r))$ by passively accepting the offered market price r, while a price-anticipating tenant i explicitly

predicts the price $r(\mathbf{b})$ as a function of all the submitted bids and decides b_i to maximize $u_i^a(b_i;\mathbf{b}_{-i}) = r(\mathbf{b}) \cdot s_i(b_i, r(\mathbf{b})) - c_i(s_i(b_i, r(\mathbf{b})))$, where $\mathbf{b}_{-i} = (b_1, \cdots, b_{i-1}, b_{i+1}, \cdots, b_N)$ is the bidding profile except b_i, for $i = 1, 2, \cdots, N$. The price-taking scenario normally applies when tenants all have similar sizes and no one can impact the market price too much, whereas the price-anticipating model is suitable when there exist a few dominant tenants. With $u_i^t(b_i;r)$ and $u_i^a(b_i;\mathbf{b}_{-i})$ defined above, we next provide the definition of equilibrium.

Definition. A bidding profile $\mathbf{b}^* = (b_1^*, b_2^*, \cdots, b_N^*)$ is a price-taking equilibrium if it satisfies $u_i^t(b_i^*;r) \geq u_i^t(b_i;r)$, $\forall b_i \geq 0$ and $i = 1, 2, \cdots, N$, and a price-anticipating equilibrium if it satisfies $u_i^a(b_i^*;\mathbf{b}_{-i}^*) \geq u_i^a(b_i;\mathbf{b}_{-i}^*)$, $\forall b_i \geq 0$ and $i = 1, 2, \cdots, N$. ■

The choice of supply function affects the total performance cost efficiency at an equilibrium. Under a set of simplifying assumptions, prior studies [13,14] have proved that choosing the supply function as $s_i(b_i, r) = \delta_i - \frac{b_i}{r}$ results in an efficient equilibrium with a bounded deviation from the optimum. We use a modified supply function $s_i(b_i, r) = \left[\delta_i - \frac{b_i}{r}\right]^+$ to avoid *negative* power reduction and demonstrate its efficiency for multi-level power capping via experiments.

9. REFERENCES

[1] A. Venkatraman, "Global census shows datacentre power demand grew 63% in 2012," in *ComputerWeekly.com*, 2012.

[2] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel, "The cost of a cloud: Research problems in data center networks," *SIGCOMM Comput. Commun. Rev.*, vol. 39, Dec. 2008.

[3] L. A. Barroso, J. Clidaras, and U. Hoelzle, *The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines*. Morgan & Claypool, 2013.

[4] D. Wang, C. Ren, A. Sivasubramaniam, B. Urgaonkar, and H. Fathy, "Energy storage in datacenters: what, where, and how much?," in *SIGMETRICS*, 2012.

[5] D. Wang, C. Ren, and A. Sivasubramaniam, "Virtualizing power distribution in datacenters," in *ISCA*, 2013.

[6] X. Fu, X. Wang, and C. Lefurgy, "How much power oversubscription is safe and allowed in data centers," in *ICAC*, 2011.

[7] Ponemon Institute, "2013 cost of data center outages," 2013, `http://goo.gl/6mBFTV`.

[8] Emerson Network Power, "Addressing the leading root causes of downtime," 2013, `http://goo.gl/b14XaF`.

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

[10] D. Wang, S. Govindan, A. Sivasubramaniam, A. Kansal, J. Liu, and B. Khessib, "Underprovisioning backup power infrastructure for datacenters," in *ASPLOS*, 2014.

[11] H. Lim, A. Kansal, and J. Liu, "Power budgeting for virtualized data centers," in *USENIX ATC*, 2011.

[12] Uptime Institute, "Data center industry survey," 2014.

[13] R. Johari and J. N. Tsitsiklis, "Parameterized supply function bidding: Equilibrium and efficiency," *Oper. Res.*, vol. 59, pp. 1079–1089, Sept. 2011.

[14] Y. Xu, N. Li, and S. H. Low, "Demand response with capacity constrained supply function bidding," *IEEE Transactions on Power Systems*, 2015.

[15] DatacenterMap, "Colocation USA," `http://www.datacentermap.com/usa/`.

[16] NRDC, "Scaling up energy efficiency across the data center industry: Evaluating key drivers and barriers," *Issue Paper*, Aug. 2014.

[17] Akamai, "Environmental sustainability policy," `http://www.akamai.com/html/sustainability/our_commitment.html`.

[18] Microsoft, "Global infrastructure," `http://www.globalfoundationservices.com/`.

[19] A. C. Riekstin, S. James, A. Kansal, J. Liu, and E. Peterson, "No more electrical infrastructure: Towards fuel cell powered data centers," *SIGOPS Oper. Syst. Rev.*, vol. 48, pp. 39–43, May 2014.

[20] CBRE, "Q4 2013: National data center market update," 2013.

[21] V. Kontorinis, L. E. Zhang, B. Aksanli, J. Sampson, H. Homayoun, E. Pettis, D. M. Tullsen, and T. S. Rosing, "Managing distributed ups energy for effective power capping in data centers," in *ISCA*, 2012.

[22] W. Zheng and X. Wang, "Data center sprinting: Enabling computational sprinting at the data center level," in *ICDCS*, 2015.

[23] I. Manousakis, I. n. Goiri, S. Sankar, T. D. Nguyen, and R. Bianchini, "Coolprovision: Underprovisioning datacenter cooling," in *SoCC*, 2015.

[24] Google, "Compute Engine Incident #15056," `https://status.cloud.google.com/incident/compute/15056`.

[25] S. Govindan, D. Wang, A. Sivasubramaniam, and B. Urgaonkar, "Leveraging stored energy for handling power emergencies in aggressively provisioned datacenters," in *ASPLOS*, 2012.

[26] J. dePreaux, "Wholesale and retail data centers - North America and Europe - 2013," *IHS*, Jul. 2013, `https://technology.ihs.com/api/binary/492570`.

[27] D. Lo, L. Cheng, R. Govindaraju, L. A. Barroso, and C. Kozyrakis, "Towards energy proportionality for large-scale latency-critical workloads," in *ISCA*, 2014.

[28] M. Lin, A. Wierman, L. L. H. Andrew, and E. Thereska, "Dynamic right-sizing for power-proportional data centers," in *IEEE Infocom*, 2011.

[29] I. Goiri, R. Bianchini, S. Nagarakatte, and T. D. Nguyen, "Approxhadoop: Bringing approximations to mapreduce frameworks," in *ASPLOS*, 2015.

[30] Internap, "Colocation services and SLA," `http://www.internap.com/internap/wp-content/uploads/2014/06/Attachment-3-Colocation-Services-SLA.pdf`.

[31] A. Ghodsi, M. Zaharia, B. Hindman, A. Konwinski, S. Shenker, and I. Stoica, "Dominant resource fairness: Fair allocation of multiple resource types," in *NSDI*, 2011.

[32] S. M. Zahedi and B. C. Lee, "Ref: Resource elasticity fairness with sharing incentives for multiprocessors," in *Proceedings of the 19th International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS, 2014.

[33] M. Guevara, B. Lubin, and B. C. Lee, "Navigating heterogeneous processors with market mechanisms," in *HPCA*, 2013.

[34] Z. Liu, I. Liu, S. Low, and A. Wierman, "Pricing data center demand response," in *SIGMETRICS*, 2014.

[35] C. Wang, N. Nasiriani, G. Kesidis, B. Urgaonkar, Q. Wang, L. Y. Chen, A. Gupta, and R. Birke, "Recouping energy costs from cloud tenants: Tenant demand response aware pricing design," in *e-Energy*, 2015.

[36] N. Chen, X. Ren, S. Ren, and A. Wierman, "Greening multi-tenant data center demand response," in *IFIP Performance*, 2015.

[37] L. Zhang, S. Ren, C. Wu, and Z. Li, "A truthful incentive mechanism for emergency demand response in colocation data centers," in *INFOCOM*, 2015.

[38] C. Wang, B. Urgaonkar, G. Kesidis, U. V. Shanbhag, and Q. Wang, "A case for virtualizing the electric utility in cloud data centers," in *HotCloud*, 2014.

[39] M. A. Islam, H. Mahmud, S. Ren, and X. Wang, "Paying to save: Reducing cost of colocation data center via rewards," in *HPCA*, 2015.

[40] T. Roughgarden, "Algorithmic game theory," *Commun. ACM*, vol. 53, pp. 78–86, July 2010.

[41] L. Zhang, S. Ren, C. Wu, and Z. Li, "A truthful incentive mechanism for emergency demand response in colocation data centers," in *INFOCOM*, 2015.

[42] "CloudSuite - The Search Benchmark," `http://parsa.epfl.ch/cloudsuite/`.

[43] J. R. Lorch and A. J. Smith, "Pace: A new approach to dynamic voltage scaling," *IEEE Trans. Computers*, vol. 53, pp. 856–869, July 2004.

[44] P. X. Gao, A. R. Curtis, B. Wong, and S. Keshav, "It's not easy being green," *SIGCOMM Comput. Commun. Rev.*, 2012.

[45] D. Wang, C. Ren, S. Govindan, A. Sivasubramaniam, B. Urgaonkar, A. Kansal, and K. Vaid, "Ace: Abstracting, characterizing and exploiting peaks and valleys in datacenter power consumption," in *SIGMETRICS*, 2013.

[46] X. Wang, M. Chen, C. Lefurgy, and T. W. Keller, "Ship: Scalable hierarchical power control for large-scale data centers," in *PACT*, 2009.

[47] A. A. Bhattacharya, D. Culler, A. Kansal, S. Govindan, and S. Sankar, "The need for speed and stability in data center power capping," in *IGCC*, 2012.

[48] L. Liu, C. Li, H. Sun, Y. Hu, J. Gu, T. Li, J. Xin, and N. Zheng, "Heb: Deploying and managing hybrid energy buffers for improving datacenter efficiency and economy," in *ISCA*, 2015.

[49] M. Skach, M. Arora, C.-H. Hsu, Q. Li, D. Tullsen, L. Tang, and J. Mars, "Thermal time shifting: Leveraging phase change materials to reduce cooling costs in warehouse-scale computers," in *ISCA*, 2015.

[50] S. Ren and M. A. Islam, "Colocation demand response: Why do I turn off my servers?," in *ICAC*, 2014.