Fair Online Power Capping for Emergency Handling in Multi-Tenant Cloud Data Centers

Qihang Sun, Shaolei Ren, and Chuan Wu

Abstract—In view of the high capital expense for scaling up power capacity to meet the escalating demand, maximizing the utilization of built capacity has become a top priority for multi-tenant data center operators, where many cloud providers house their physical servers. The traditional power provisioning guarantees a high availability, but is very costly and results in a significant capacity under-utilization. On the other hand, power oversubscription (*i.e.*, deploying more servers than what the capacity allows) improves utilization but offers no availability guarantees due to the necessity of power reduction to handle the resulting power emergencies. Given these limitations, we propose a novel hybrid power provisioning approach, called HyPP, which provides a combination of two different power availabilities to tenants: capacity with a very high availability (100% or nearly 100%), plus additional capacity with a medium availability that may be unavailable for *up to* a certain amount during each billing period. For HyPP, we design an online algorithm for the operator to coordinate tenants' power reduction at runtime when the tenants' aggregate power demand exceeds the power capacities. Our algorithm aims at achieving long-term fairness in tenants' power reduction (defined as the ratio of total actual power reduction by a tenant to its contracted reduction budget over a billing period). We analyze the theoretical performance of our online algorithm and derive a good competitive ratio in terms of fairness compared to the offline optimum. We also validate our algorithm through simulations under realistic settings.

Index Terms—Power management, online fairness, hybrid power provisioning, multi-tenant cloud data centers.

1 Introduction

With the emergence of a plethora of IT services, Internet of Things, and cloud computing, multi-tenant data center (also sometimes called "colocation") has been in an unprecedented high demand worldwide. Being a shared facility where multiple tenants house and manage their own physical servers, multi-tenant data centers provide a cost-effective and scalable data center solution to almost all industry sectors. In particular, cloud providers have the strongest demand for multi-tenant data centers to deploy their physical servers and quickly expand global cloud services, wherever it is impractical and uneconomical to build their own data centers (*i.e.*, setting up all data center infrastructures, including facility, cooling and power systems). For example, major cloud providers, including Amazon, Google, and Microsoft, have recently leased large capacities in multi-tenant data centers for service expansion [1], whereas Apple houses approximately 25% of its servers in multi-tenant data centers [2]. In addition, government agencies have also been consolidating different units (each viewed as a "tenant") into shared *multi-tenant* data centers for cost efficiency, as attested to by the recent U.S. Federal Data Center Consolidation Initiative [3].

- *Q. Sun is with the Department of Computer Science, The University of Hong Kong*
  E-mail: qhsun@cs.hku.hk
- *S. Ren is with the Department of Electrical and Computer Engineering, University of California, Riverside, USA*
  E-mail: sren@ece.ucr.edu
- *C. Wu is with the Department of Computer Science, The University of Hong Kong*
  E-mail: cwu@cs.hku.hk

While the demand is escalating, scaling up the multi-tenant data center power capacity (often measured in protected power delivered to servers) has become increasingly more challenging. The capital expense for building data center power infrastructure is very high and even exceeds 1.5 times of the total energy cost over a 15-year lifespan [4]. Additionally, local grid capacity and long time-to-market cycle (often several months or even years) are also limiting data center construction and/or expansion to meet the demand. For these reasons, maximizing utilization of the built power capacity has become a top priority for data centers [5], especially for multi-tenant data centers that already consume nearly as five times energy as Google-type data centers combined altogether [6].

Traditionally, multi-tenant data centers provide service level agreement (SLA) to tenants by leasing power capacity with a very high availability guarantee (99.9999% or even 100%) [7]. The actual provisioned data center infrastructure is sized to support the maximum total leased power capacity. While this achieves a high availability, it results in a significantly low utilization of data center power capacity (only 60% or even lower [4], [8]) at most times, failing to fully capitalize on the expensive infrastructure. The reason is that power consumption by different tenants rarely peaks simultaneously [4], [5].

More recently, multi-tenant data centers have been commonly *oversubscribing* the infrastructure by selling the power capacity to more tenants than what is allowed. This is equivalent to under-provisioning the infrastructure below the total leased power capacity, thus cutting the high capital expense. Nonetheless, a dangerous consequence is the occasional power emergency when tenants' aggregate power demand exceeds the provisioned capacity. Though

TABLE 1
Pros & Cons of Different Power Provisioning

Approach	Capacity Utilization	Availability SLA	Emergency Handling Capability
Traditional	Low	✓	N/A
Oversubscription	Medium	✗	✗
HyPP	High	✓	✓

rare, power emergency drastically compromises data center availability and, if not properly handled, can even result in costly outage incidents (an average of $901,560 per incident [9]), damaging the operator's reputation and causing a high churn rate.

Recent research has proposed incentive mechanisms to coordinate tenants' load shedding during an emergency [5], [10]. They are purely *best-effort* designs to handle emergencies by relying on tenants' voluntary power reduction, thus providing no assurance to the operator that enough power will be cut [5], [10]. Even assuming that tenants will contribute, no (worst-case) SLA is guaranteed: tenants may be asked to cut power very frequently and/or by a large amount, causing an unacceptable degradation in their workload performance. Consequently, these severely limit the oversubscription level for improving utilization and raise concerns with the applicability of the proposed solutions in practice [5], [10].

In view of the limitations (summarized in Table 1), we propose a novel contract-based *hybrid* power provisioning, called HyPP, which oversubscribes the infrastructure to improve utilization and contracts enough power reduction to handle power emergencies. HyPP focuses on the operator side for improving the infrastructure utilization, orthogonal to tenants' performance, differing from Amazon spot instance model which focuses on the tenant's side and concerns tenant's utility. Specifically, HyPP provisions two types of power capacities to tenants: high-availability capacity (100% or nearly 100% availability [7]), plus medium-availability capacity that has a lower SLA and may be unavailable for *up to* a certain amount during each billing period. As HyPP still operates the power under the designed capacity, some indices shown in traditional data centers, *e.g.*, SLA, and reliability, are not compromised.

Many power management algorithms and interfaces (e.g., CPU frequency scaling [4], [11] and Intel Rack Scale Design [12]) are readily available to scale down tenants' power consumption when medium-availability power capacity is cut. Thus, by leasing medium-availability capacity on top of the guaranteed capacity, tenants can lower their power subscription costs yet have an SLA assurance. In fact, hybrid SLA has also been quickly emerging in other contexts to meet the diverse needs of users (e.g., Google Cloud offers both low-latency online storage for "hot" data and medium-latency *nearline* storage for "cold" data [13]). Nonetheless, HyPP focuses on multi-tenant data centers and provides a novel hybrid SLA to tenants in terms of power provisioning, which has remained under-explored in the literature.

While HyPP is appealing, turning it into practice creates multifaceted challenges. In particular, when tenants' aggregate power occasionally exceeds the capacity (i.e., emergency), which tenants' power provisioning should be capped and by how much? Additionally, power capping can degrade tenants' performance, and hence another central question is how to achieve fairness in power capping for different tenants? Last but not least, power capping decisions must be made online without knowing future power emergencies, subject to the power provisioning SLA which requires that the total amount of unavailable power for a tenant be below a certain threshold over a billing period.

To address the above challenges, we propose an online fair power capping algorithm that judiciously makes power reduction decisions whenever a power oversubscription emergency occurs, maximizing the fairness of power reduction among the tenants over a billing period and satisfying the power provisioning SLA (with a bounded violation even in the worst case). We novelly design the online algorithm based on a primal-dual framework and dual fitting technique. We rigorously analyze the competitive ratio of the algorithm, as compared with the offline optimum (*i.e.*, the best solution when knowing all future information). We also run trace-based simulations under realistic system settings to validate our algorithm. We show that under realistic settings, our algorithm is always able to handle power emergencies subject to SLA, outperforms baselines, and is near the optimum in terms of fairness.

2 Preliminaries and Problem Formulation

2.1 Power Architecture in Data Centers

Multi-tenant cloud data centers typically employ a tree-type power hierarchy. As illustrated in Fig. 1, high-voltage grid power first enters the data center through an automatic transfer switch (ATS), which will switch to backup generators during grid failures. Then, power passes the uninterrupted power supply (UPS) system. The UPS-protected power, also called IT critical power, is fed to multiple power distribution units (PDUs), which each have a capacity of 200-300kW and output power at suitable levels to support server racks. Finally, the rack-level power strip (also called rack-level PDU) directly supplies power to the servers.

There are power capacity constraints throughout the hierarchy: UPS, PDU, and rack levels. The UPS and PDUs are very expensive and hence often aggressively oversubscribed to cut the capital expense [14], [15]. While an emergency may not instantly lead to an outage due to system redundancy [11], ignoring it can significantly increase downtime risks. Thus, the data center operator must keep the tenants' aggregate power consumption below the PDU and UPS capacities at all times [4], [14].[1]

In a multi-tenant data center, each tenant typically manages multiple colocated server racks. These racks, however, may be physically connected through underfloor cabling to different PDUs shared with other tenants to exploit multiplexing effects (*i.e.*, the power consumption of different tenants rarely peaks simultaneously). In fact, some emerging design has even enabled dynamically connecting servers/racks to PDUs to improve capacity utilization [14].

1. A tenant may also oversubscribe its own contracted rack-level capacity, but all the induced "emergencies" (exceeding the contracted capacity) must be taken care of by itself to avoid additional charges by the operator.

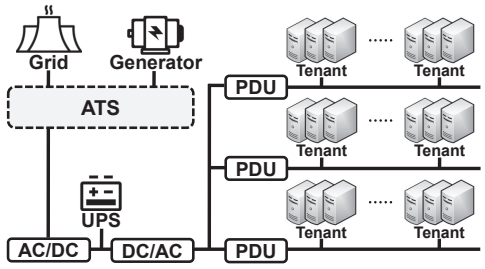


Fig. 1. Power Architecture in a Multi-tenant Data Center

2.2 Model Basics

Consider a multi-tenant data center with one (centralized) UPS that supports M PDUs. We use index 0 to refer to the UPS and indices $1, 2, \ldots, M$ to denote the PDUs. The UPS capacity is denoted by $\mathcal{C}_0$, while the j-th PDU capacity is $\mathcal{C}_j$ for $j = 1, \cdots, M$. There are N tenants in the multi-tenant data center. We consider a time-slotted model, where each time slot can be 5 minutes (duration of a typical power emergency event [5], [15]) and T time slots constitute a billing cycle (e.g., one month). For notational convenience, we use $[X]$ to denote the set of $[1, 2, \ldots, X]$, where X can be different for different sets (e.g., $[N]$ for the set of N tenants).

2.3 Hybrid Power Provisioning

In HyPP, tenants can subscribe two types of capacities through a purchasing contract: *high*-availability capacity (high cost but guaranteed with a 100%, or nearly 100%, availability, as traditionally provided [7]), and *medium*-availability capacity (lower cost but may be cut subject to a SLA defined later). More formally, HyPP specifies for each tenant $i \in [N]$ the following three values: (1) C_i^g, the amount of guaranteed high-availability capacity which cannot be compromised at any time (otherwise, the operator is entailed to financially compensate affected tenants [7]); (2) C_i^f, the amount of flexible medium-availability capacity which may be cut subject to an SLA; (3) B_i, the SLA parameter defined as the maximum amount of accumulated power reduction (also referred to as *power reduction budget*) that can be imposed on tenant i throughout a billing period. For example, if a tenant's medium-availability capacity can be cut by at most 30% on average during all events, then the monthly budget of a tenant subscribing 20kW medium-availability capacity can be set as $20(\text{kW}) \times 24(\text{hours}) \times 30(\text{days}) \times 3\% \times 30\% = 129.6(\text{kWh/month})$, supposing power emergency events occur in about 3% of all the time slots. Note that the power capacities of both C_i^g and C_i^f represent tenant i's total subscription and are evenly split across tenant i's server racks due to the rack-level capacity constraint in the data center power hierarchy [4], [11]. For example, if tenant i owns k racks in total in the data center, its per-rack high-availability capacity and medium-availability capacity are $\frac{C_i^g}{k}$ and $\frac{C_i^f}{k}$, respectively. The average capacity split is practical, deploying the racks with same power densities together, for the ease of heating management. Although additional constraints (*e.g.*, accumulated power reduction per rack) can be incorporated; we leave them out of our investigation to allow more flexibilities to both tenants and the operator.

HyPP is a "win-win" solution. On the one hand, the data center operator can safely oversubscribe the power capacity by contracting enough reduction of medium-availability power to handle emergencies. On the other hand, medium-availability power capacity is a type of transient power supply (but with SLA), and tenants can use numerous techniques to dynamically modulate power consumption to follow *transient* power supplies [16], [17]. Thus, by participating in HyPP, a tenant has a low-cost option for power subscription yet enjoys an SLA guarantee in terms of the maximum accumulated power unavailability specified by B_i per billing period. During runtime, for each tenant, only the power of the medium-availability capacity might be reduced, and the power of the high-availability capacity always be guaranteed. During the power reduction, if tenants are slow to the reduction request in mild violation, the system redundancy can handle. If the tenants are completely unresponsive to the requests, the operator can also cut their power directly for preventing outage of the entire system.

The data center operator sizes its PDU/UPS to ensure that the total high-availability power sold to tenants does not exceed the UPS capacity (*i.e.*, $\sum_{i\in[N]} C_i^g \leq \mathcal{C}_0$) and that the total high-availability power provisioned to the racks connected to each PDU is also below the respective PDU capacity. Thus, the data center operator can sell guaranteed capacity to tenants as usual (e.g., at a market price of US$150-200/kW per month [18]).

Meanwhile, the operator sells medium-availability power capacity using oversubscription. The specifications of C_i^f and B_i as well as pricing are determined through a business process to ensure that the resulting power emergencies, if any, can be handled via contracted power reduction by tenants subject to SLA. This process depends on several factors, such as PDU-/UPS-level power usage statistics, how aggressively the operator oversubscribes the infrastructure, and tenant i's energy agility (measuring how well servers' power consumption follows the transient power supplies, *e.g.*, by workload shifting [16]). In general, with a more aggressive oversubscription, the operator would sell more medium-availability power capacity to tenants. The SLA parameter B_i specifies the total power reduction budget and can be related to the frequency of emergencies according to PDU-/UPS-level power usage statistics. Given a fixed medium-availability power subscription C_i^f, a larger B_i means a worse SLA and possibly a cheaper price. It warrants a separate economic analysis to optimally specify C_i^g, C_i^f and B_i. In this paper, we view these values as orthogonal and focus on the data center operation at runtime: *how to fairly decide tenants' power reduction online to handle emergencies?*

2.4 Problem Formulation

We consider a general two-level oversubscription: both shared PDUs and UPS are oversubscribed (*i.e.*, the total power consumption by the racks may exceed the shared PDU capacity, while the total capacity of all PDUs may

exceed the UPS capacity) [5], [15]. In case of an emergency,[2] the operator caps power by asking tenants to cut medium-availability power consumption of their racks connected to the affected PDU(s) and/or UPS.

We use R_j^t to denote the amount of total power overload in t at the UPS ($j = 0$) from all the tenants, or at the j-th PDU ($j = 1, \ldots, M$) from the tenants which have racks connected to the PDU. In case of no power emergency at a certain PDU/UPS in t, we have $R_j^t = 0$ for the corresponding j index. Let e_{ij}^t denote the actual usage of medium-availability power by tenant i's racks served by PDU j in t, which can be tracked by meters on racks. We have $e_{ij}^t = 0$ if the actual usage is zero, or tenant i has no racks served by PDU j. In practice, the data center operator continuously monitors tenants' rack-level power usage [4], [5]. Thus, R_j^t can be easily obtained by the operator by deducting the physical capacity of the UPS/PDU from the actual aggregate power usage drawn from the respective UPS/PDU. Meanwhile, e_{ij}^t can be calculated as the difference between the total power usage of tenant i's racks served by PDU j and the total contracted high-availability power capacity allocated to these racks. We use x_{ij}^t to denote the percentage of e_{ij}^t that tenant i is asked to reduce from its racks served by PDU j at time t, $\forall i \in [N], j = 1, 2, \cdots, M$, which represent the decisions that the operator should judiciously make in time slot t when a power emergency occurs.

Fairness. Cutting medium-availability power capacity at runtime can affect tenants' workload performance and must be fairly exercised subject to SLA. Hence, we define a *min-max* fairness objective as follows, which aims at minimizing the maximum ratio of the actual overall power reduction to the contracted power reduction budget per billing period among all the tenants.

$$\text{minimize} \quad \max_{i \in [N]} \{ \frac{\sum_{t \in [T]} \sum_{j \in [1,M]} e_{ij}^t x_{ij}^t}{B_i} \} \tag{1}$$

$$\textbf{s.t.} \quad x_{ij}^t \in [0,1], \forall i \in [N], \forall j \in [1,M], \forall t \in [T] \tag{2}$$

This fairness index takes inspiration from and also extends the widely-applied max-min fairness [19], [20] (usually used for dividing scarce resources in communication network). Minimizing the fair index is equivalent to not only prevent the operator from reducing most of the power from few specific tenants but also to ensure the tenant who reduces most has its remaining budget as much as possible. Rather than fairness in a single time slot, however, our index emphasizes long-term fairness of power reduction (e.g., over an entire billing period in our study), since tenants typically stay in a multi-tenant data center for a long time and a billing period is at least one month. In fact, maximizing fairness for a single time slot may not lead to the maximum long-term fairness, as will be shown in Section 4.

In our fairness definition, a smaller value means "fairer". Consider an example of three emergency events in Table 2, where we show the total power overload that needs to be cut at one shared PDU (*i.e.*, total reduction targets), three involved tenants' medium-availability power usage, and amounts of power reduction (*i.e.*, the numbers following the sign " $\downarrow$") along with their overall power reduction budgets. All the values have the same unit of kW for illustration. Consider a heuristic approach which always reduces a tenant's medium-availability power usage proportional to its total budget, to meet reduction targets during each emergency. In this way, the three tenants reduce power by $(1, 1, 1)$ in the first time slot (emergency event), by $(3, 3, 0)$ in the second time slot, and by $(3, 0, 0)$ in the third time slot, respectively. Consequently, the fairness index of power reduction defined in (1) is $\frac{7}{40} = 0.175$. Nonetheless, the fairest power reduction strategy is that during the three emergencies, the three tenants reduce power by $(0, 0, 3)$, $(1.5, 4.5, 0)$, and $(3, 0, 0)$, respectively, yielding a fairness index of $\frac{4.5}{40} = 0.1125$.

TABLE 2
An Example of Three Power Emergency Events

	Slot#1	Slot#2	Slot#3	Budget
Overall Power Overload	3	6	3	/
Tenant#1	1 ($\downarrow$ 1)	3 ($\downarrow$ 3)	6 ($\downarrow$ 3)	40
Tenant#2	1 ($\downarrow$ 1)	9 ($\downarrow$ 3)	0	40
Tenant#3	10 ($\downarrow$ 1)	0	0	40

While achieving fairness, the following constraints should be met at all times.

PDU/UPS Power Capacity Limitation. The total power reduction by the involved tenants (*i.e.*, e_{ij}^t) should meet the overall power reduction demand at each affected PDU/UPS (*i.e.*, R_j^t and R_0^t) at each time t.

$$\sum_{i \in [N]} e_{ij}^t x_{ij}^t \geq R_j^t, \forall j \in [1,M], \forall t \in [T] \tag{3}$$

$$\sum_{j \in [1,M]} \sum_{i \in [N]} e_{ij}^t x_{ij}^t \geq R_0^t, \forall t \in [T] \tag{4}$$

Power Provisioning SLA Guarantee. The SLA requires that the total accumulated reduction of medium-availability power usage by a tenant be below the threshold (B_i) over a billing period:

$$\sum_{t \in [T]} \sum_{j \in [1,M]} e_{ij}^t x_{ij}^t \leq B_i, \forall i \in [N] \tag{5}$$

Separating our study from the previous research [5], [15], the power provisioning SLA specifies that only medium-availability power usage (i.e., e_{ij}^t) may be reduced and meanwhile eases tenants' concerns of being asked to cut power by too much during emergencies.

Optimization Problem. The fair power capping problem can be formulated by optimizing the fairness objective in (1), subject to constraints in (2), (3), (4) and (5). Let $\lambda = \max_{i \in [N]} \{ \frac{\sum_{t \in [T]} \sum_{j \in [1,M]} e_{ij}^t x_{ij}^t}{B_i} \}$. An *equivalent* formulation of the optimization problem can be expressed as follows:

$$\text{minimize} \quad \lambda \tag{6}$$

2. Readers are referred to [4], [5] for how to detect power emergencies.

subject to:

$$\sum_{i\in[N]} \frac{e_{ij}^t}{R_j^t} x_{ij}^t \geq 1, \forall j \in [1,M], \forall t \in [T] : R_j^t > 0 \quad (6a)$$

$$\sum_{j\in[1,M]} \sum_{i\in[N]} \frac{e_{ij}^t}{R_0^t} x_{ij}^t \geq 1, \forall t \in [T] : R_0^t > 0 \quad (6b)$$

$$\sum_{t\in[T]} \sum_{j\in[1,M]} \frac{e_{ij}^t}{B_i} x_{ij}^t \leq \lambda, \forall i \in [N] \quad (6c)$$

$$x_{ij}^t \leq 1, \forall i \in [N], \forall j \in [1,M], \forall t \in [T] \quad (6d)$$

$$x_{ij}^t \geq 0, \forall i \in [N], \forall j \in [1,M], \forall t \in [T] \quad (6e)$$

$$\lambda \leq 1, \quad (6f)$$

$$\lambda \geq 0, \quad (6g)$$

where (6), (6c), (6f) and (6g) together express (1) and (5) equivalently. (6a) and (6b) correspond to (3) and (4) by removing those trivial constraints where $R_j^t = 0$ or $R_0^t = 0$ and normalizing the LHS and RHS (left-/right-hand side) by the respective power reduction demand in constraints where $R_j^t > 0$ or $R_0^t > 0$. In (6), when all x_{ij}^t are decided, the minimum of variable λ represents the maximum fair index among all tenants. In the final solution, if $\lambda = 1$, it means at least one tenant uses up its reduction budget.

The fair power capping problem in (6) is formulated assuming complete knowledge of power usage over the whole billing period T. We refer to this problem as *offline* optimization when presenting our online algorithm design. In practice, the offline optimization problem is not feasible to solve, because some input variables and constraints only emerge gradually as time progresses. For example, upon the arrival of time slot t, the values of e_{ij}^t, R_j^t, $\forall i \in [N], j \in [1,M]$, and R_0^t for this t are obtained by the data center operator (as discussed in the second paragraph of this subsection); there is a set of new variables $x_{ij}^t, \forall i \in [N], j \in [1,M]$, subject to constraints (6a)(6b)(6d)(6e) for this t. The operator must decide immediately the amount of power reduction by each involved tenant for time t without future knowledge, while respecting the (long-term) SLA and fairness both defined for the entire billing period.

In the following, we design an online algorithm based on the primal-dual framework (The transformation from the primal to the dual is shown in [21]). To begin with, we formulate the dual problem of (6) as follows, where y_j^t, y_0^t, z_i, ϕ_{ij}^t and ξ are dual variables corresponding to constraints (6a) - (6d) and (6f), respectively:

$$\text{maximize} \sum_{t\in[T]} y_0^t + \sum_{t\in[T]} \sum_{j\in[1,M]} y_j^t - \sum_{t\in[T]} \sum_{j\in[1,M]} \sum_{i\in[N]} \phi_{ij}^t - \xi \quad (7)$$

subject to:

$$\frac{e_{ij}^t}{R_j^t} y_j^t + \frac{e_{ij}^t}{R_0^t} y_0^t \leq \frac{e_{ij}^t}{B_i} z_i + \phi_{ij}^t, \quad \forall i \in [N], \forall j \in [1,M], \forall t \in [T] \quad (7a)$$

$$\sum_{i\in[N]} z_i - \xi \leq 1 \quad (7b)$$

$$y_0^t, y_j^t, z_i, \phi_{ij}^t, \xi \geq 0, \forall i \in [N], \forall j \in [1,M], \forall t \in [T] \quad (7c)$$

TABLE 3
Notation Table

Var	Definition
R_j^t	amount of total power overload in t at the UPS ($j = 0$) or at the j-th PDU ($j = 1, \ldots, M$)
e_{ij}^t	actual usage of medium-availability power by tenant i's racks served by PDU j in t.
x_{ij}^t	percentage of e_{ij}^t that tenant i is asked to reduce from its racks served by PDU j at time t
B_i	overall power reduction budget of tenant i
λ	$\lambda = \max_{i\in[N]}\{\frac{\sum_{t\in[T]}\sum_{j\in[1,M]} e_{ij}^t x_{ij}^t}{B_i}\}$
$\mathcal{N}_i$	total number of emergencies that involve tenant i in the entire data center in $[T]$
$\mathrm{U_{OPT}}$	an estimated upper bound of optimal λ

3 Online Algorithm For Fair Power Capping

We next present an online algorithm for the data center operator to make tenants' power reduction decisions upon power emergency events at the UPS and/or PDUs. We then analyze its performance in terms of the *competitive ratio* achieved, computed as the worst-case ratio of the fairness index in (6) derived by our online algorithm, by the offline optimal fairness index, computed with full knowledge of the system over the entire billing period.

3.1 Online Algorithm

Basic Idea. Our algorithm design is based on the primal-dual optimization framework [22] and the dual fitting technique in approximation algorithm design [23], as shown in **Alg.** 1. To meet the power reduction demand in a PDU where power emergency occurs, the algorithm iteratively selects the current best candidate (a tenant) from all tenants which have racks in the PDU and use their medium-availability power in the PDU at the time, and asks this candidate to reduce medium-availability power usage. The current best candidate is one with the smallest ratio of cumulative power reduction so far (since time slot 1 in all PDUs) over its overall power reduction budget, among the remaining candidates whose power reduction has not been decided in a round. After addressing power reduction demands at all PDUs where emergencies occur, if the UPS power capacity is still exceeded, the algorithm repeats a similar procedure: the current best candidate, a (tenant, PDU) pair, is picked among all tenants which are still using their medium-availability power in some PDU(s), and asked to further reduce medium-availability power consumption, until the UPS power capacity is respected. The current best (tenant, PDU) pair is decided by selecting the tenant with the smallest ratio of cumulative power reduction so far over its overall power reduction budget, and the PDU is randomly picked among all PDUs where the tenant is still using medium-availability power.

The amount of medium-availability power that the operator asks a selected tenant to reduce, *i.e.*, primal variable $\mathbf{x}$, is updated by multiplying a well-designed factor in each round, and dual variables $\mathbf{y}$ are updated along (dual fitting), to retain primal and dual feasibility (of all/most constraints) at all times, while minimizing primal objective (the fair index in (6)). Further, a good bound between primal and

Algorithm 1 Fair Online Power Capping Algorithm in t

1: $\mathbf{x} = \mathbf{0}$; $\mathbf{y} = \mathbf{0}$;
2: /*First, handle all PDU emergencies if any*/
3: **for all** PDU $j \in [1, M]$, $R_j^t > 0$ (power emergency) **do**
4: $\kappa = \max_{i\in[N]}\{\mathcal{N}_i\}$, where $\mathcal{N}_i$ represents total # of emergencies involving tenant i in entire data center in $[T]$
5: $\rho = \frac{\max_{i\in[N],j\in[1,M],t\in[T]}\{e_{ij}^t/B_i\}}{\min_{i\in[N],j\in[1,M],t\in[T]:e_{ij}^t>0}\{e_{ij}^t/B_i\}}$
6: $\delta = \min_{t\in[T],j\in[1,M]:R_j^t>0}\{\max_{i\in[N]}\frac{e_{ij}^t}{R_j^t}n_j\}$, where n_j represents no. of tenants in PDU j in t
7: $\sigma = \frac{e}{\ln(2)}\ln(2\kappa\rho\delta)$
8: $\Gamma = 2\sigma \mathsf{U}_{\mathrm{OPT}}$, where $\mathsf{U}_{\mathrm{OPT}}$ is an estimated upper bound of optimal λ
9: $\beta = \frac{1+2\sigma\ln(N)}{\frac{1}{\alpha}+2\sigma\ln(N)}$
10: $\mu = 1 + \frac{1}{(1+2\beta)\ln(eN)}$
11: $x_{ij}^t = \frac{1}{\kappa\rho\delta}, \forall i \in [N] : e_{ij}^t > 0$
12: **while** $\sum_{i\in[N]} e_{ij}^t x_{ij}^t < R_j^t$ **do**
13: $\mathcal{A} = \mathsf{getCandidateSet}(j, t)$
14: Compute $\mathsf{rate}_{ij}^t(x)$ using (10), $\forall i \in [N] : e_{ij}^t > 0$
15: $\epsilon_j^t(\mathbf{x}) = (\mu - 1)\min_{i\in[N]:e_{ij}^t>0}\{\frac{R_j^t \mathsf{rate}_{ij}^t(\mathbf{x})}{e_{ij}^t}\}$
16: **while** $(\mathcal{A} \neq \varnothing)$ and $(\sum_{i\in[N]} e_{ij}^t x_{ij}^t < R_j^t)$ **do**
17: $(i^*, j^*) = \mathsf{nextCandidate}(\mathcal{A}, \mathbf{x}, t)$
18: $x_{i^*j^*}^t = \min\{1, x_{i^*j^*}^t(1 + \epsilon_j^t(\mathbf{x})\frac{e_{i^*j^*}^t}{R_{j^*}^t \mathsf{rate}_{i^*j^*}^t(\mathbf{x})})\}$
19: $\mathcal{A} = \mathcal{A} \setminus (i, j)$
20: **end while**
21: $y_j^t = y_j^t + e\epsilon_j^t(\mathbf{x})$
22: **end while**
23: **end for**
24: /*Next, handle UPS emergency if any*/
25: **if** $R_0^t > 0$ **then**
26: $x_{ij}^t = \frac{1}{\kappa\rho\delta}, \forall i \in [N], \forall j \in [1, M] : e_{ij}^t > 0, x_{ij}^t = 0$
27: **while** $\sum_{j\in[1,M]}\sum_{i\in[N]} e_{ij}^t x_{ij}^t < R_0^t$ **do**
28: $\mathcal{A} = \mathsf{getCandidateSet}(0, t)$
29: Compute $\mathsf{rate}_{ij}^t(x)$ using (10), $\forall i, j : e_{ij}^t > 0$
30: $\epsilon_0^t(\mathbf{x}) = (\mu - 1)\min_{i\in[N],j\in[1,M]:e_{ij}^t>0}\{\frac{R_0^t \mathsf{rate}_{ij}^t(\mathbf{x})}{e_{ij}^t}\}$
31: **while** $(\mathcal{A} \neq \varnothing)$ and $(\sum_{j\in[1,M]}\sum_{i\in[N]} e_{ij}^t x_{ij}^t < R_0^t)$ **do**
32: $(i^*, j^*) = \mathsf{nextCandidate}(\mathcal{A}, \mathbf{x}, t)$
33: $x_{i^*j^*}^t = \min\{1, x_{i^*j^*}^t(1 + \epsilon_0^t(\mathbf{x})\frac{e_{i^*j^*}^t}{R_0^t \mathsf{rate}_{i^*j^*}^t(\mathbf{x})})\}$
34: $\mathcal{A} = \mathcal{A} \setminus (i, j)$
35: **end while**
36: $y_0^t = y_0^t + e\epsilon_0^t(\mathbf{x})$
37: **end while**
38: **end if**

dual objective values can be guaranteed (according to weak duality [22]), which leads to a bounded competitive ratio of the online algorithm.

Algorithm Steps. In **Alg.** 1, PDU power emergencies are handled first (lines 3-23). Lines 4-11 define parameters for initializing and updating primal/dual variables, which are well designed to ensure primal/dual feasibility and the competitive ratio. The While loop in lines 12-22 handles medium-availability power usage reduction among tenants using PDU j. A candidate set is constructed (line 13), which includes all tenants that use medium-availability power in PDU j. We repeatedly identify the current best candidate (line 17) among the remaining tenants in the set and reduce its power usage in the PDU by a multiplicative factor > 1 (line 18).

The multiplicative factor is designed based on the following rationale: the factor is smaller (the tenant is to reduce less power usage) if a unit increase of the respective power reduction amount leads to a larger increase of the fairness index in (6). We evaluate the marginal increase of the fairness index, *i.e.*

$$\lambda(\mathbf{x}) = \max_{i\in[N]}\{\frac{\sum_{t\in[T]}\sum_{j\in[1,M]} e_{ij}^t x_{ij}^t}{B_i}\} \tag{8}$$

using partial derivative in the respective x_{ij}^t. However, since $\lambda(\mathbf{x})$ is not differentiable, we use the following function which approximates $\lambda(\mathbf{x})$:

$$\mathrm{est}(\mathbf{x}) = \Gamma\ln(\sum_{i\in[N]}\exp(\sum_{t\in[T]}\sum_{j\in[1,M]}\frac{e_{ij}^t}{B_i\Gamma}x_{ij}^t)) \tag{9}$$

The reason is that for any series of non-negative real numbers $a_1, a_2, \ldots, a_n$, we have $\max_{i=1,\ldots,n} a_i \leq \ln(\sum_{i=1}^n \exp(a_i)) \leq \max_{i=1,\ldots,n} a_i + \ln(n)$. Here Γ is defined in line 8 of **Alg.** 1, according to an estimated upper bound $\mathsf{U}_{\mathrm{OPT}}$ of the offline optimal fair index. $\mathsf{U}_{\mathrm{OPT}}$ is estimated such that the optimal fair index falls in the range of $[\frac{1}{\alpha}\mathsf{U}_{\mathrm{OPT}}, \mathsf{U}_{\mathrm{OPT}}]$, where $\alpha \geq 1$ (we will show how this estimation influences competitive ratio in Sec. 3.2). Γ is related to β defined in line 9 and then μ in line 10. μ appears in the multiplicative factors for updating primal and dual variables (lines 15, 18 and 21). β appears in μ and will be part of our competitive ratio (to be shown in Theorem 1). The partial derivative of $\mathrm{est}(\cdot)$ in x_{ij}^t is

$$\mathsf{rate}_{ij}^t(\mathbf{x}) = \frac{\partial \mathrm{est}(\mathbf{x})}{\partial x_{ij}^t} = \frac{\frac{e_{ij}^t}{B_i}\exp(\sum_{\bar{t}\in[T]}\sum_{\bar{j}\in[1,M]}\frac{e_{i\bar{j}}^{\bar{t}}}{B_i\Gamma}x_{i\bar{j}}^{\bar{t}})}{\sum_{\bar{i}\in[N]}\exp(\sum_{\bar{t}\in[T]}\sum_{\bar{j}\in[1,M]}\frac{e_{\bar{i}\bar{j}}^{\bar{t}}}{B_{\bar{i}}\Gamma}x_{\bar{i}\bar{j}}^{\bar{t}})} \tag{10}$$

Then the update to $\mathbf{x}$ in line 18 is (where $\epsilon_j^t(\mathbf{x})$ is defined in line 15 of **Alg.** 1):

$$\begin{aligned} x_{ij}^t &= x_{ij}^t(1 + \epsilon_j^t(\mathbf{x})\frac{e_{ij}^t}{R_j^t \mathsf{rate}_{ij}^t(\mathbf{x})}) \\ &= x_{ij}^t(1 + (\mu - 1)\frac{\min_{\bar{i}\in[N]:e_{\bar{i}j}^t>0}\{\mathsf{rate}_{\bar{i}j}^t(\mathbf{x})/e_{\bar{i}j}^t\}}{\mathsf{rate}_{ij}^t(\mathbf{x})/e_{ij}^t}) \end{aligned}$$

Therefore, if $\mathsf{rate}_{ij}^t/e_{ij}^t = \min_{\bar{i}\in[N]:e_{\bar{i}j}^t>0}\{\mathsf{rate}_{\bar{i}j}^t(\mathbf{x})/e_{\bar{i}j}^t\}$, the multiplicative factor is μ; otherwise, the multiplicative factor is smaller ($< \mu$).

The inner While loop in lines 16-20 goes through each candidate tenant once. If the overall medium-availability power reduction has still not met the reduction demand (R_j^t), the algorithm goes through another round of the outer While loop (line 12) to reduce medium-availability power usage at tenants in PDU j again, until the reduction demand is fulfilled. Then the algorithm checks if the UPS-level power capacity has been respected due to power reduction in the PDUs (line 27). If the overall medium-availability power

Algorithm 2 getCandidateSet$(\cdot)$

1: **function** getCandidateSet(j, t)
2: $\mathcal{S} = \varnothing$
3: **if** $j = 0$ **then** ▷ handle UPS power emergency
4: **for all** $(i, j), i \in [N], j \in [M], e_{ij}^t > 0$ **do**
5: $\mathcal{S} = \mathcal{S} \cup (i, j)$
6: **end for**
7: **else** ▷ handle PDU power emergency
8: **for all** $i \in [N], e_{ij}^t > 0$ **do**
9: $\mathcal{S} = \mathcal{S} \cup (i, j)$
10: **end for**
11: **end if**
12: **return** $\mathcal{S}$
13: **end function**

usage in the data center is still positive, similar procedures, as in handling a PDU power emergency, are carried out to further cut medium-availability power consumption at all relevant tenants in the data center (lines 27-37).

The algorithm (lines 36, 21) also updates dual variables y_0^t and y_j^t by the output of algorithms—*i.e.*, y_j^t is updated by ϵ_j^t which is decided by the algorithm, which are constructed for helping analyzing performance bound but do not influence the online decisions of $\mathbf{x}$ directly.

We note that **Alg.** 1 requires $\mathcal{N}_i$, the total number of emergencies that involve tenant i in the entire data center over $[T]$, and $\mathrm{U_{OPT}}$, an estimated upper bound of the optimal fair index, in the computation of κ in line 4 and Γ in line 8, respectively. The exact value of $\mathcal{N}_i$ (and hence κ) is not known before all time slots have passed. Instead, we adopt an estimated κ in our online algorithm, *e.g.*, based on past experience. The estimation of $\mathrm{U_{OPT}}$ influences the theoretical competitive ratio, which will be shown in our analysis in the following section. Further, we will evaluate the impact of inaccurate estimation of these quantities on the performance of our online algorithm in practical settings in the simulation.

Algorithm 3 nextCandidate$(\cdot)$

1: **function** nextCandidate$(\mathcal{A}, \mathbf{x}, t)$
2: $\text{ratio}_{min} = \infty$
3: **for all** $(i, j) \in \mathcal{A}$ **do**
4: $\text{ratio}_i \leftarrow (\sum_{\tau \in [1,t]} \sum_{j \in [1,M]} e_{ij}^\tau x_{ij}^\tau)/B_i$
5: **if** $\text{ratio}_i < \text{ratio}_{min}$ **then**
6: $i_{min} = i$
7: $j_{min} = j$
8: $\text{ratio}_{min} = \text{ratio}_i$
9: **end if**
10: **end for**
11: **return** (i_{min}, j_{min})
12: **end function**

3.2 Theoretical Analysis

We now analyze the competitive ratio achieved by our online algorithm in **Alg.** 1. We first present a few lemmas giving bounds on primal/dual objective values, which lead to the competitive ratio. The detailed proofs of lemmas and theorems can be found in the appendices.

We use OPT to denote the offline optimal fair index in (6), computed by solving the fair power capping problem exactly based on full knowledge of the entire billing period.

Lemma 1. *$\lambda(\mathbf{x}_{\text{initial}}) \le OPT$ and $\text{est}(\mathbf{x}_{\text{initial}}) \le OPT + \Gamma \ln(N)$, where $\lambda(\cdot)$ and $\text{est}(\cdot)$ are defined in* (8) *and* (9)*, respectively, $\mathbf{x}_{\text{initial}}$ denotes the initial values of $x_{ij}^t, \forall i \in [N], j \in [1, M], t \in [T]$, that we set in line 11 and 26 of **Alg.** 1, and Γ is defined in line 8 of **Alg.** 1.*

Lemma 1 shows that the initial fair index set by **Alg.** 1 and the initial value of the $\text{est}(\cdot)$ function we use to approximate $\lambda(\cdot)$, present lower bounds of OPT.

Lemma 2. *The increase in $\sum_{t \in [T]} y_0^t + \sum_{t \in [T]} \sum_{j \in [1,M]} y_j^t$ (part of the dual objective function in* (7)*) upper bounds the increase in $\text{est}(x)$ after every round of the* While *loop in lines 12-22 of **Alg.** 1, or every round of the* While *loop in lines 31-35 of **Alg.** 1.*

In **Alg.** 1, we update primal variables x_{ij}^t, and also increase dual variables y_j^t and y_0^t using ϵ_j^t and ϵ_0^t in line 21 and line 36, in the respective While loop. Lemma 2 shows that the increase of the dual objective value bounds the increase of the (approximated) primal objective value. The following lemma presents how the $\sum_{t \in [T]} y_0^t + \sum_{t \in [T]} \sum_{j \in [1,M]} y_j^t$ part of the dual objective value bounds the (approximated) primal objective value, after the online algorithm has finished running in all time slots $t \in [T]$.

Lemma 3. *$\sum_{t \in [T]} y_0^t + \sum_{t \in [T]} \sum_{j \in [1,M]} y_j^t \ge \text{est}(\mathbf{x}) - \beta(OPT + \Gamma \ln(N))$ where β is defined in line 9 of **Alg.** 1.*

Lemma 3 shows the value of $\sum_{t \in [T]} y_0^t + \sum_{t \in [T]} \sum_{j \in [1,M]} y_j^t$ (part of the dual objective function in (7)) always upper bounds the value of $\text{est}(x)$ minus a constant number.

Lemma 4. *$\sum_{i \in [N]} z_i - \xi \le 1$, if the values of dual variables z_i and ξ are set as*

$$z_i = \frac{w^{max}}{\ln(eN) + \frac{\lambda^{max}}{\Gamma}}, \quad \xi = 0, \quad \forall i \in [N] \tag{11}$$

*where λ^{max} is the largest fair index $\lambda(\mathbf{x})$ that ever appears throughout the update of $\mathbf{x}$ in **Alg.** 1 over $[T]$, and w^{max} denotes the largest value of $\frac{\exp(\sum_{\bar{t} \in [T]} \sum_{\bar{j} \in [1,M]} \frac{e_{i\bar{j}}^{\bar{t}}}{B_i \Gamma} x_{i\bar{j}}^{\bar{t}})}{\sum_{\bar{i} \in [N]} \exp(\sum_{\bar{t} \in [T]} \sum_{\bar{j} \in [1,M]} \frac{e_{\bar{i}\bar{j}}^{\bar{t}}}{B_{\bar{i}} \Gamma} x_{\bar{i}\bar{j}}^{\bar{t}})}$ throughout the update of $\mathbf{x}$ in **Alg.** 1 over $[T]$.*

Lemma 4 shows that z_i's and ξ given in (11) satisfy dual constraint (7b).

Lemma 5. *$(\frac{e_{ij}^t}{R_j^t} y_j^t + \frac{e_{ij}^t}{R_0^t} y_0^t) \le (\frac{e_{ij}^t}{B_i} z_i + \phi_{ij}^t) \cdot \sigma(\ln(eN) + \frac{\lambda^{max}}{\Gamma})$, if the value of dual variable ϕ_{ij}^t is set as*

$$\phi_{ij}^t = 0, \forall i \in [N], j \in [1, M], t \in [T] : R_j^t > 0, R_0^t > 0 \tag{12}$$

*where σ is defined in line 7 of **Alg.** 1.*

Comparing the inequality in Lemma 5 with dual constraint (7a), we know that if we scale $\mathbf{y}$ obtained by running **Alg.** 1 over $[T]$ by a factor $\sigma(\ln(eN) + \frac{\lambda^{max}}{\Gamma})$, the scaled $\mathbf{y}$ together with $\mathbf{z}$, ξ in (11) and $\vec{\phi}$ in (12) give a set of feasible dual solutions. These dual feasible solutions are useful in deriving the competitive ratio given in the following theorem, using weak duality (*i.e.*, the optimal primal

objective value is lower bounded by the dual objective value computed using any dual feasible solutions).

Theorem 1. ***Alg.*** *1 is* $2\sigma\left(1+\beta+\alpha\beta\right)\ln(eN)$*-competitive, where* $\sigma = \frac{e}{\ln(2)}\ln(2\kappa\rho\delta)$ *(defined in lines 4-6 in* ***Alg.*** *1),* $\beta = \frac{1+2\sigma\ln(N)}{\frac{1}{\alpha}+2\sigma\ln(N)}$*, and* α *satisfies* $\frac{1}{\alpha}\mathrm{U_{OPT}} \leq OPT \leq \mathrm{U_{OPT}}$.

The competitive ratio means $\frac{\lambda(\mathbf{x})}{OPT} \leq 2\sigma\left(1+\beta+\alpha\beta\right)\ln(eN)$, where $\lambda(\mathbf{x})$ is the final fairness index derived by running **Alg.** 1 over $[T]$. The competitive ratio is relevant to $\mathrm{U_{OPT}}$, an estimated upper bound of the offline optimal fair index OPT, used as input to the online algorithm. If $\mathrm{U_{OPT}}$ is much larger than OPT, we need a large α to ensure that OPT falls in $[\frac{1}{\alpha}\mathrm{U_{OPT}}, \mathrm{U_{OPT}}]$. If $\mathrm{U_{OPT}}$ is close to OPT, α is smaller. In our model, α and κ are the estimated parameters. According to the definition of α, the estimation of $\mathrm{U_{OPT}}$ is more accurate; the ratio is smaller. Another variable κ is implicitly shown in the parameter β; however, according to the definition of σ and β, when estimation is more inaccurate, needs to be set larger, and the σ will be smaller, and the ratio is smaller. However, κ decides the initial value of x_{ij}^t (line 11 and line 26 in **Alg.** 1). Correspondingly, the execution time of algorithms will increase according to line 18 and line 33 in **Alg.** 1.

Finally, we note that based on settings of the data center system, it is possible that $OPT \cdot 2\sigma\left(1+\beta+\alpha\beta\right)\ln(eN)$ is larger than 1, such that $\lambda(\mathbf{x})$ may potentially be larger than 1, implying that **Alg.** 1 might violate the power reduction budget constraint for some tenant(s). This is in fact common for online algorithms solving problems with mixed packing and covering constraints (like (6)) [24], due to the hardness of fulfilling covering constraints without exceeding upper limits set by packing constraints, in case that no future information is known. In our scenario, it is even harder because coefficients in the LHS of packing constraints (*i.e.*, (5)) are also unknown in advance. Nonetheless, $OPT \cdot 2\sigma\left(1+\beta+\alpha\beta\right)\ln(eN)$ can bound the degree of violation of the budget constraints, even in the worst case. For example, if $OPT \cdot 2\sigma\left(1+\beta+\alpha\beta\right)\ln(eN) = 1.2$, we have $\lambda(\mathbf{x}) \leq 1.2$, which means that B_i for some tenant i might be exceeded by at most 20%. In practice, if the operator observes the need of exceeding a tenant's budget when running **Alg.** 1, the operator can provide compensation as if the operator fails to provide the contracted guaranteed capacity to the tenant (*e.g.*, US$3/kW for each hour of unavailability [7]). Further, we will show through simulations under practical settings that, **Alg.** 1 always obtains feasible solutions with realistic power reduction budget settings.

Theorem 2. *When each emergency happens, the time complexity of* ***Alg.*** *1 is* $O(NM\frac{\log(\kappa\rho\delta)}{\log(1+\frac{1}{(1+2\alpha)\ln(eN)})})$ *where* $\sigma = \frac{e}{\ln(2)}\ln(2\kappa\rho\delta)$ *(defined in lines 4-6 in* ***Alg.*** *1) and* α *satisfies* $\frac{1}{\alpha}\mathrm{U_{OPT}} \leq OPT \leq \mathrm{U_{OPT}}$.

4 Performance Evaluation

We now evaluate our online algorithm in **Alg.** 1 via simulations, and highlight that it can fairly handle power emergencies within reduction budgets.

4.1 Simulation Setup

Power capacities. By default, we consider a standard medium-size multi-tenant cloud data center with 1 UPS, 4 PDUs, and 10 tenants. Each PDU serves 5-10 tenants, while each tenant deploys server racks across 2-4 PDUs. Later, we will also change the settings for sensitivity studies. Each PDU has a power capacity of 300kW (*i.e.*, hosting around one thousand servers) and the UPS has a power capacity of 1091kW (set to allow 110% oversubscription at the UPS by PDUs, following the recent literature [5], [15]). We assume for simplicity that the tenants equally share the UPS capacity as their guaranteed high-availability capacities (*i.e.*, $C_i^g = \frac{1091}{10}$kW), and each tenant buys medium-availability capacity (C_i^f) equal to 10%-20% of its guaranteed high-availability capacity. C_i^g and C_i^f are evenly divided among PDUs where tenant i has server racks. We simulate a *one-month* billing period (T). Each time slot is 5 minutes long.

Tenant power usage. We simulate tenants' power usage based on workload trace from Google clusters [25], which has been widely cited in the studies about data center workloads. Moreover, besides deploying servers in their own data centers, deploys servers in multi-tenant data centers at hundreds of locations. The trace contains resource usage (CPU, RAM, and Disk) and the execution time slots of each job. We select representative jobs in the trace, assign them to tenants' server racks in different PDUs, and then scale the CPU usage trace to get tenants' power usage in all PDUs, such that power emergencies (at PDU(s) and/or the UPS) occur in about 3% of all the time slots. Fig. 2 illustrates the number of produced power emergencies at the PDUs and the UPS. This setting is consistent with real-world measurements and also used in the prior research [4], [5], [15].

Power reduction budget. We set the power reduction budget such that on average a tenant would reduce at most 30% of its overall medium-availability capacity during all emergencies.

Comparisons. We compare our online algorithm with the following schemes based on reasonable and best-known studies. (1) OPT: the offline optimum computed by solving (6) exactly using `Gurobi` [26] and assuming full knowledge of power usage during the entire billing period. (2) PropR1: when an emergency happens, the operator reduces the power from involved tenants that are using their medium-availability power in proportion to their budgets B_i. (3) PropR2: the operator reduces involved tenants' medium-availability power in proportion to their medium-availability power usage at the current slot. (4) RandR: the operator repeatedly randomly picks one tenant that is using its medium-availability power, and cuts its medium-availability power usage as much as possible, until diminishing the emergency.

We run our algorithms on an Intel Core i7-6820HQ processor. For each emergency, under the typical setup, the running time is less than one second. Moreover, in real-world environment, using Intel RAPL [27], the device power can be reduced in a few milliseconds.

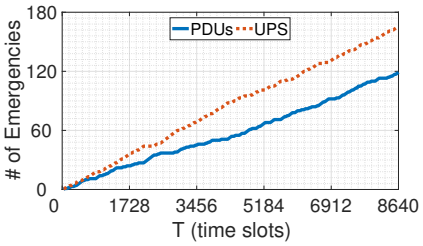


Fig. 2. Cumulative number of emergencies.

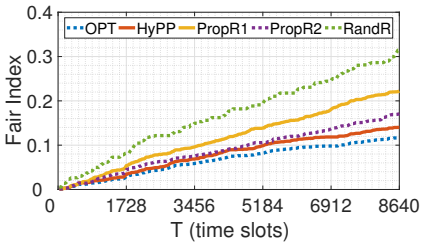


Fig. 3. Fair index: different schemes.

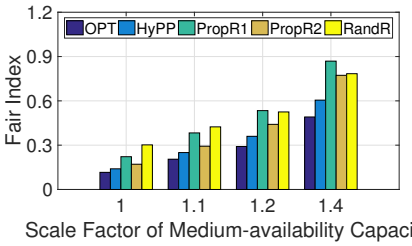


Fig. 4. Fair index: medium-availability capacity.

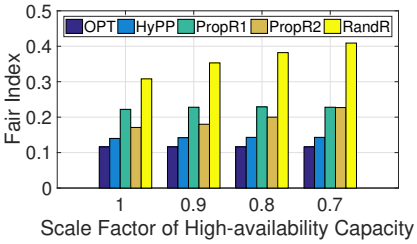


Fig. 5. Fair index: different high-availability capacities.

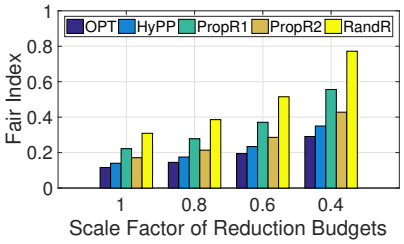


Fig. 6. Fair index: different power reduction budgets.

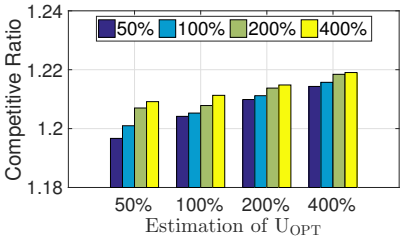


Fig. 7. Competitive ratio: different estimation accuracies of $\mathrm{U_{OPT}}$ and different κ (in legend).

4.2 Evaluation Results

Fairness index. In **Fig.** 3, we compare the fairness index (defined in (6)) achieved over time among the different schemes. At the end of the one-month period, our algorithm obtains a fair index of 0.14, and the optimum is 0.116—*i.e.*, a ratio of around 1.2. We can see that our mechanism is "fairer" than the baseline schemes (except for the offline optimum, which is not attainable in practice).

Impact of medium-availability capacity levels. In **Fig.** 4, we evaluate the impact of different medium-availability capacity levels at tenants, by multiplying a factor to scale the default value. Correspondingly, we also scale tenants' power usage, while keeping their high-availability capacities unchanged. We can see our algorithm is close to the offline optimum and better than other baselines in terms of fairness. When the operator sells more medium-availability capacities, emergencies occur more frequently, leading to more power reduction at tenants (*i.e.*, fairness index increases for all the schemes, but the relative comparison remains the same).

Impact of high-availability capacity levels. In **Fig.** 5, we decrease tenants' high-availability capacities by multiplying a scale factor. We keep the tenants' power usage and total power capacities the same (increasing medium-availability capacities). In this way, when an emergency occurs, the operator has more freedom to decide power reduction from tenants. Nonetheless, as the tenants' power usage remains same which results in the same emergencies, the offline optimum does not change. As shown in **Fig.** 5, our algorithm and PropR1 are not affected by high-availability capacity levels. However, PropR2 and RandR become worse, because during certain time slots they might "mistakenly" reduce relatively more power from one tenant according to their reduction policies.

Impact of power reduction budget. In **Fig.** 6, we scale down tenants' power reduction budgets from our default setting. We can see that even if the budget reduces to 40% of the default setting (*i.e.*, 12% of tenant's overall medium-availability capacity during all emergencies), the worst RandR scheme still only uses at most 80% of a tenant's budget. This implies that budget feasibility can be easily guaranteed under realistic settings.

Impact of estimations of $\mathrm{U_{OPT}}$ and $\mathcal{N}_i$. Our previous experiments are conducted by setting $\mathrm{U_{OPT}}$ to OPT and $\mathcal{N}_i$'s to the actual values. We next evaluate the impact of different estimation accuracies of their values. $\mathcal{N}_i$'s decide κ in line 4 of **Alg.** 1. In **Fig.** 7, we vary $\mathrm{U_{OPT}}$ to different percentages of OPT, while changing κ among different percentages of its actual value. We observe that the ratio of fair index obtained by **Alg.** 1 to OPT only varies slightly around that obtained by setting $\mathrm{U_{OPT}}$ to OPT and κ to the actual value (caused by the changes of updating steps shown in line 18 in **Alg.** 1, related to rate_{ij}^t, Γ, σ), showing that our online algorithm is insensitive to inaccurate estimation.

Impact of emergency probabilities. In **Fig.** 8, a higher emergency probability means that there are power emergencies in more time slots (our default is 3%). When emergency ratio increases, fair indices of all benchmarks increase. Moreover, the difference between algorithms and the optimum increases too because the algorithms make more "mistakes" (compared with the optimum) without future information with more emergencies.

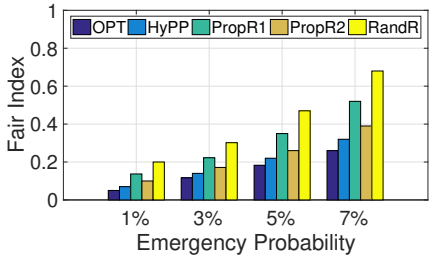


Fig. 8. Fair index: different emergency probabilities.

Impact of tenant number. In **Fig.** 9, we evaluate the impact of the number of tenants, by multiplying the tenant numbers by a factor (up to 4 times greater than the default setting). Correspondingly, we proportionally scale down

tenants' capacities, power usage, and budgets according to UPS/PDU capacities. We can see that the fairness indices of OPT, HyPP, PropR1, and PropR2 are only slightly affected, showing that our algorithm is not sensitive to the number of tenants; meanwhile, the index of RandR becomes much better since the random selection becomes more "average" with more tenants.

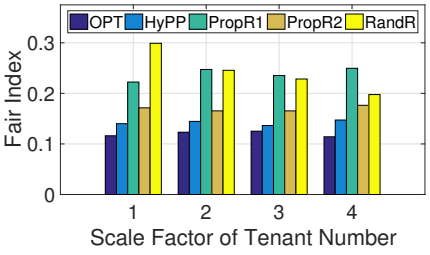


Fig. 9. Fair index: different numbers of tenants.

Utilization of power capacity of the entire data center. In **Fig.** 10, we show the power utilization of the entire data center with/without HyPP with one-day trace. In our setting, for selling 20% of the guaranteed capacity as the medium-availability capacity, there is improvement of the utilization because HyPP utilizes the remanent of power capacity when different tenants do not arrive their peaks simultaneously.

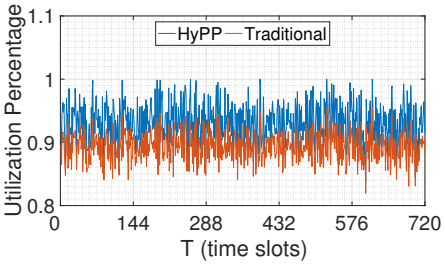


Fig. 10. Power Utilization: HyPP and the traditional.

5 Related Work

Energy efficiency has remained a focal point of research, and many solutions have been proposed, including dynamic server provisioning [28], geographic load balancing [29], incorporating renewables to reduce brown energy usage [17], power provisioning over distributed data centers in grids [30], [31], among others. These studies focus on minimizing the operating expense, and are complementary to maximizing the data center capacity utilization for cutting the capital expense.

Power oversubscription has been commonly applied to improve utilization of expensive power infrastructure [4], [11], [15]. To handle the resulting emergencies, [11] proposes coordinated rack-level power capping, [8] studies statistic-based power profiling, and [4] implements a priority-based power capping in large systems. These solutions focus on an owner-operated data center (or single *tenant*). In addition, they study power capping only for a single time slot without considering (long-term) fairness or power provisioning SLA. Discharging energy storage devices (ESD) [15] has also been proposed to handle emergencies. Our solution is complementary and, when combined with ESD discharging, can improve power provisioning SLA with less frequent power capping.

For multi-tenant data centers, various incentive mechanisms have been proposed to coordinate tenants' power usage for energy cost saving [18], [32], demand response [10] and power capping [5], etc. While [5] handles emergencies, it relies on tenants' voluntary power reduction on a best-effort basis, neglects fairness and provides no power provisioning SLA. Our solution addresses these limitations and is provably fair, even compared to the offline optimum.

Fairness is a key consideration for resource allocation. Several fairness indices have been studied, including max-min fairness [19] (considering the one-slot fairness for networks), long-term fairness [33] (fairly charging users according to their resource usage), and recently dominant resource fairness [20] (considering the one-slot fairness in terms of the dominant resource among different types of resources). These studies do not apply to our problem, because the tenants' medium-availability power usage is constrained by the *interdependent* multi-level PDU/UPS power capacity constraints, and the goal is to achieve bounded online fairness with the long-term (*i.e.*, temporal coupling) power reduction budget constraints. The long-term budgets have never been considered in the existing work. Our work is remotely related to the growing literature about spot instances, but our work also has several fundamental differences, *e.g.*, the long-time guarantee and the segregation of resources. For applying our approach to a production environment, some existing interfaces and accounting mechanisms can be used for tracking the budget usage as important pieces; and the algorithms can also be regulated according to specific situations for speeding up. Furthermore, to our knowledge, hybrid power provisioning and online power capping in multi-tenant data centers have not been considered in the prior literature.

6 Concluding Remarks

This paper proposes HyPP, a novel hybrid power provisioning approach that provides two different power availabilities in a multi-tenant cloud data center: high-availability capacity and medium-availability capacity subject to SLA. We also design an online algorithm for the operator to coordinate tenants' power reduction in cases of power emergencies, targeting long-term fairness among tenants. We prove the competitive ratio of the online algorithm and run simulations to validate the analysis, showing that our algorithm achieves a good fairness, as compared to different alternative schemes.

7 Acknowledgements

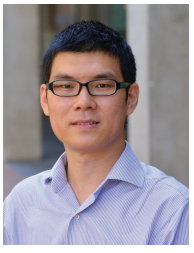
This work was supported in part by grants from Hong Kong RGC under the contracts HKU 718513, 17204715, 17225516, C7036-15G (CRF), and by the U.S. NSF under grants CNS-1551661, CNS-1565474, and ECCS-1610471.

References

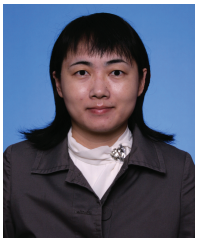
[1] Y. Sverdlik, "Google to build and lease data centers in big cloud expansion," in *DataCenterKnowledge*, April 2016.

[2] Apple, "Environmental responsibility report," 2016.
[3] Data Center Consolidation and Optimization, https://cio.gov/drivingvalue/data-center-consolidation/.
[4] Q. Wu, Q. Deng, L. Ganesh, C.-H. R. Hsu, Y. Jin, S. Kumar, B. Li, J. Meza, and Y. J. Song, "Dynamo: Facebook's data center-wide power management system," in *ISCA*, 2016.
[5] M. A. Islam, X. Ren, S. Ren, A. Wierman, and X. Wang, "A market approach for handling power emergencies in multi-tenant data center," in *HPCA*, 2016.
[6] NRDC, "Scaling up energy efficiency across the data center industry: Evaluating key drivers and barriers," *Issue Paper*, Aug. 2014.
[7] Internap, "Colocation services and SLA," http://goo.gl/lSCddj.
[8] S. Govindan, J. Choi, B. Urgaonkar, and A. Sivasubramaniam, "Statistical profiling-based techniques for effective power provisioning in data centers," in *EuroSys*, 2009.
[9] Ponemon Institute, "2013 cost of data center outages," 2013, http://goo.gl/6mBFTV.
[10] L. Zhang, S. Ren, C. Wu, and Z. Li, "A truthful incentive mechanism for emergency demand response in colocation data centers," in *INFOCOM*, 2015.
[11] X. Fu, X. Wang, and C. Lefurgy, "How much power oversubscription is safe and allowed in data centers," in *ICAC*, 2011.
[12] "Intel Rack Scale Design," https://www-ssl.intel.com/content/www/us/en/architecture-and-technology/rack-scale-design-overview.html.
[13] Google Cloud Storage Nearline, https://cloud.google.com/storage-nearline/.
[14] S. Pelley, D. Meisner, P. Zandevakili, T. F. Wenisch, and J. Underwood, "Power routing: Dynamic power provisioning in the data center," in *ASPLOS*, 2010.
[15] D. Wang, C. Ren, and A. Sivasubramaniam, "Virtualizing power distribution in datacenters," in *ISCA*, 2013.
[16] S. Subramanya, Z. Mustafa, D. Irwin, and P. Shenoy, "Beyond energy-efficiency: Evaluating green datacenter applications for energy-agility," in *ICPE*, 2016.
[17] I. Goiri, R. Beauchea, K. Le, T. D. Nguyen, M. E. Haque, J. Guitart, J. Torres, and R. Bianchini, "Greenslot: scheduling energy consumption in green datacenters," in *SuperComputing*, 2011.
[18] M. A. Islam, H. Mahmud, S. Ren, and X. Wang, "Paying to save: Reducing cost of colocation data center via rewards," in *HPCA*, 2015.
[19] B. Radunović and J.-Y. L. Boudec, "A unified framework for max-min and min-max fairness with applications," *IEEE/ACM Trans. Netw.*, vol. 15, no. 5, pp. 1073–1083, Oct. 2007.
[20] 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.
[21] "Duality in Linear Programming," http://web.mit.edu/15.053/www/AMP-Chapter-04.pdf.
[22] N. Buchbinder and J. Naor, "The design of competitive online algorithms via a primal: dual approach," *Foundations and Trends in Theoretical Computer Science*, vol. 3, no. 2–3, pp. 93–263, 2009.
[23] V. V. Vazirani, *Approximation algorithms*. Springer Science & Business Media, 2013.
[24] Y. Azar, U. Bhaskar, L. Fleischer, and D. Panigrahi, "Online mixed packing and covering," in *SODA*, 2013.
[25] C. Reiss, J. Wilkes, and J. L. Hellerstein, "Google cluster-usage traces: format + schema," Google Inc., Mountain View, CA, USA, Technical Report, Nov. 2011.
[26] Gurobi Optimization, http://www.gurobi.com/.
[27] "Running Average Power Limit RAPL," https://01.org/zh/blogs/2014/running-average-power-limit-%E2%80%93-rapl.
[28] M. Lin, A. Wierman, L. L. H. Andrew, and E. Thereska, "Dynamic right-sizing for power-proportional data centers," in *INFOCOM*, 2011.
[29] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs, "Cutting the electric bill for internet-scale systems," in *SIGCOMM*, 2009.
[30] L. Yu, T. Jiang, and Y. Zou, "Distributed real-time energy management in data center microgrids," *IEEE Transactions on Smart Grid*, 2016.
[31] ——, "Price-sensitivity aware load balancing for geographically distributed internet data centers in smart grid environment," *IEEE Transactions on Cloud Computing*, 2016.
[32] 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 *eEnergy*, 2015.
[33] S. Tang, B.-s. Lee, B. He, and H. Liu, "Long-term resource fairness: Towards economic fairness on pay-as-you-use computing systems," in *ICS*, 2014.

Qihang Sun received the B.S. degree from Wuhan University, China, in 2014, in Computer Science and Technology. He is currently working towards the Ph.D. degree in the Computer Science Department at The University of Hong Kong. His research interests include power-aware computing, online optimization, and cloud computing.



Shaolei Ren is an Assistant Professor of Electrical and Computer Engineering at University of California, Riverside. He received his B.E. from Tsinghua University in 2006, M.Phil. from Hong Kong University of Science and Technology in 2008, and Ph.D. from University of California, Los Angeles in 2012, all in electrical and computer engineering. His research interests include cloud computing, data centers, and network economics. He was a recipient of the U.S. NSF Faculty Early Career Development (CAREER) Award in 2015. He also received best paper awards from several conferences, including ACM e-Energy'16, IEEE ICC'16 and IEEE ICC'09.



Chuan Wu received her B.E. and M.E. degrees in 2000 and 2002 from Department of Computer Science and Technology, Tsinghua University, China,and her Ph.D. degree in 2008 from the Department of Electrical and Computer Engineering, University of Toronto, Canada. She is currently an associate professor in the Department of Computer Science,The University of Hong Kong, China. Her research interests include cloud computing and online/mobile social network.

APPENDIX A
PROOF OF LEMMA 1

Proof. When a new covering constraint arrives (*i.e.*, a new power emergency, R_j^t, happens in one PDU), the system must satisfy it, *i.e.*, $\sum_{i\in[N]} e_{ij}^t x_{ij}^t \geq R_j^t$. Hence, in the optimal solution $\mathbf{x}^*$, for any j,t (*i.e.*, any power emergency signal in PDUs), we have $\max_i\{e_{ij}^t\}\sum_i x_{ij}^{t*} \geq \sum_{i\in[N]} e_{ij}^t x_{ij}^{t*} \geq R_j^t$, and then there must exist variable $x_{base}^* \geq \max_{j\in[1,M],t\in[T]:R_j^t>0}\{\frac{1}{\max_{i\in[N]}\{e_{ij}^t/R_j^t\}\cdot n_j}\}$. Therefore,

$$\begin{aligned}
OPT &= \max_{i\in[N]}\{\sum_{j\in[1,M]}\sum_{t\in[T]}\frac{e_{ij}^t}{B_i}(\mathbf{x}^*)_{ij}^t\}\\
&\geq \min_{i\in[N],j\in[1,M],t\in[T]:e_{ij}^t>0}\{e_{ij}^t/B_i\}x_{base}^*\\
&\geq \min_{i\in[N],j\in[1,M],t\in[T]:e_{ij}^t>0}\{e_{ij}^t/B_i\}\\
&\quad\cdot \max_{j\in[1,M],t\in[T]:R_j^t>0}\{\frac{1}{\max_{i\in[N]}\{e_{ij}^t/R_j^t\}\cdot n_j}\}.
\end{aligned}$$

Using $\rho = \frac{\max_{i\in[N],j\in[1,M],t\in[T]}\{e_{ij}^t/B_i\}}{\min_{i\in[N],j\in[1,M],t\in[T]:e_{ij}^t>0}\{e_{ij}^t/B_i\}}$ and $\delta = \min_{t\in[T],j\in[1,M]:R_j^t>0}\{\max_{i\in[N]}\frac{e_{ij}^t}{R_j^t}n_j\}$

$$OPT \geq \max_{i\in[N],j\in[1,M],t\in[T]:e_{ij}^t>0,R_j^t>0}\{\frac{e_{ij}^t}{B_i}\}\frac{1}{\rho\delta}. \tag{13}$$

When a tenant needs to reduce power in PDU j at time slot t at its first time, $(\mathbf{x}_{\text{initial}})_{ij}^t = \frac{1}{\kappa\rho\delta}$, and thus

$$\begin{aligned}
\lambda(\mathbf{x}_{\text{initial}}) &\leq \max_{i\in[N]}\{\mathcal{N}_i\cdot \max_{j\in[1,M],t\in[T]:R_j^t>0}\{\frac{e_{ij}^t}{B_i}\}\cdot(x_{ij}^t)_{\text{initial}}\}\\
&\leq \max_{i\in[N]}\{\mathcal{N}_i\}\cdot \max_{i\in[N],j\in[1,M],t\in[T]:R_j^t>0}\{\frac{e_{ij}^t}{B_i}\}\cdot(x_{ij}^t)_{\text{initial}}\\
&\leq \max_{i\in[N]}\{\mathcal{N}_i\}\cdot \max_{i\in[N],j\in[1,M],t\in[T]:R_j^t>0}\{\frac{e_{ij}^t}{B_i}\}\cdot\frac{1}{\kappa\rho\delta}\\
&\leq OPT.
\end{aligned}$$

where the last inequality is due to (13). Thus, based on the property of est($\mathbf{x}$)—*i.e.*, est($\mathbf{x}$) $\leq \lambda + \Gamma\ln(N)$, we have est($\mathbf{x}_{\text{initial}}$) $\leq OPT + \Gamma\ln(N)$. □

APPENDIX B
PROOF OF LEMMA 2

Proof. When new emergencies (*i.e.*, R_j^t or R_0^t) appear at time t, PDU j or UPS must reduce tenants' medium-availability power usage. Let P_j denote the set of tenants that PDU j includes, and let P_0 denote the set of all tenants. We also divide the power reduction over the whole bill period by multiple rounds indexed by l—the updates of variables within line 13-21 or line 28-36 belong to the same round. $\mathbf{x}^{(l)}$ and $\mathbf{x}^{(l+1)}$ represents the vector of $\mathbf{x}$ before and after round l, and $(\mathbf{x}^{(l)})_{ij}^t$ represents the value of x_{ij}^t in $\mathbf{x}^{(l)}$. Thus, if round l corresponds to a PDU emergency R_j^t, $\{i\in P_j : (\mathbf{x}^{(l+1)})_{ij}^t > (\mathbf{x}^{(l)})_{ij}^t\}$ denotes the tenants that reduce their power for PDU j; otherwise (*i.e.*, if round l corresponds to a UPS emergency R_0^t, $\{(i,j)\in P_0 : (\mathbf{x}^{(l+1)})_{ij}^t > (\mathbf{x}^{(l)})_{ij}^t\}$ denotes the tenants that reduce their power for UPS.

Next, we analyze the incremental of est($\mathbf{x}$) by line 18 and 33 in **Alg.** 1. Let est^l and est^{l+1} denote the value of est($\mathbf{x}$) before and after round l. We formulate one function $g_{ij}^t(u) = (\mathbf{x}^l)_{ij}^t + ((\mathbf{x}^{l+1})_{ij}^t - (\mathbf{x}^l)_{ij}^t)u$ subject to $u\in[0,1]$ where $g_{ij}^t(0) = (\mathbf{x}^l)_{ij}^t$ and $g_{ij}^t(1) = (\mathbf{x}^{l+1})_{ij}^t$, and let $\mathbf{g}(u)$ denote the vector of $g_{ij}^t(u)$. For simplicity, we let est(u) = est$(\mathbf{g}(u)) = \ln(\sum_{i\in[N]}\exp(\sum_{t\in[T]}\sum_{j\in[M]}\frac{e_{ij}^t}{B_i\Gamma}g_{ij}^t(u)))$ and rate$_{ij}^t(u)$ = rate$_{ij}^t(\mathbf{g}(u))$, and we have

$$\text{rate}_{ij}^t(u) = \text{rate}_{ij}^t(\mathbf{g}(u)) = \frac{\partial\text{est}(u)}{\partial g_{ij}^t(u)}. \tag{14}$$

Then, we have

$$\frac{\text{dest}(u)}{\text{d}u} = \sum_{i\in[N]}\sum_{j\in[M]}\frac{\partial\text{est}(u)}{\partial g_{ij}^t(u)}\frac{\text{d}g_{ij}^t(u)}{\text{d}u} = \sum_{i\in[N]}\sum_{j\in[M]}\text{rate}_{ij}^t(u)\frac{\text{d}g_{ij}^t(u)}{\text{d}u}$$

by the chain rule and (14). As each round l only belongs to one specific time slot t, we only sum up $i\in[N]$ and $j\in[M]$ with the specific t. Specifically, for one PDU emergency, *e.g.*, R_j^t, we only need to sum up $i\in P_j$ with specific j; and for one UPS emergency, *e.g.*, R_0^t, we need to sum up $\forall(i,j)\in P_0$. The analyses are similar, and we first analyze the PDU emergency. We have

$$\begin{aligned}
\text{est}^{l+1} - \text{est}^l &= \int_{u=0}^1\frac{\text{dest}(u)}{\text{d}u}\text{d}u\\
&= \sum_{i\in P_j:(\mathbf{x}^{(l+1)})_{ij}^t>(\mathbf{x}^{(l)})_{ij}^t}\int_{u=0}^1\text{rate}_{ij}^t(u)\frac{\text{d}g_{ij}^t(u)}{\text{d}u}\text{d}u.
\end{aligned}$$

By Lemma 7 (in Appendix H), for any $0\leq u\leq 1$, rate$_{ij}^t(u) \leq e\cdot$ rate$_{ij}^t(0)$. Hence,

$$\begin{aligned}
&\text{est}^{l+1} - \text{est}^l\\
&\leq e\cdot\sum_{i\in P_j:(\mathbf{x}^{(l+1)})_{ij}^t>(\mathbf{x}^{(l)})_{ij}^t}\text{rate}_{ij}^t(\mathbf{x}^{(l)})\int_{u=0}^1\frac{\text{d}g_{ij}^t(u)}{\text{d}u}\text{d}u\\
&= e\cdot\sum_{i\in P_j:(\mathbf{x}^{(l+1)})_{ij}^t>(\mathbf{x}^{(l)})_{ij}^t}(\text{rate}_{ij}^t(\mathbf{x}^{(l)})\cdot((\mathbf{x}^{(l+1)})_{ij}^t - (\mathbf{x}^{(l)})_{ij}^t)).
\end{aligned}$$

As in each round l each variable $x_{ij}^t, \forall i\in P_j$ with positive coefficient e_{ij}^t is updated by line 18 in **Alg.** 1, we have

$$\begin{aligned}
&\text{est}^{l+1} - \text{est}^l\\
&\leq e\cdot\sum_{i\in P_j:(\mathbf{x}^{(l+1)})_{ij}^t>(\mathbf{x}^{(l)})_{ij}^t}(\text{rate}_{ij}^t(\mathbf{x}^{(l)})(\epsilon_j^t(\mathbf{x}^{(l)})\frac{e_{ij}^t}{R_j^t\text{rate}_{ij}^t})(\mathbf{x}^{(l)})_{ij}^t)\\
&\leq e\cdot\epsilon_j^t\sum_{i\in P_j:(\mathbf{x}^{(l+1)})_{ij}^t>(\mathbf{x}^{(l)})_{ij}^t}(\frac{e_{ij}^t}{R_j^t}(\mathbf{x}^{(l)})_{ij}^t)\\
&\leq e\cdot\epsilon_j^t,
\end{aligned}$$

where the last inequality is due to not meeting the power reduction demand in the power emergency before updating $\mathbf{x}^{(l)}$ (*i.e.*, there is a power emergency—an unsatisfied covering constraint—$\sum_{i\in P_j:(\mathbf{x}^{(l+1)})_{ij}^t>(\mathbf{x}^{(l)})_{ij}^t}(\frac{e_{ij}^t}{R_j^t}(\mathbf{x}^{(l)})_{ij}^t) < 1$).

Hence, according to line 21, the increase of y_j^t bounds the increase of est($\mathbf{x}$) in the round corresponding to a PDU emergency.

Following a similar approach, we analyze the increase of y_0^t in the round corresponding to an UPS emergency. We have

$$\begin{aligned}
&\text{est}^{l+1} - \text{est}^l \\
&= \sum_{(i,j)\in P_0:(\mathbf{x}^{(l+1)})_{ij}^t > (\mathbf{x}^{(l)})_{ij}^t} \int_{u=0}^{1} \text{rate}_{ij}^t(u) \frac{\mathrm{d}g_{ij}^t(u)}{\mathrm{d}u} \mathrm{d}u \\
&\le e \cdot \sum_{(i,j)\in P_0:(\mathbf{x}^{(l+1)})_{ij}^t > (\mathbf{x}^{(l)})_{ij}^t} (\text{rate}_{ij}^t(\mathbf{x}^{(l)})((\mathbf{x}^{(l+1)})_{ij}^t - (\mathbf{x}^{(l)})_{ij}^t)) \\
&\le e \cdot \epsilon_0^t \sum_{(i,j)\in P_0:(\mathbf{x}^{(l+1)})_{ij}^t > (\mathbf{x}^{(l)})_{ij}^t} (\frac{e_{ij}^t}{R_0^t}(\mathbf{x}^{(l)})_{ij}^t) \\
&\le e \cdot \epsilon_0^t \sum_{i\in[N]} \sum_{j\in[M]} (\frac{e_{ij}^t}{R_0^t}(\mathbf{x}^{(l)})_{ij}^t) \\
&\le e \cdot \epsilon_0^t,
\end{aligned}$$

and thus according to line 36 in **Alg.** 1, the increase of y_0^t bounds the increase of est($\mathbf{x}$) in the round for UPS power emergency.

Therefore, the increase of $\sum_{t\in[T]} y_0^t + \sum_{t\in[T]} \sum_{j\in[M]} y_j^t$ is an upper bound of the increase of est($\mathbf{x}$). □

APPENDIX C
PROOF OF LEMMA 3

Proof. By Lemma 2, the increase of $(\sum_{t\in[T]} y_0^t + \sum_{t\in[T]} \sum_{j\in[M]} y_j^t)$ bounds the increase of est($\mathbf{x}$) during the updating process of x_{ij}^t multiplied by a factor (Line 18 and 33 in **Alg.** 1). By Lemma 1 and Lemma 6 (in Appendix G), we know that $\beta(OPT/\Gamma + \ln(N))$ bounds the increase of est($\mathbf{x}$) by online initialization of x_{ij}^t (line 11 in **Alg.** 1). Hence, summing up these two parts of the increase of est($\mathbf{x}$), we have $\sum_{t\in[T]} y_0^t + \sum_{t\in[T]} \sum_{j\in[M]} y_j^t + \beta(OPT/\Gamma + \ln(N)) \ge \text{est}(\mathbf{x})$. □

APPENDIX D
PROOF OF LEMMA 4

Proof. Each packing constraint corresponds to one tenant's reduction budget, and according to (11) let $\phi(i)$ denote the index of updating round that decides z_i, and let L denote the set including all round indices l. Therefore,

$$\phi(i) = \arg\max_{l\in L} \frac{\exp(\sum_{\bar{t}\in[T]} \sum_{\bar{j}\in[M]} \frac{e_{i\bar{j}}^{\bar{t}}}{B_i\Gamma}(\mathbf{x}^{(l)})_{i\bar{j}}^{\bar{t}})}{\sum_{\bar{i}\in[N]} \exp(\sum_{\bar{t}\in[T]} \sum_{\bar{j}\in[1,M]} \frac{e_{\bar{i}\bar{j}}^{\bar{t}}}{B_{\bar{i}}\Gamma} x_{\bar{i}\bar{j}}^{\bar{t}})}.$$

Then, we reorder the tenants in order to ensure $\phi(1) \le \phi(2) \le \cdots \le \phi(N)$, and we know that for any $r, k \in [N]$ with $r \le k$ we have $\phi(r) \le \phi(k)$ and $\exp(\sum_{\bar{t}\in[T]} \sum_{\bar{j}\in[1,M]} \frac{e_{r\bar{j}}^{\bar{t}}}{B_r\Gamma} x_{r\bar{j}}^{\bar{t}}) \le \exp(\sum_{\bar{t}\in[T]} \sum_{\bar{j}\in[1,M]} \frac{e_{r\bar{j}}^{\bar{t}}}{B_r\Gamma} x_{r\bar{j}}^{\bar{t}})$—*i.e.*, the values of the terms in later rounds are always no less than that in the early rounds. Thus, we have

$$\begin{aligned}
&\sum_{r\in[N]} \exp(\sum_{\bar{t}\in[T]} \sum_{\bar{j}\in[1,M]} \frac{e_{r\bar{j}}^{\bar{t}}}{B_r\Gamma} x_{r\bar{j}}^{\bar{t}}) \\
&\ge \sum_{r\le k} \exp(\sum_{\bar{t}\in[T]} \sum_{\bar{j}\in[1,M]} \frac{e_{r\bar{j}}^{\bar{t}}}{B_r\Gamma} x_{r\bar{j}}^{\bar{t}}) \qquad (15) \\
&\ge \sum_{r\le k} \exp(\sum_{\bar{t}\in[T]} \sum_{\bar{j}\in[1,M]} \frac{e_{r\bar{j}}^{\bar{t}}}{B_r\Gamma} x_{r\bar{j}}^{\bar{t}}).
\end{aligned}$$

According to (11) and (15), we have

$$\begin{aligned}
(\ln(eN) + \frac{\lambda^{max}}{\Gamma}) z_k &= \frac{\exp(\sum_{\bar{t}\in[T]} \sum_{\bar{j}\in[1,M]} \frac{e_{k\bar{j}}^{\bar{t}}}{B_k\Gamma} x_{k\bar{j}}^{\bar{t}})}{\sum_{r\in[N]} \exp(\sum_{\bar{t}\in[T]} \sum_{\bar{j}\in[1,M]} \frac{e_{r\bar{j}}^{\bar{t}}}{B_r\Gamma} x_{r\bar{j}}^{\bar{t}})} \\
&\le \frac{\exp(\sum_{\bar{t}\in[T]} \sum_{\bar{j}\in[1,M]} \frac{e_{k\bar{j}}^{\bar{t}}}{B_k\Gamma} x_{k\bar{j}}^{\bar{t}})}{\sum_{r\le k} \exp(\sum_{\bar{t}\in[T]} \sum_{\bar{j}\in[1,M]} \frac{e_{r\bar{j}}^{\bar{t}}}{B_r\Gamma} x_{r\bar{j}}^{\bar{t}})}.
\end{aligned}$$

To simplify the analysis, let a_i denote $\exp(\sum_{\bar{t}\in[T]} \sum_{\bar{j}\in[1,M]} \frac{e_{i\bar{j}}^{\bar{t}}}{B_i\Gamma} x_{i\bar{j}}^{\bar{t}})$, and then we utilize Lemma 8 (in Appendix I). We have

$$\begin{aligned}
&(\ln(eN) + \frac{\lambda^{max}}{\Gamma}) \sum_{i\in[N]} z_i \\
&\le 1 + \ln(\frac{\sum_{i\in[N]} \exp(\sum_{\bar{t}\in[T]} \sum_{\bar{j}\in[1,M]} \frac{e_{i\bar{j}}^{\bar{t}}}{B_i\Gamma} x_{i\bar{j}}^{\bar{t}})}{\exp(\sum_{\bar{t}\in[T]} \sum_{\bar{j}\in[1,M]} \frac{e_{1\bar{j}}^{\bar{t}}}{B_1\Gamma} x_{1\bar{j}}^{\bar{t}})}) \\
&\le \ln(eN) + \max_l \tilde{\lambda}(\mathbf{x}^{(l)})
\end{aligned}$$

where the last inequality holds due to $\ln(eN) \ge 1$, $\exp(\sum_{\bar{t}\in[T]} \sum_{\bar{j}\in[1,M]} \frac{e_{1\bar{j}}^{\bar{t}}}{B_1\Gamma} x_{1\bar{j}}^{\bar{t}}) \ge 1$, and $\ln(\exp(\sum_{\bar{t}\in[T]} \sum_{\bar{j}\in[1,M]} \frac{e_{i\bar{j}}^{\bar{t}}}{B_i\Gamma} x_{i\bar{j}}^{\bar{t}})) \le \max_{l\in L} \frac{1}{\Gamma}\lambda(\mathbf{x}^{(l)})$. Hence, we have $\sum_{i\in[N]} z_i - \xi \le 1$ where ξ is set as $\xi = 0$. □

APPENDIX E
PROOF OF LEMMA 5

Proof. Consider a round l executed upon arrival of a power emergency signal (*i.e.*, a covering constraint). In this round, y_j^t is incremented by $(e\epsilon_j^t)$. Let L_j^t and L_0^t denote the set of rounds during PDU emergency R_j^t and UPS emergency R_0^t, and the increment occurs in every round in L_j^t. Hence,

$$\frac{e_{ij}^t}{R_j^t} y_j^t + \frac{e_{ij}^t}{R_0^t} y_0^t = e(\frac{e_{ij}^t}{R_j^t} \sum_{l\in L_j^t} \epsilon_j^t(\mathbf{x}^{(l)}) + \frac{e_{ij}^t}{R_0^t} \sum_{l\in L_0^t} \epsilon_0^t(\mathbf{x}^{(l)})),$$

where $\epsilon_j^t(\mathbf{x}^{(l)})$ is the value of ϵ_j^t computed in line 15 in **Alg.** 1 based on $\mathbf{x}^{(l)}$. As the initial value of x_{ij}^t is $\frac{1}{\kappa\rho\delta}$, and x_{ij}^t is updated by $(1 + \epsilon_j^t(\mathbf{x}^{(l)}) \frac{e_{ij}^t}{R_j^t \text{rate}_{ij}^t})$ in each round; for any $i \in [N], j \in [M], t \in [T]$,

$$
\begin{aligned}
&\mu \geq \mu x_{ij}^t \\
&\geq \frac{1}{\kappa\rho\delta}\Big(\prod_{l\in L_0^t}(1+\epsilon_0^t(\mathbf{x}^{(l)})\frac{e_{ij}^t}{R_0^t \text{rate}_{ij}^t})\Big)\Big(\prod_{l\in L_j^t}(1+\epsilon_j^t(\mathbf{x}^{(l)})\frac{e_{ij}^t}{R_j^t \text{rate}_{ij}^t})\Big) \\
&\geq \frac{1}{\kappa\rho\delta}\Big(\prod_{l\in L_0^t}(1+\epsilon_0^t(\mathbf{x}^{(l)})\frac{e_{ij}^t}{R_0^t \text{rate}_{ij}^t})\Big)\Big(\prod_{l\in L_j^t}\exp(\ln(2)\epsilon_j^t(\mathbf{x}^{(l)})\frac{e_{ij}^t}{R_j^t \text{rate}_{ij}^t}\Big) \\
&\geq \frac{1}{\kappa\rho\delta}\prod_{l\in L_0^t}\exp(\ln(2)\epsilon_0^t(\mathbf{x}^{(l)})\frac{e_{ij}^t}{R_0^t \text{rate}_{ij}^t})\prod_{l\in L_j^t}\exp(\ln(2)\epsilon_j^t(\mathbf{x}^{(l)})\frac{e_{ij}^t}{R_j^t \text{rate}_{ij}^t}),
\end{aligned}
$$

where in the last two inequalities we replace $(1+\epsilon_0^t(\mathbf{x}^{(l)})\frac{e_{ij}^t}{R_0^t \text{rate}_{ij}^t})$ and $(1+\epsilon_j^t(\mathbf{x}^{(l)})\frac{e_{ij}^t}{R_j^t \text{rate}_{ij}^t})$ with $\exp(\ln(2)\epsilon_0^t(\mathbf{x}^{(l)})\frac{e_{ij}^t}{R_0^t \text{rate}_{ij}^t})$ and $\exp(\ln(2)\epsilon_j^t(\mathbf{x}^{(l)})\frac{e_{ij}^t}{R_j^t \text{rate}_{ij}^t})$ respectively. According to line 10 and line 15 in **Alg.** 1, we have $\epsilon_j^t(\mathbf{x}^{(l)})\frac{e_{ij}^t}{R_j^t \text{rate}_{ij}^t} \leq \frac{1}{(1+2\beta)\ln(eN)} \leq 1$. Moreover, for any $0 \leq \varrho \leq A$, we have $e^{\frac{\ln(1+A)}{A}\varrho} \leq 1+\varrho$ where $A=1$ and $\varrho = \epsilon_j^t(\mathbf{x}^{(l)})\frac{e_{ij}^t}{R_j^t \text{rate}_{ij}^t}$.

Next, we multiply both sides by $(\kappa\rho\delta)$ and take the natural logarithm, and then we have

$$
\ln(\mu\kappa\rho\delta) \geq \ln(2)\Big(\sum_{l\in L_0^t}(\epsilon_0^t(\mathbf{x}^{(l)})\frac{e_{ij}^t}{R_0^t \text{rate}_{ij}^t}) + \sum_{l\in L_j^t}(\epsilon_j^t(\mathbf{x}^{(l)})\frac{e_{ij}^t}{R_j^t \text{rate}_{ij}^t})\Big).
$$

After rearranging some terms and multiplying $\max_{l\in\{L_0^t\cup L_j^t\}}\{\text{rate}_{ij}^t(\mathbf{x}^{(l)})\}$ on both sides, we have

$$
\begin{aligned}
&\frac{1}{\ln(2)}\ln(\mu\kappa\rho\delta)\cdot\max_{l\in\{L_0^t\cup L_j^t\}}\{\text{rate}_{ij}^t(\mathbf{x}^{(l)})\} \\
&\qquad\geq \frac{e_{ij}^t}{R_0^t}\sum_{l\in L_0^t}\epsilon_0^t(\mathbf{x}^{(l)}) + \frac{e_{ij}^t}{R_j^t}\sum_{l\in L_j^t}\epsilon_j^t(\mathbf{x}^{(l)}).
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
\frac{e_{ij}^t}{R_j^t}y_j^t + \frac{e_{ij}^t}{R_0^t}y_0^t &\leq \frac{e}{\ln(2)}\ln(\mu\kappa\rho\delta)\cdot\max_{l\in\{L_0^t\cup L_j^t\}}\{\text{rate}_{ij}^t(\mathbf{x}^{(l)})\} \\
&\leq \frac{e}{\ln(2)}\ln(\mu\kappa\rho\delta)\cdot\frac{e_{ij}^t}{B_i}z_i\cdot(\ln(eN)+\frac{\lambda^{max}}{\Gamma}) \\
&\leq (\frac{e_{ij}^t}{B_i}z_i + \phi_{ij}^t)\cdot\sigma(\ln(eN)+\frac{\lambda^{max}}{\Gamma})
\end{aligned}
$$

where the first inequality follows by line 21 and 36 in **Alg.** 1, the second inequality follows by (11), and last inequality follows by $\phi_{ij}^t \geq 0$ and $1<\mu<2$. □

APPENDIX F
PROOF OF THEOREM 1

Proof. Let $\mathbf{x}^*$ and $(\mathbf{y}^*,\mathbf{z}^*,\boldsymbol{\phi}^*,\xi^*)$ denote the optimal primal and dual solution, and let $\mathbf{x}^f$ and $(\mathbf{y}^f,\mathbf{z}^f,\boldsymbol{\phi}^f,\xi^f)$ denote the primal and dual solution obtained by our algorithm. We have

$$
\begin{aligned}
OPT &= \lambda(\mathbf{x}^*) \\
&= \sum_{t\in[T]}(\mathbf{y}^*)_0^t + \sum_{t\in[T]}\sum_{j\in[M]}(\mathbf{y}^*)_j^t - \sum_{t\in[T]}\sum_{j\in[M]}\sum_{i\in[N]}(\boldsymbol{\phi}^*)_{ij}^t - \xi^*
\end{aligned}
$$

by the LP strong duality. In our solution, the dual variables $(\mathbf{y}^f,\mathbf{z}^f,\boldsymbol{\phi}^f,\xi^f)$ might not be feasible. Then, according to Lemma 4 and 5, $(\mathbf{y}^f/(\sigma\Omega),\mathbf{z}^f,\mathbf{0},0)$ is a feasible dual solution—$\mathbf{y}^f/(\sigma\Omega)$ represents that each element of $\mathbf{y}$ is divided by $(\sigma\Omega)$—where $\Omega = \ln(eN)+\frac{\lambda^{max}}{\Gamma} = \ln(eN)+\tilde{\lambda}(\mathbf{x}^f)$ and $\tilde{\lambda}(\mathbf{x}) = \frac{\lambda^{max}}{\Gamma}$. Due to weak duality, the optimal primal objective value is always no less than dual objective values of feasible dual solutions. Hence, we have

$$
\begin{aligned}
OPT &\geq \frac{1}{\sigma\Omega}\Big(\sum_{t\in[T]}y_0^t + \sum_{t\in[T]}\sum_{j\in[M]}y_j^t\Big) \\
&\geq \frac{\Gamma}{\sigma\Omega}\Big(\frac{\text{est}(\mathbf{x})}{\Gamma} - \beta OPT/\Gamma - \beta\ln(N)\Big).
\end{aligned}
$$

After rearranging terms, we have

$$
\frac{\sigma\Omega}{\Gamma}OPT + \frac{\beta}{\Gamma}OPT + \beta\ln(N) \geq \frac{\text{est}(\mathbf{x})}{\Gamma},
$$

and we substitute Ω and have

$$
\frac{\sigma}{\Gamma}OPT(\ln(eN)+\tilde{\lambda}(\mathbf{x}^f)) + \frac{\beta}{\Gamma}OPT + \beta\ln(N) \geq \tilde{\lambda}(\mathbf{x}^f). \quad (16)
$$

As $\Gamma = 2\sigma \text{U}_{\text{OPT}}$ and $\text{U}_{\text{OPT}} \geq OPT \geq \frac{1}{\alpha}\text{U}_{\text{OPT}}$, we have

$$
2\sigma\alpha OPT \geq \Gamma \geq 2\sigma OPT. \quad (17)
$$

Then, for **Eq.** (16), we replace Γ with its lower bound (*i.e.*, $2\sigma OPT$) and as $\sigma > 1$ we have

$$
\begin{aligned}
\frac{1}{2}(\ln(eN)+\tilde{\lambda}(\mathbf{x}^f)) + \frac{\beta}{2} + \beta\ln(N) &\geq \tilde{\lambda}(\mathbf{x}^f) \\
\ln(eN) + \beta + 2\beta\ln(N) &\geq \tilde{\lambda}(\mathbf{x}^f) \\
(1+2\beta)\ln(eN) &\geq \tilde{\lambda}(\mathbf{x}^f),
\end{aligned}
$$

which shows that $\tilde{\lambda}(\mathbf{x}^f)$ is always no more than $(1+2\beta)\ln(eN)$ which is consistent with Lemma 7 during the proof of competitive ratio.

According to $\tilde{\lambda}(\mathbf{x}^f) \leq (1+2\beta)\ln(eN)$ and the upper bound of Γ (*i.e.*, $2\sigma\alpha OPT$), for (16), we have

$$
\begin{aligned}
&\sigma OPT(\ln(eN)+\tilde{\lambda}(\mathbf{x}^f)) + \beta OPT + \Gamma\beta\ln(N) \geq \Gamma\tilde{\lambda}(\mathbf{x}^f) = \lambda(\mathbf{x}) \\
&\left(\sigma\ln(eN) + \sigma(1+2\beta)\ln(eN) + \beta + 2\sigma\alpha\beta\ln(N)\right)OPT \geq \lambda(\mathbf{x}).
\end{aligned}
$$

According to $2\sigma\alpha > 1$, we replace that β term with $2\sigma\alpha\beta$ and have

$$
\begin{aligned}
\left((2\sigma+2\sigma\beta)\ln(eN) + 2\sigma\alpha\beta\ln(eN)\right)OPT &\geq \lambda(\mathbf{x}) \\
2\sigma\left(1+\beta+\alpha\beta\right)\ln(eN)OPT &\geq \lambda(\mathbf{x})
\end{aligned}
$$

□

APPENDIX G
PROOF OF LEMMA 6

Lemma 6. *$\beta(OPT/\Gamma + \ln(N))$ bounds the increase of* $\text{est}(\mathbf{x})$ *by online initialization of* x_{ij}^t.

Proof. For simplification, we divide the value of the estimation function $\text{est}(\mathbf{x})$ into two parts: one part is the λ, which is the objective value of (6) (*i.e.*, $\lambda = \max_{i\in[N]}\{\frac{\sum_{t\in[T]}\sum_{j\in[1,M]} e_{ij}^t x_{ij}^t}{B_i}\}$); another part is equal to $\text{est}(\mathbf{x}) - \lambda$, which represents the estimation error. According to (9), the increase of $\text{est}(\mathbf{x})$ caused by initialization is equal to the increase of λ plus the increase of the estimation error. Accumulating the whole initialization processes, according to (13), the total increase of λ caused by initialization is up to U_{OPT} (*i.e.*, the upper bound of OPT); and according to (9), the estimation error is up to $\Gamma\ln(N)$. Therefore, the increase of $\text{est}(\mathbf{x})$ during initialization is less than $\text{U}_{\text{OPT}} + \Gamma\ln(N) \le \beta(OPT + \Gamma\ln(N))$ where $\beta = \frac{1+2\sigma\ln(N)}{\frac{1}{\alpha}+2\sigma\ln(N)} = \frac{\text{U}_{\text{OPT}}+\Gamma\ln(N)}{\frac{1}{\alpha}\text{U}_{\text{OPT}}+\Gamma\ln(N)}$ (Note that U_{OPT} and $\frac{1}{\alpha}\text{U}_{\text{OPT}}$ represent the upper bound of OPT and lower bound of OPT respectively). □

APPENDIX H
PROOF OF LEMMA 7

Lemma 7. *Given x_{ij}^t and $x_{ij}^t{}'$ and $0 \le x_{ij}^t \le x_{ij}^t{}' \le 1$, we have* $\text{rate}_{ij}^t(\mathbf{x}') \le e \cdot \text{rate}_{ij}^t(\mathbf{x})$.

Proof. By definition of rate_{ij}^t in (10), and as $x_{ij}^t \le x_{ij}^t{}' \le \mu x_{ij}^t$, we have

$$\begin{aligned}
\text{rate}_{ij}^t(\mathbf{x}') &= \frac{\frac{e_{ij}^t}{B_i}\exp\left(\sum_{\bar{t}\in[T]}\sum_{\bar{j}\in[M]}\frac{e_{i\bar{j}}^{\bar{t}}}{B_i\Gamma}x_{i\bar{j}}^{\bar{t}}{}'\right)}{\sum_{\bar{i}\in[N]}\exp\left(\sum_{\bar{t}\in[T]}\sum_{\bar{j}\in[M]}\frac{e_{\bar{i}\bar{j}}^{\bar{t}}}{B_{\bar{i}}\Gamma}x_{\bar{i}\bar{j}}^{\bar{t}}{}'\right)} \\
&\le \frac{\frac{e_{ij}^t}{B_i}\exp\left(\sum_{\bar{t}\in[T]}\sum_{\bar{j}\in[M]}\frac{e_{i\bar{j}}^{\bar{t}}}{B_i\Gamma}x_{i\bar{j}}^{\bar{t}}{}'\right)}{\sum_{\bar{i}\in[N]}\exp\left(\sum_{\bar{t}\in[T]}\sum_{\bar{j}\in[M]}\frac{e_{\bar{i}\bar{j}}^{\bar{t}}}{B_{\bar{i}}\Gamma}x_{\bar{i}\bar{j}}^{\bar{t}}\right)} \\
&\le \frac{\frac{e_{ij}^t}{B_i}\exp\left(\sum_{\bar{t}\in[T]}\sum_{\bar{j}\in[M]}\mu\frac{e_{i\bar{j}}^{\bar{t}}}{B_i\Gamma}x_{i\bar{j}}^{\bar{t}}\right)}{\sum_{\bar{i}\in[N]}\exp\left(\sum_{\bar{t}\in[T]}\sum_{\bar{j}\in[M]}\frac{e_{\bar{i}\bar{j}}^{\bar{t}}}{B_{\bar{i}}\Gamma}x_{\bar{i}\bar{j}}^{\bar{t}}\right)},
\end{aligned}$$

As for any $i \in [N]$, $\tilde{\lambda}(\mathbf{x}) \le (1+2\beta)\ln(eN) \Rightarrow (\sum_{\bar{t}\in[T]}\sum_{\bar{j}\in[M]}\frac{e_{i\bar{j}}^{\bar{t}}}{B_i\Gamma}x_{i\bar{j}}^{\bar{t}}) \le (1+2\beta)\ln(eN)$ and the definition of μ, we have $(\sum_{\bar{t}\in[T]}\sum_{\bar{j}\in[M]}\mu\frac{e_{i\bar{j}}^{\bar{t}}}{B_i\Gamma}x_{i\bar{j}}^{\bar{t}}) \le (\sum_{\bar{t}\in[T]}\sum_{\bar{j}\in[M]}\frac{e_{i\bar{j}}^{\bar{t}}}{B_i\Gamma}x_{i\bar{j}}^{\bar{t}}) + (\sum_{\bar{t}\in[T]}\sum_{\bar{j}\in[M]}\frac{e_{i\bar{j}}^{\bar{t}}}{B_i\Gamma}x_{i\bar{j}}^{\bar{t}})/((1+2\beta)\ln(eN)) \le (\sum_{\bar{t}\in[T]}\sum_{\bar{j}\in[M]}\frac{e_{i\bar{j}}^{\bar{t}}}{B_i\Gamma}x_{i\bar{j}}^{\bar{t}}) + 1$, then we have

$$\begin{aligned}
\text{rate}_{ij}^t(\mathbf{x}') &\le \frac{\frac{e_{ij}^t}{B_i}\exp\left(1+\sum_{\bar{t}\in[T]}\sum_{\bar{j}\in[M]}\frac{e_{i\bar{j}}^{\bar{t}}}{B_i\Gamma}x_{i\bar{j}}^{\bar{t}}\right)}{\sum_{\bar{i}\in[N]}\exp\left(\sum_{\bar{t}\in[T]}\sum_{\bar{j}\in[M]}\frac{e_{\bar{i}\bar{j}}^{\bar{t}}}{B_{\bar{i}}\Gamma}x_{\bar{i}\bar{j}}^{\bar{t}}\right)} \\
&\le \mathrm{e}\cdot\frac{\frac{e_{ij}^t}{B_i}\exp\left(\sum_{\bar{t}\in[T]}\sum_{\bar{j}\in[M]}\frac{e_{i\bar{j}}^{\bar{t}}}{B_i\Gamma}x_{i\bar{j}}^{\bar{t}}\right)}{\sum_{\bar{i}\in[N]}\exp\left(\sum_{\bar{t}\in[T]}\sum_{\bar{j}\in[M]}\frac{e_{\bar{i}\bar{j}}^{\bar{t}}}{B_{\bar{i}}\Gamma}x_{\bar{i}\bar{j}}^{\bar{t}}\right)} \\
&\le e\cdot\text{rate}_{ij}^t(\mathbf{x}),
\end{aligned}$$

proving the lemma. □

APPENDIX I
PROOF OF LEMMA 8

Lemma 8. *For $a_1, a_2, \ldots, a_N \in \mathbb{R}_+$ with $N > 0$, we have* $\sum_{i\in[N]}\frac{a_i}{\sum_{j\le i}a_j} \le 1 + \ln(\frac{\sum_{i\in[N]}a_i}{a_1})$

Proof. It is trivially true when $N = 1$. For $N > 1$, for simplifying, we define $b_i = \sum_{j\le i}a_j$ and so $a_i = b_i - b_{i-1}$, thus we have

$$\sum_{i\in[N]}\frac{a_i}{\sum_{j\le i}a_j} = 1 + \sum_{i\in[2,N]}(1 - \frac{b_{i-1}}{b_i}) \tag{18}$$

For simplification, and we denote $r_i = \frac{b_i}{b_{i+1}}$. We next analyze the second term of (18), and we define $\chi = \sum_{i\in[1,N-1]}(1-r_i) = (N-1) - \sum_{i\in[1,N-1]}r_i$. Since $r_i \le 1$ and $\prod_{i\in[1,N-1]}r_i = \frac{b_1}{b_N}$ (in the production, the denominator of one term cancels the numerator of the next term), the term $\sum_{i\in[1,N-1]}r_i$ is minimized when $r_i = \frac{b_1}{b_N}^{\frac{1}{N-1}}$ and the minimum value (of term $\sum_{i\in[1,N-1]}r_i$) is $(N-1)(\frac{b_1}{b_N})^{\frac{1}{N-1}}$. Therefore, we have

$$\chi \le (N-1) - (N-1)(\frac{b_1}{b_N})^{\frac{1}{N-1}}. \tag{19}$$

For obtaining the upper bound of χ (*i.e.*, the RHS of (19)), we denote $\Theta = (N-1) - (N-1)\vartheta^{\frac{1}{N-1}}$ where $\vartheta = \frac{b_1}{b_N}$. Differentiating Θ with respect to ϑ, we have

$$\begin{aligned}
\frac{\partial\Theta}{\partial\vartheta} &= \vartheta^{\frac{1}{N-1}}\frac{\ln(\vartheta)}{N-1} - \vartheta^{\frac{1}{N-1}} + 1 \\
\frac{\partial^2\Theta}{\partial\vartheta^2} &= -\frac{\vartheta^{\frac{1}{N-1}}\ln^2(\vartheta)}{(N-1)^3} < 0
\end{aligned}$$

As the second partial derivative of Θ with respect to ϑ (*i.e.*, $\frac{\partial^2\Theta}{\partial\vartheta^2}$) is always less than zero, Θ is maximized when $\frac{\partial\Theta}{\partial\vartheta} = 0$ that is $\vartheta^{\frac{1}{N-1}}\ln(\vartheta) - (N-1)\vartheta^{\frac{1}{N-1}} + (N-1) = 0 \Rightarrow (N-1) - (N-1)\vartheta^{\frac{1}{N-1}} = \ln(\frac{1}{\vartheta})$, which is consistent with the RHS of (19). As $\vartheta = \frac{b_1}{b_N} \le 1$, we have

$$\chi \le \ln(\frac{1}{\vartheta}) = \ln(\frac{b_N}{b_1}) = \ln(\frac{\sum_{i\in[N]}a_i}{a_1}). \tag{20}$$

According to (18), (19), and (20), we have

$$\sum_{i\in[N]}\frac{a_i}{\sum_{j\le i}a_j} \le 1 + \ln(\frac{\sum_{i\in[N]}a_i}{a_1})$$

□

APPENDIX J
PROOF OF THEOREM 2

Proof. Algorithm 1 is an iterate program; and our algorithm iteratively updates x_{ij}^t at line 18 and line 33. And the initial value of x_{ij}^t is $\frac{1}{\kappa\rho\delta}$. According to the definitions of ϵ_j^t, μ, and β, and $\beta \le \alpha$, the maximum updating times of one variable x_{ij}^t are up to $\frac{\log(\kappa\rho\delta)}{\log(1+\frac{1}{(1+2\alpha)\ln(eN)})}$. Moreover, for these two sub-functions shown in Algorithm 2 and Algorithm 3, is both $O(NM)$, looking up the best candidate from all tenants over all PDUs. Therefore, the time-complexity of algorithm is $O(NM\frac{\log(\kappa\rho\delta)}{\log(1+\frac{1}{(1+2\alpha)\ln(eN)})})$ □