Heat Behind the Meter: A Hidden Threat of Thermal Attacks in Edge Colocation Data Centers

Zhihui Shao
University of California, Riverside
zshao006@ucr.edu

Mohammad A. Islam
University of Texas at Arlington
mislam@uta.edu

Shaolei Ren
University of California, Riverside
sren@ece.ucr.edu

***Abstract*—The widespread adoption of Internet of Things and latency-critical applications has fueled the burgeoning development of edge colocation data centers (a.k.a., edge colocation) — small-scale data centers in distributed locations. In an edge colocation, multiple entities/tenants house their own physical servers together, sharing the power and cooling infrastructures for cost efficiency and scalability. In this paper, we discover that the sharing of cooling systems also exposes edge colocations' potential vulnerabilities to cooling load injection attacks (called thermal attacks) by an attacker which, if left at large, may create thermal emergencies and even trigger system outages. Importantly, thermal attacks can be launched by leveraging the emerging architecture of built-in batteries integrated with servers that can conceal the attacker's actual server power (or cooling load). We consider both one-shot attacks (which aim at creating system outages) and repeated attacks (which aim at causing frequent thermal emergencies). For repeated attacks, we present a foresighted attack strategy which, using reinforcement learning, learns on the fly a good timing for attacks based on the battery state and benign tenants' load. We also combine prototype experiments with simulations to validate our attacks and show that, for a small 8kW edge colocation, an attacker can potentially cause significant losses. Finally, we suggest effective countermeasures to the potential threat of thermal attacks.**

I. Introduction

In the wake of the Internet of Things and ubiquitous computing demand, edge computing has recently emerged as a game-changing paradigm that brings computation to the Internet edge, thereby enabling ultra-low latencies for many critical applications such as augmented reality and assisted driving [1]. Consequently, the rise of edge computing spurs the burgeoning development of multi-tenant edge colocation data centers (a.k.a., edge colocation). An edge colocation is a small-scale shared colocation data center built at numerous distributed locations for hosting latency-ultrasensitive workloads such as assisted driving [2]. In such a colocation, the operator provides power and cooling resources to multiple entities (i.e., tenants) for housing their own physical servers. Thus, this fundamentally differs from a multi-tenant cloud platform where users/tenants share the cloud resources without *owning* the physical servers.

Edge colocations have become the preferred choice for edge service providers. For example, Vapor IO, an edge colocation operator, is rolling out thousands of edge colocations in partnership with wireless tower companies [3]. Moreover, a recent Uptime Institute survey [4] shows that more than 75% of the respondents will use edge colocations to house their physical servers and deploy edge applications.

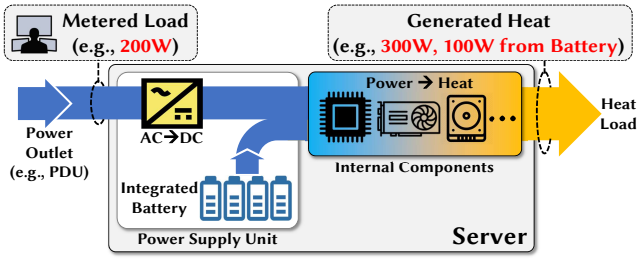


Fig. 1. An attacker uses its built-in batteries to stealthily inject additional heat to overload the cooling system.

The criticality of hosted applications, such as assisted driving [3], clearly mandates a high level of security for edge colocations. While securing servers and networks from cyber attacks remains a key issue, recent research has also identified critical vulnerabilities in data center physical infrastructures. More concretely, the practice of infrastructure oversubscription exposes data centers to well-timed power load attacks that aim at overloading the power capacity and compromising the data center availability [5]–[8]. Likewise, data center cooling system removes server heat to avoid overheating and hence is also crucial for service uptime. If not properly managed, malicious workloads can create more hot spots that expose servers to an adverse thermal environment and thus more thermal emergencies [9]. Importantly, cooling system has emerged as a leading root cause for downtime incidents in state-of-the-art data centers (e.g., Microsoft's) [10], [11].

To meet the power capacity constraints and avoid outages [12]–[14], the colocation operator has power meters to continuously monitor tenants' server power usage. Meanwhile, power meters are also used as a proxy to measure servers' cooling loads,[1] ensuring that the designed cooling capacity is not violated. The reason is that nearly 100% server power is eventually converted into heat or cooling load [6], [15], [16]. Therefore, with proper heat dissipation, meeting the power capacity constraints also implicitly means meeting the cooling capacity constraints [16].

This work was supported in part by the NSF CNS-1551661 (CAREER).

[1]Heat generated by servers is "cooling load" for the cooling system.

Contributions. In this paper, we study an under-explored threat of thermal attacks — injecting additional cooling loads to overload the cooling system — in an edge colocation. While edge colocations have been generally considered as secure due to tenants' full control of their own servers, we discover that the way tenants' cooling loads are measured (i.e., using power meters as proxies) is potentially vulnerable to thermal attacks. More concretely, as illustrated in Fig. 1, an attacker can tap into the emerging architecture of built-in battery units and generate additional cooling loads (i.e., heat), yet without violating the power capacity enforced by the colocation operator. If left neglected, successful thermal attacks may create significant damages: (1) service outage for benign tenants due to overheating (which we call *one-shot* attack); or (2) more frequent thermal emergencies that result in tenants' performance degradation (which we call *repeated* attacks). While various defenses (e.g., measuring servers' outlet temperatures and air flows) are readily available, they have yet to be included in standard practices for many data centers. As such, despite non-trivial efforts needed by thermal attacks, our study serves as a precaution for strengthening cooling system management in edge colocations.

A common practice in today's colocations is to tightly monitor tenants' power usage as well as their server inlet/outlet temperature. Nonetheless, if other effective defense mechanisms (in Section VII) are not properly implemented, built-in batteries integrated with servers' power supply units can assist an attacker with launching thermal attacks that are difficult to trace. To provide better energy efficiency and reliability [17]–[19], vendors have begun to integrate built-in batteries with servers' power supply units (e.g., Supermicro BBP [17]). Such built-in batteries can conceal the attacker's actual cooling load from the operator's power meters — by discharging built-in batteries to supply additional power, the attacker's servers can consume more actual power and hence generate more heat than the operator measures using power meters. Moreover, this additional cooling load may not be promptly pinpointed by only monitoring the servers' inlet and outlet temperatures. Consequently, indiscernible additional cooling loads can be injected by an attacker to exceed the shared cooling capacity, thus triggering thermal emergencies. While we focus on edge colocation data center, such thermal attacks may be mounted against larger colocation data centers as well, albeit the attacker needs to commit more resources.

Meanwhile, before automatic system shutdown [12], [20], handling thermal emergencies require tenants' power/cooling load reduction through clock rate throttling and/or workload re-routing to other unaffected data centers, which can adversely affect tenants' performance in terms of application response time.

While successful thermal attacks can create an adverse environment for hosting servers in edge colocations, they need non-trivial efforts. As a prerequisite for a good timing, the attacker needs to estimate benign tenants' power/cooling loads based on a voltage side channel [5]. For a one-shot attack with the goal of shutting down an entire edge data center, the attacker can install a large built-in battery and inject sufficient cooling loads continuously, resulting in overheating and triggering automatic system shutdown. For repeated attacks that aim at benign tenants' performance degradation, the attacker needs to repeatedly trigger thermal emergencies by charging and discharging its battery at appropriate times. We propose a foresighted policy based on batch Q-learning that learns on the fly a good timing for repeated attacks based on the battery state and benign tenants' load: thermal attacks are launched only when both the benign tenants' loads are sufficiently high and the remaining battery energy is more than a threshold.

We run prototype experiments to validate the potential threat of thermal attacks. To evaluate the effectiveness of our proposed repeated attack strategies, we run year-long simulations based on computational fluid dynamics (CFD) analysis. Our results demonstrate that for an 8kW edge colocation, an attacker subscribing 10% of the capacity can cause thermal emergencies for more than 3% of the year, degrading benign tenants' performance. Finally, while the existing practices may render edge colocations vulnerable, battery-assisted thermal attacks can be fairly easily detected and nullified using a reasonable amount of efforts. We discuss such defense strategies in Section VII.

In conclusion, while batteries have been exploited (such as for smoothing power demand [21]), our study makes a novel contribution by leveraging servers' built-in batteries for an under-explored malicious purpose — thermal attacks that can potentially result in service outage or performance degradation in edge colocations — and serves as a precaution despite its futility when proper defensive measures are enforced.

II. Preliminaries on Edge Colocations

Colocations represent a critical segment and account for nearly 40% of the total energy consumption by data centers [12], serving almost all industry sectors. To complement their own megascale data centers that are typically built in rural areas, even top-brand companies like Google and Microsoft rely on third-party colocations for better performances due to close proximity to end users [22]. Importantly, in the context of edge computing, colocations play an even more crucial role, as it is not economical for individual companies to fully manage small-scale data centers in numerous locations [23].

The data center capacity includes both power and cooling capacities. Power capacity is quantified by the amount of UPS-protected power (a.k.a. critical power) that is delivered to the servers, excluding other power consumption such as UPS power losses and cooling system power. As nearly 100% server power consumption (except for fan power) is converted into heat or *cooling load*, the cooling system capacity is often sized based on the colocation's power capacity and usually also measured in kilowatt [15], [16], [24]. The data center design may also leave some "headroom" in the cooling capacity to handle, if any, irregular heat generation and/or hot spots due to certain servers generating more heat than expected. In such cases, the cooling system utilization may sometimes still be high because of the increasingly common practice of power

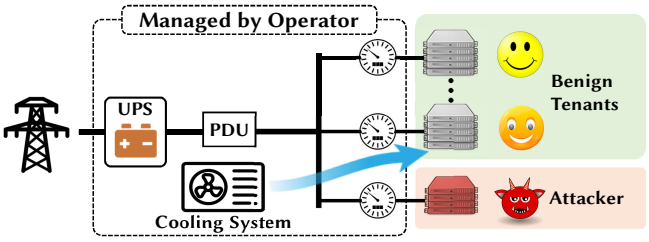


Fig. 2. An edge colocation data center with an attacker.

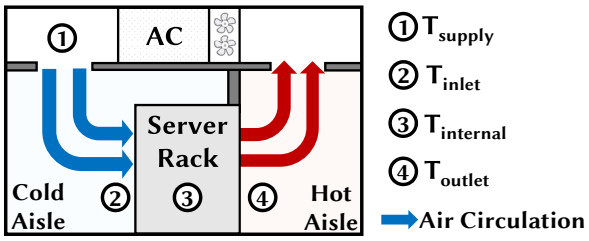


Fig. 3. Overview of a cooling system in an edge colocation.

oversubscription for capital cost saving in modern data centers (e.g., Facebook aggressively oversubscribes its power capacity by 47% on average) [13], [25]–[27].

The colocation operator provides non-IT infrastructure support (i.e., power and cooling systems), while each tenant brings and controls its own physical servers.[2] The non-IT infrastructure is expensive and/or time-consuming to construct, taking nearly 60% of the total cost of ownership over a 10-year lifespan for a colocation operator [12], [15], [21]. Thus, like network bandwidth, the operator's power and cooling infrastructure capacity is a limited resource carefully sized based on the tenants' demand.

A. Power Infrastructure

As illustrated in Fig. 2, typically, an edge colocation data center uses a tree-type power hierarchy with total capacity in the range of a few kilowatts to a few tens of kilowatts shared by multiple tenants. Utility power first enters the data center through an uninterruptible power supply (UPS). Then, the UPS-protected power goes into a power distribution unit (PDU), which distributes the power to its downstream servers.

B. Cooling Infrastructure

While various cooling methods (e.g., computer room air conditioner, chiller, and "free" outside air cooling) are available [28], an edge colocation usually uses a computer room air conditioner to remove servers' heat due to its small size and often rugged deployment (e.g., outdoor with a wireless tower). Fig. 3 illustrates a typical cooling system in an edge colocation. For the best cooling efficiency, today's edge colocations also implement hot/cold aisle containment to prevent the hot air from mixing with the cold air [23], [29].

There are four different notions of temperature in a data center: supply air temperature T_{sup}, server inlet temperature T_{inlet} (i.e., temperature of cold air entering a server), server internal temperature $T_{internal}$ (e.g., CPU temperature), and server outlet temperature T_{outlet} (i.e., temperature of hot air exiting a server). With heat containment installed, all the servers' inlet temperature is nearly identical to the supply air temperature. Thus, supply air temperature and server inlet temperature are the lowest and baseline, whose increase will lead to increases in server internal and outlet temperatures. Server outlet temperature is typically elevated by 10+°C compared to the inlet temperature while server internal temperature is the highest and regulated by servers' internal fans. Hence, with heat containment, we have the following [30]–[32]:

$$T_{inlet} \approx T_{sup} < T_{outlet} < T_{internal}. \quad (1)$$

In a data center, *server inlet temperature is the most important thermal metric* [31], [33], because servers' internal temperature control uses the inlet temperature as a reference [34]. For example, in modern data centers, server inlet temperature is conditioned at 27°C for cooling efficiency, as recommended by ASHRAE [33], [35]. Also note that, while server heat is responsible for increase in internal and outlet temperatures, neither $T_{internal}$ nor T_{outlet} is a reliable indicator for a server's cooling load since they depend on the server's internal heat management (e.g., fan speed) and air flow rate.

III. Thermal Attack

The main focus of this section is to present the potential threat of battery-assisted thermal attacks (when concealed cooling loads are *behind the meter* and not promptly detected) and help strengthen edge colocations. As a precursor, we first introduce our threat model that outlines the scenario considered for thermal attacks. We then present the potential impacts on edge colocations. Finally, we introduce two possible attack strategies followed by discussions on their feasibility.

A. Threat Model

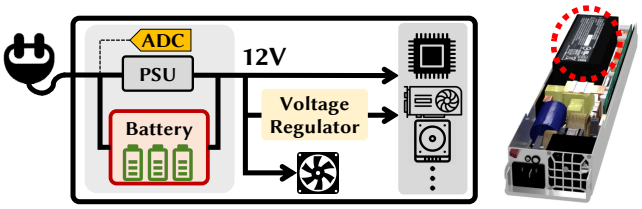
We consider an edge colocation data center with a total power/cooling capacity of C, housing a few racks of servers owned by multiple tenants. There exists a malicious tenant (i.e., attacker) that runs artificial workloads without real values and has bad intentions.

What the attacker can do. The attacker houses its own physical servers in the edge colocation, sharing the power and cooling infrastructures with benign tenants. As illustrated in Figs. 1 and 4(a), the attacker's server power supply units has built-in battery units, which can conceal the attacker's actual server power/cooling load from the operator's power meters. Fig. 4(a) shows an overview of the attacker's server.

The attacker subscribes a data center capacity of c_a from the colocation operator and keeps its power drawn from the operator's PDU below c_a at all times (even during an attack), in order to meet the operator's requirement.

When launching a thermal attack, the attacker runs power-hungry applications (e.g., intensive computation) to increase

[2] A tenant can share fraction of a rack space with other tenants.



(a) (b)

Fig. 4. (a) Attacker's server with built-in batteries. (b) Supermicro's power supply [17]. The built-in battery module is highlighted in a red circle.

its actual server power consumption to $p_a > c_a$ where c_a amount comes from the operator's PDU and the rest from its built-in battery. In practice, when running at the peak load, a single server equipped with multiple CPUs and/or GPUs can easily consume several hundred watts, even more than 1kW [36]. Thus, the attacker can inject an additional cooling load of $p_b = p_a - c_a$ beyond its subscribed capacity by discharging built-in battery units (which can be achieved by a dual-source power supply that can simultaneously draw power from the PDU and the battery units [21], [37]–[40].)

The attacker uses a voltage side channel, as proposed in [5], to estimate benign tenants' real-time total server load with a high accuracy (Fig. 5(b)).

What the attacker cannot do. We do not consider naive attacks, such as self-explosion and tampering with the physical infrastructures, which are beyond the scope of our work. Moreover, other attacks, such as network DDoS attacks, are also orthogonal to our focus.

B. Impact of Thermal Attacks

Although non-trivial efforts are needed in the threat model, a successful thermal attack can overload a data center's cooling system and possibly increase the server inlet temperature to a dangerous level, triggering frequent performance degradation and even system outages [34], [41].

1) Performance degradation: Before system shutdown, a preventative mechanism is to temporarily cap the data center-wide cooling load (i.e., server power) below the cooling capacity [20], [42]. Specifically, when the server inlet temperature exceeds a threshold (e.g., 32°C) for a certain amount of time [20], it is considered that a data center exception, called *thermal emergency*, has occurred and servers are forcibly put in a low power state. The wait-time between inlet temperature violation and thermal emergency declaration depends on operator's risk management policy. The temperature threshold for a thermal emergency is set lower than the server's automatic shutdown temperature to proactively handle an emergency. For example, in a Google-type data center, disk speeds and/or CPUs are throttled to lower the server power load (i.e., cooling load) in the event of a thermal emergency [20], [43]. Similar mechanisms also exist in multi-tenant colocations to handle a thermal emergency. Concretely, without controlling tenants' servers, the operator sends signals to tenants' own server management systems such that tenants can cap power loads below a certain level (a.k.a. power capping). The actual amount and duration of power capping can be either pre-determined based on SLA terms [35] or decided at runtime through a dynamic coordination mechanism [12], [44].

Nonetheless, handling a thermal emergency by capping tenants' server power (through, e.g., CPU throttling) inevitably results in performance degradation, which can in turn cause user dissatisfaction, revenue loss, and/or SLA violation [12], [13], [18], [45]. Some workloads may be re-routed to other unaffected data centers for service continuity, but this comes at a higher latency since otherwise those workloads would have been processed in the preferred site to achieve the best performance without being re-routed.

2) System outage: In order to prevent permanent hardware damage, if the server inlet temperature continues rising despite cooling load capping, automatic system shutdown may occur, leading to a system outage (e.g., the shared PDU can power off when the inlet temperature reaches 45°C) and service interruptions [33]. Such system outages can cause loss of working data sets, and also suffer from long restart waiting time. Financially, a system outage can cost thousands of dollars every minute [10]. For latency-critical applications, an outage event may cause even more catastrophic consequences such as decreased safety in edge-assisted driving [46].

We also run a prototype experiment to demonstrate the potential impact of thermal attacks on benign tenants, and the results are in Appendix A.

C. Attack Strategies

We introduce two possible strategies for battery-assisted thermal attacks with different goals.

One-shot attack. It aims at creating a system outage by increasing the server inlet temperature beyond the safety limit (e.g., 45°C [33]). It can also be coordinated across multiple edge colocations for a wide-area service interruption. Even successfully launched only once, the caused damage may be significant, especially for safety-critical applications (e.g., edge-assisted driving) [46].

Repeated attacks. Instead of aggressively overheating and shutting down the entire edge colocation, repeated attacks aim at frequently degrading performance of benign tenants' latency-sensitive applications over a long period (e.g., one year) by triggering thermal emergencies and cooling load capping. Thus, repeated attacks compromise the long-term cooling system availability in edge colocations.

In general, one-shot attack requires a higher battery capacity to support more intense attack loads (which may still be feasible as shown in Section VI). On the other hand, repeated attacks require relatively less (still a considerable amount of) resource, but they require more sophisticated timing of the attacks and can be easy to detect.

D. Feasibility of Thermal Attacks

Motivation for thermal attacks. One-shot attack is as motivating as traditional DDoS attacks, as it can potentially create service outages. Likewise, repeated attacks can result

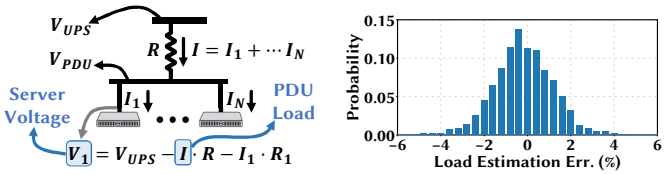
in frequent performance degradation for latency-sensitive applications, which in turn causes user dissatisfaction, revenue loss, and/or SLA violation. Thus, although the cost barrier is non-trivial, battery-assisted thermal attack might still be inviting for potential attackers, such as the target colocation's ill-intentioned competitor or state-sponsored attackers.

Attacker's malicious cooling load. In recent years, vendors have integrated built-in batteries into servers' power supply units as an emerging backup power solution (e.g., Supermicro BBP [17] shown in Fig. 4(b)). Thus, an attacker can discharge built-in batteries to supply additional power to its servers, generating malicious cooling loads without being monitored by the colocation operator's power meters. Moreover, without air flow meters, temperature sensors that only monitor server inlet/outlet temperature cannot reliably locate the malicious cooling load. Consequently, if left neglected, thermal attacks can be launched *behind* the meter. This is also illustrated in Fig. 1: an attacker generates 300W cooling load, but the colocation operator only measures 200W from the power meter and the additional 100W load is supported by the attacker's internal batteries.

Availability of off-the-shelf hardware. Servers with built-in batteries are commercially available (e.g., Supermicro [17]). The current battery energy density is enough to fit into servers and supply sufficient additional power to mask the attacker's malicious cooling loads [47], even for an one-shot attack that requires more attack loads than repeated attacks. Moreover, servers with large peak-to-average ratios are also available for generating a large amount of heat during an attack. For example, Dell manufactured PowerEdge R740/R740xd servers can be equipped with up to three Nvidia Tesla GPUs each with 225W peak and 20W idle power [48], [49].

Voltage side channel to time thermal attacks. Due to time-varying loads, the attacker needs to find a good timing for successful attacks (especially for repeated attacks) when benign tenants' aggregate power load (or cooling load) is high. The attacker can utilize a side channel — voltage side channel in our study — to estimate benign tenants' power draw from the shared PDU. The voltage side channel is robust against changes in the environment and provides high accuracy due to its *wired* signal [5]. Utilizing the voltage side channel requires one analog-to-digital converter (ADC) that can fit on a server's power supply unit (as demonstrated in an orthogonal study for USB-powered IoT devices [50]). As shown in Fig. 4(a), the ADC taps into the server's input voltage to sample the PDU-level voltage.

For the readers' understanding, we show in Fig. 5(a) the fundamental principle behind the voltage side channel as recently proposed in [5]. The key idea is that because of the voltage drop along the shared power cable, the total load information (proportional to current) is contained in the voltage signal, e.g., V_1, entering any servers connected to the PDU. Meanwhile, all today's servers have power factor correction (PFC) circuits that generate high-frequency voltage ripples, whose amplitude is strongly correlated with the server load. Thus, the attacker can sense the incoming voltage signal, extract the voltage ripples,



(a) (b)

Fig. 5. (a) Server voltage carries the servers' load information. (b) Load estimation error of the voltage side channel.

and estimate the total load at runtime.

We run a 24-hour real-world workload trace in our prototype and collect the voltage signal using a NI digital data acquisition (DAQ) as an ADC proxy to extract the servers' total power load. We plot in Fig. 5(b) the probability distribution of load estimation errors, confirming that the voltage side channel can be leveraged for precisely timing thermal attacks.

Possibility of being detected. Detection of battery-assisted thermal attacks is not difficult, but contingent upon the edge colocation operator's practice of environment monitoring. Specifically, if the operator solely relies on power meters for monitoring tenants' loads and temperature sensors for conditioning the thermal environment, thermal attacks may possibly remain undetected until they cause damages. A service outage (due to one-shot attack) or more frequent thermal emergencies (due to repeated attacks) can trigger a thorough inspection, thus exposing the attacker. In order to proactively prevent such damages in advance, as discussed in Section VII, the operator can install additional monitoring apparatus such as server outlet air flow meters, which are not widely used in many data centers. Thus, although thermal attacks do not have a high degree of stealthiness, there is a need of attention to potential thermal attacks.

Relationship to power attacks. Power attacks exploit oversubscribed power capacity and can be launched *without* the need of battery [5]–[8], [51]. On the other hand, our proposed thermal attacks are launched with the help of built-in battery for concealment of malicious cooling loads. Moreover, for repeated attacks, thermal attacks are *stateful* due to battery charging/discharging that results in temporal correlation of battery states, whereas power attacks are *stateless* and can be launched at any time without being constrained by the available battery energy. Thus, our thermal attacks are complementary to power attacks and present a potential threat by leveraging servers' built-in battery for a malicious purpose.

IV. Learning an Attack Policy

An one-shot attack is a special case of repeated attacks if the attacker sets a sufficiently high threshold on benign tenant's load (above which an attack is launched) and greedily use up its large built-in battery energy. Thus, we now study a general repeated attack policy, Foresighted, by formulating it as a discrete-time Markov decision process (MDP) and using reinforcement learning. The repeated attack policy has

a structural property: *attack when both the benign tenants' server load and the battery energy level are sufficiently high.*

A. MDP formulation

We divide the entire time horizon into time slots (e.g., 1 minute each) indexed by $k = 0, 1, 2, \cdots, \infty$, and present our MDP formulation below.

- System state: $s = (b, u) \in \mathcal{S}$
- Action: $a(s) \in \mathcal{A}(s)$
- State transition probabilities: $P(s, a, s')$
- Reward function: $R(s, a, s')$
- Discount factor: $\gamma \in (0, 1)$

The tuple (s, a, s') means that, given an action a, the system state evolves from s to s'. In our problem, the system state includes two sub-states: battery state (the amount of remaining energy b in the batter units) and the attacker's estimated benign tenants' load state u (using a voltage side channel in Section III-D [5]). Note that we consider the estimated load as part of the system state, because the true value of servers' total load is not available to the attacker. We consider three actions: **(1)** charging the battery units; **(2)** launching a thermal attack by running the servers at peak power and discharging batteries; and **(3)** standby, i.e., running dummy workloads without charging or discharging batteries. The battery's charging rate is fixed at the vendor recommended value, while the effective discharging rate (i.e., power actually delivered to servers, excluding battery losses) is set to p_b which, if combined with the attacker's subscribed capacity c_a, can support the attacker's total server power consumption p_a for thermal attacks. The state transitions are governed by benign tenants' load that is exogenous to the attacker and the battery energy evolution which is controlled by the attacker's charging/discharging decision.

We define the attacker's reward function as follows:

$$R(s, a, s') = w \cdot [T(s, a) - T_0]^+ - \beta(a), \tag{2}$$

where $T(s, a)$ is the resulting server inlet temperature, T_0 is the server inlet temperature conditioned by the operator without attacks, $\beta(a)$ is a cost term, and the operator $[\cdot]^+$ means $\max(\cdot, 0)$. Note that the attacker can easily sense the resulting inlet temperature $T(s, a)$, because today's servers have built-in temperature sensors to monitor the server inlet temperature for safety reasons (i.e., if the server inlet temperature is too high, the server may shut down by itself [34]). Clearly, after discharging batteries, the attacker needs to recharge them, which hence draws more energy from the operator's PDU than otherwise. To account for this, we add a normalized *cost* term: $\beta(a) = 1$ during an attack and $\beta(a) = 0$ otherwise. The cost is normalized to 1, because the attacker discharges a fixed amount of energy for each attack. The weight $w \geq 0$ governs the tradeoff between server inlet temperature increase and total battery usage (or attack time): the larger w, the more importance of server inlet temperature increase and hence more attacks.

In a standard MDP, the goal is to find an optimal policy $\pi^* : \mathcal{S} \rightarrow \mathcal{A}$ (i.e., deciding an optimal action given each system state) which maximizes the total discounted reward $\sum_{k=0}^{\infty} \gamma^k R(s_k, a_k, s_{k+1})$. The discount factor $\gamma \in (0, 1)$ is imposed to ensure the convergence of summation and implies in practice that future rewards are relatively less important than immediate rewards [52]. Nonetheless, the resulting server inlet temperature $T(s, a)$ is an involved function that also depends on external factors such as the edge colocation layout, and the dynamics of benign tenants' power usage is unknown to the attacker. Thus, we need an online *learning* approach to identify the optimal policy π^* on the fly.

B. Batch Q-learning

Reinforcement learning can effectively assist an agent with finding optimal actions in an unknown environment. The cooling load state is essentially uncontrollable and exogenous to the attacker. On the other hand, the battery state is fully controllable and, with simplification, can be approximated as $b_{k+1} = \min(b_k + e_k, \bar{B})$, where e_k is the charged energy during one time slot (a negative value means battery discharging for attacks) and $\bar{B}$ is the total battery capacity. Thus, we adopt batch Q-learning [53], by extending the widely-used standard Q-learning [52], [53]. Concretely, by introducing an intermediate state (also called *post state* $\tilde{s}_k$), we have two state transition processes: from s_k to $\tilde{s}_k$, we only update the battery state whose transition, according to the attacker's action, is fully determined; then, from $\tilde{s}_k$ to s_{k+1}, we will update the cooling demand state based on observations. More specifically, for each time slot k, our proposed batch Q-learning works as follows:

$$a_k \leftarrow \arg \max_{a \in \mathcal{A}(s_k)} [Q(s_k, a) + \theta V(\tilde{s}_k(s_k, a))] \tag{3}$$

$$\tilde{s}_k(s_k, a_k) \leftarrow f(s_k, a_k) \tag{4}$$

$$Q(s_k, a_k) \leftarrow (1 - \delta) Q(s_k, a_k) + \delta R(s_k, a_k, s_{k+1}) \tag{5}$$

$$C(s_k) = \max_a [Q(s_k, a) + \gamma V(\tilde{s}_k)] \tag{6}$$

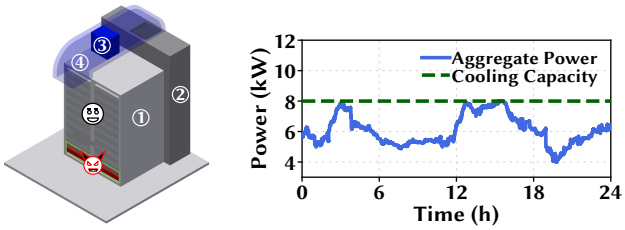
$$V(\tilde{s}_k) = (1 - \delta)\, V(\tilde{s}_k) + \delta C(s_{k+1}) \tag{7}$$

where $\delta \in (0, 1)$ is the learning rate, and only the battery state is updated based on the attacker's charging/discharging action when setting the post state $\tilde{s}_k(s_k, a)$ in Eqn. 4.

Unlike standard Q-learning, three different *value matrixes* are used for batch learning: **state-action value** $Q(s_k, a_k)$, **post-state value** $V(\tilde{s}_k)$, and **normal state value** $C(S_k)$. First, after observing the system state s_k, the attacker makes an action a based on $Q(s_k, a)$ and post-state value $V(\tilde{s}_k(s_k, a))$ according to Eqn. 3. Then, post state s_k can be obtained based on attacker's action. Next, the reward R_k is obtained based on attacker's observed server inlet temperature and its reward function in Eqn. (2). Meanwhile, the next state s_{k+1} is obtained by estimating the cooling state through a voltage side channel as discussed in Section III-D. Thus, the three value matrixes can be updated recursively according to Eqns. (5), (6) and (7), respectively, making the learning process converge more quickly.

TABLE I
LIST OF PARAMETERS WITH THE DEFAULT VALUES.

Parameter	Value
Data Center Capacity	8 kW
Number of Tenants	4
Number of Servers	40
Number of Server Racks	2
Attacker's Capacity (c_a)	0.8 kW
Attacker's Total Battery Capacity (B)	0.2 kWh
Attack Thermal Load from Battery	1 kW
Charging Rate of the Battery	0.2 kW
Temperature Threshold for Emergency (T_{th})	32°C
Q-learning Discount Factor (γ)	0.99
Q-learning Learning Rate ($\delta(t)$)	$1/t^{0.85}$



(a) (b)

Fig. 6. (a) Data center layout. ① Server racks. ② Heat containment. ③ Air conditioner. ④ Supply air duct. (b) 24-hour snapshot of the power trace.

V. EVALUATION METHODOLOGY

In this section, we first present the default simulation settings and evaluation metrics, and then validate our simulation model.

A. Settings

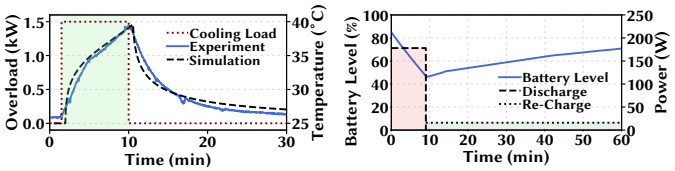
It is practically challenging, if possible at all, to evaluate different thermal attack strategies over a timescale of years. Thus, we resort to a simulation-based approach based on the well-established computational fluid dynamics (CFD) analysis [6], [31], [32], [54] to simulate thermal dynamics. This is also the state-of-the-art methodology in data center-scale research [6], [13], [14], [31]. Prior to simulations, we will also validate our simulation model with real experiments on our scaled-down prototype of 14 servers. We list the default simulation parameters in Table I.

Edge colocation infrastructure. We consider a containerized modular data center design, which is particularly suitable for edge colocations due to its self-contained design. We follow the specification of the Vertiv SmartMod container data center with two server racks, each holding 20 servers [55]. We consider there are four tenants (including the attacker) with a total subscribed power (i.e., the power capacity) of 8kW, where each server's maximum power consumption is 200W. The attacker has 4 servers with a total subscribed capacity of 0.8 kW while the other three benign tenants each subscribe to 2.4kW. The attacker's servers are shown in red shades in Fig. 6(a). Note that, while we place the attacker's servers at the bottom of the rack, their location within the rack does not play any significant role in the attack since the cooling load is determined by server power. The data center employs heat containment for the hot exhaust air returning to the AC. The AC supplies cold air at 27°C, with a cooling capacity of 8kW. Fig. 6(a) shows the layout used in our experiment.

Thermal environment. The CFD analysis provides the most detailed thermal dynamics of a data center (e.g., even Google uses CFD analysis to predict thermal distributions [54]). However, it is computationally exhaustive to run transient CFD analysis for long experiments (e.g., a year) [31]. Therefore, following the literature [6], [31], we model the data center's heat flow using a heat distribution matrix, for which we only need to obtain the matrix parameters using shorter CFD analysis. Specifically, to extract the heat distribution matrix, we test the data center with a heat spike from each server and measure the resulting temperature impact for 10 minutes. We repeat the process for all servers to completely build the matrix. We use the 10-minute window to allow the heat convection through the air and capture the gradual temperature build-up from sustained server heat generation. We limit the CFD analysis of each heat spike to 10 minutes since we find no measurable impact beyond this time horizon. The accuracy of CFD analysis and the heat distribution model has been extensively verified with real systems [31], [56], and will also be validated against our prototype in Section V-B. Because of the well-insulated environment, we do not incorporate the impact of outside temperature. Even with low outside temperature, if overloaded, data center's cooling system cannot remove all server heat.

Attacker. The attacker has built-in batteries integrated with the servers' power supply units. While the capacity of each server subscribed from the operator is 200W, each of the attacker's servers can run at a peak power of 450W by discharging built-in batteries to supply the additional 250W. Thus, the attacker can inject up to 1kW cooling load for repeated attacks. When recharging, the built-in batteries have a total charging rate of 0.2 kW. We use battery specification of [47] with a suitable size for placing inside a server and set the attacker's default total battery capacity to 0.2kWh with 0.05kWh (i.e., 200W for 15 minutes) per server. If the attacker aims at an one-shot attack, each of its four servers has a peak power of 950W, resulting in a total attack load of 3KW. This can be achieved by using multiple power hungry GPUs (e.g., Nvidia RTX 3080, each with a full power of 320W [57]) in each server. The current battery energy density [47] is enough to support the additional load for an one-shot attack, as each attack only lasts a few minutes.

Thermal emergency and system outage. A thermal emergency is considered to arise when the server inlet temperature exceeds 32°C for at least 2 minutes. We consider 32°C as the threshold temperature, because it is the maximum allowed temperature based on the ASHRAE guideline for data centers with enterprise-grade servers and storage [33]. To handle a thermal emergency, each server (including attacker's servers) is required to cap its power below 120W (60% of capacity) to prevent more serious impacts. As a precaution, load capping



(a) (b)

Fig. 7. Experimental validation of our simulation model.

lasts for 5 minutes for each thermal emergency. If the inlet temperature continues rising to reach 45°C, automatic shutdown occurs (e.g., the shared PDU can power off), creating a system outage and service interruptions [33].

Power trace. For the three benign tenants, we use workload traces from Facebook and Baidu [13], [14], and generate a year-long synthetic power trace from request-level log using server power models validated in real systems [58]–[60]. The total power usage is scaled to have a 75% average utilization in our 8kW data center. We show a 24-hour snapshot of the power trace in Fig 6(b). To demonstrate its robustness across different load patterns, we also run an alternate power trace and show in Fig. 13 in Section VI-F.

Application performance. For delay-sensitive workloads, high-percentile latency is the most critical metric [61]. Here, we consider 95-percentile response time as the performance metric and model the tenants' performance based on experiments on our small cluster (Fig. 15 in Appendix A).

Q-learning parameters. Following the literature [62], we set the default discount factor $\gamma = 0.99$ and a dynamic learning rate that is updated everyday using $\delta(t) = 1/t^{0.85}$, where t is the number of days elapsed. We use one minute as each time slot, and show the other parameters when presenting the results in Section VI. To initialize the table of Q values, we use random power traces offline based on an initial attack policy. Our results show that during the online learning stage, the action policy can converge quickly (often within 1-4 weeks).

Evaluation metrics. For the adverse thermal environment, we consider the average server inlet temperature increase, the probability distribution of the temperature, and the total emergency hours due to repeated thermal attacks. For benign tenants, we examine their performance degradation. We also study the average response time during the emergency periods normalized to that of without any emergencies.

B. Experimental Validation of Our Simulation Model

While simulation-based evaluation is widely used in data center research [6], [9], [21], we validate our simulation model using real experiments on our prototype consisting of 14 servers and a 600VA CyberPower UPS battery. We look into the two important aspects of our simulation model — thermal dynamics and battery charging/discharging model.

Temperature dynamics. We place our server rack in a sealed environment with a comparable dimension to an edge data center. The rack is cooled by the building's central cooling system and has air vents on the top. We create an additional 1.5kW thermal overload beyond the limit that can be handled by the top air vents. We obtain the heat distribution model based on CFD analysis. In Fig. 7(a), we show the monitored server inlet temperature change along with our temperature change simulated using our model. We see that both the heat distribution model and temperature sensor readings exhibit very similar dynamics. This is expected since we adopt well-established CFD-based simulation [6], [31].

Battery energy dynamics. In our Q-learning and the simulation, we need to validate that the linear battery model $b_{k+1} = \min(b_k + e_k, \bar{B})$, where b_k is the battery level at time k, is accurate to model the battery energy changes with respect to the charging/discharging decisions. For this, we connect two Dell desktops with a total load of ∼175W to our UPS battery. We connect a power meter between the UPS and the AC power outlet to measure the total power consumption of the battery and the desktops. We connect another power meter between the UPS and the desktops to record the total power of the two desktops. Subtracting the later from the former gives the total power consumption of the UPS. To demonstrate the battery dynamics, we first run the UPS on the battery discharging mode by unplugging it from the AC outlet. After 10 minutes, we reconnect the UPS to the AC outlet, which puts it in the battery charging mode. We show the battery energy levels in Fig. 7(b). In our experiment, the charging rate is lower than the discharging rate, because of the additional UPS loss to power the running desktops. This experiment conforms to our choice of a linear battery energy model. While even more complicated and detailed battery models (e.g., impact of ambient temperature) may be adopted [63], it does not offer much additional insight for our purpose and our observations still hold.

To sum up, our simulation methodology (i.e., using CFD-based analysis for modeling temperature dynamics and using a linear charging/discharing model for battery energy dynamics) matches well with the real-world observations and hence can be used to evaluate thermal attacks with a good confidence.

VI. Evaluation Results

We first show an example of one-shot attack. Then, for repeated attacks, we compare Foresighted with another attack policy, Myopic, that launches thermal attacks in a greedy manner whenever there is enough energy in the battery and the benign tenants' aggregate load is sufficiently high. Besides Myopic and Foresighted, we also consider Random as a benchmark, where the attacker randomly launches thermal attacks whenever it has enough battery energy without considering benign tenants' power loads.

A. Thermal Attack Demonstration

1) One-shot attack: We consider a 30-minute snapshot and demonstrate an one-shot attack in Fig. 8 where the attacker injects 3kW of intense attack load at around the 18th minute, causing the server inlet temperature to rise quickly. At around the 21st minute, a thermal emergency is triggered and power capping is applied, limiting the total metered load below 5KW.

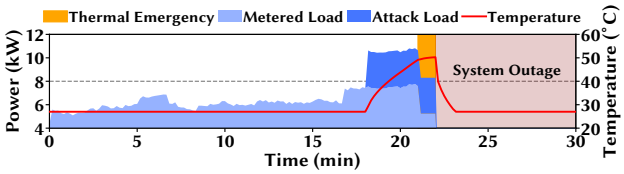


Fig. 8. Demonstration of a one-shot attack.

Nonetheless, the attack load remains to keep the server inlet temperature high enough beyond the safety threshold of 45°C [33], successfully resulting in a system outage. This is also consistent with other orthogonal studies that demonstrate a very quick rise of inlet temperature in case of a cooling system malfunction [41]. If the one-shot attack is coordinated across multiple colocations, a service interruption may occur and create significant damages.

2) Repeated attacks: We illustrate how repeated attacks create emergencies under different attack policies in Fig. 9 by considering a four-hour snapshot when the total power/cooling load is relatively higher. In our illustration, Random launches attacks for 8% of the times, Myopic sets the attack threshold at 7.4kW, while Foresighted uses a weight $w = 14$. These settings are chosen to yield similar attack times (i.e., 8% of the time) across different attack policies. The total power drawn from the operator's PDU is shown as "Metered Power", while the actual server power consumption also includes the contribution from the attacker's batteries ("Attack Load") and hence is larger than the metered power during the attacks. On the other hand, the actual server power is smaller than the metered power during battery charging. The discrepancy between the metered power and actual server power highlights the attacker's "behind-the-meter" cooling loads that are not monitored by the operator.

We see in Fig. 9 that thermal attacks using Random, which remains oblivious of the high cooling load, fail to create any thermal emergencies. Note that, Random's attacks look sparser in Fig. 9 since they are more spread over time while Myopic and Foresighted's attacks are concentrated in the high power/cooling load periods. Myopic exploits the voltage side channel [5] to detect benign tenants' high power loads and launches thermal attacks between hours 0 and 1. Since the power/cooling load remains at a high level, attacks continue until the operator announces a thermal emergency. At that point, attacks are stopped and the power consumption is capped to oblige to the operator's emergency handling protocol. The power returns to a normal level after being capped for 5 minutes to handle the thermal emergency.

While it also launches thermal attacks between hours 0 and 1, Foresighted does not launch a series of unsuccessful short-duration attacks like Myopic. Instead, it waits to regain the battery energy and launches a sustained thermal attack to trigger a second thermal emergency near hour 2. This shows the benefits of reinforcement learning which considers the impact of its actions on the future for maximizing the long-term benefits. Note that, even if Myopic only launches long-duration attacks

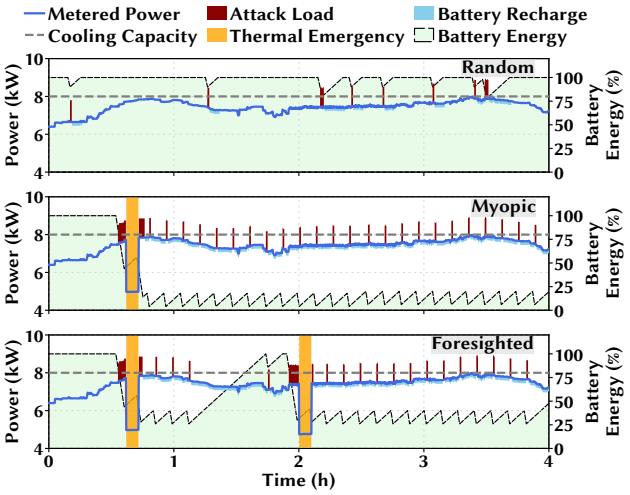
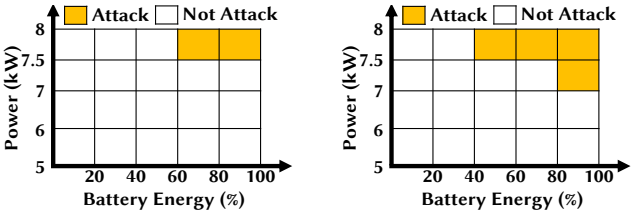


Fig. 9. 4-hour snapshot of thermal attacks.



(a) Weight $w = 9$ (b) Weight $w = 14$

Fig. 10. Attack policy learnt by Foresighted.

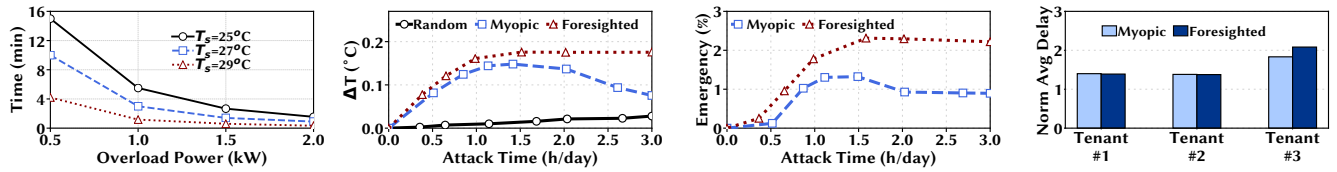
with fully charged batteries, unlike Foresighted, these attacks will more likely occur at the wrong times due to the lack of learning and accounting for battery level dynamics.

B. Attack Policy Learnt by Foresighted

We show in Fig. 10 the structural property of our repeated attack policy learnt by Foresighted: *attack when both the benign tenants' server load and the battery energy level are sufficiently high.* For illustration, we consider two different values of w (the larger w, the more weight on creating temperature increases and hence more attacks). For $w = 9$ in Fig. 10(a), attacks are launched only when the estimated power load (including the attacker's subscribed power capacity) is above 7.5kW and more than 60% of battery energy is left. For $w = 14$, we see that attacks are launched even for 40% remaining battery energy when the power is above 7.5kW. Meanwhile, Foresighted launches attacks at a lower power of 7kW when it has more than 80% battery energy.

C. Cost Estimate

Benign tenants' cost. With an one-shot attack, benign tenants can suffer from service outages, which may be costly or even indirectly cause fatal damages (e.g., decreased safety for assisted driving [46]); with repeated attacks, tenants can potentially experience more frequent performance degradation. The monetary impact of thermal attacks is generally difficult



(a) (b) (c) (d)

Fig. 11. (a) Overload time required to exceed the temperature limit of 32°C. (b) Average temperature increase vs. attack time. (c) Total attack-induced emergency vs. attack time. (d) Tenants' performance during emergencies.

to estimate. To offer an approximate point of reference, we provide a *ballpark* estimate for repeated attacks following prior studies [10], [12], [64] that calculate the cost impact resulting from the increased 95-percentile latency. Under our setting, Foresighted causes a total performance cost of roughly \$60+K/year to benign tenants in our 8kW edge colocation (roughly 80% of benign tenants' total rental costs plus amortized server costs), noting that the actual cost highly depends on the affected tenants' applications and can include additional indirect cost such as business reputation.

Attacker's cost. The attacker's cost involves the power capacity subscription cost, electricity cost, and server purchase cost: 150\$/kW/month power subscription cost, 0.1\$/kWh energy cost, and \$4500 for each server [12]. It is on a par with the cost for other related attacks [5]–[9], and can be affordable for institutional or state-sponsored attackers.

D. Impact of Thermal Attacks

For repeated attacks, we first show in Fig. 11(a) how long it takes for the server inlet temperature to exceed the 32°C threshold. Naturally, the temperature exceeds the threshold sooner with increased cooling overload. Similarly, when the data center is already running hotter (i.e., higher supply temperature T_s), its temperature reaches the limit faster. We see that it takes less than four minutes to increase the data center temperature from 27°C to 32°C with one kW of additional cooling load, demonstrating the potential danger of thermal attacks.

We then vary the total attack energy injected into the edge colocation (i.e., total attack time), while keeping the attack load from the battery fixed at 1kW. We vary the attack probability for Random from 0% to 15%, the load threshold (including the attacker's own power subscription) for launching an attack under Myopic from 6.5kW to 8.0kW, and the weight parameter for Foresighted from $w = 0$ to $w = 30$. Figs. 11(b) and 11(c) show the average server inlet temperature increase (ΔT) beyond 27°C and the amount of attack-induced emergencies (measured in % of the total time) given different average daily attack times, respectively. In Fig. 11(c), we exclude Random because it fails to create any thermal emergency.

Temperature increase. We see in Fig. 11(b) that with more attacks, the temperature increase caused by Random also rises. For Myopic and Foresighted, the temperature increase rises very fast initially, when attacks are conservatively launched. However, as more attacks are launched, the temperature increase for Myopic peaks at around attack time of 1.1 hours per day and then starts to decrease. This is because Myopic launches premature attacks which deplete the battery energy and hence miss future attack opportunities. We see a similar impact on the annual thermal emergency time in Fig. 11(c) where Myopic's performance starts to deteriorate around attack time of 1.5 hours per day.

Foresighted takes the future into account and hence retains both the average temperature increase and annual emergency time increase with more thermal attacks. However, beyond an attack time of 1.5 hours per day, Foresighted cannot create further higher temperature increases nor more thermal emergencies. This is mainly because the total available attack opportunities are limited (i.e., benign tenants do not always have high power loads) and recharging batteries takes time. Nonetheless, given any amount of thermal attacks, Foresighted can create higher server inlet temperature increases and more thermal emergencies than Myopic.

Attack-induced thermal emergencies. In Fig. 11(c), we see that the attack-induced thermal emergencies for both Myopic and Foresighted are close to zero at low attack time. This is because the operator declares a thermal emergency when the data center temperature exceeds 32°C and stays there for at least two minutes. Hence, at low attack time which also corresponds to low average temperature increases in Fig. 11(b), there are almost no thermal emergencies due to attacks.

Performance impacts. We normalize the tenants' 95-percentile response time to that of without any emergencies. We take the average of the normalized response time during the emergency periods and show the result in Fig. 11(d). We see that Myopic has a slightly higher average performance impact than Foresighted. This is because Myopic mainly captures the most prominent attack opportunities while Foresighted intelligently picks up even the subtle opportunities with relatively lower impact, resulting in a lower *average* performance impact. Nonetheless, since Foresighted seizes both the prominent and subtle attack opportunities, it results in more frequent thermal emergencies, thus resulting in a greater cost impact.

E. Sensitivity Study

We now study how the battery capacity, side channel accuracy, attack load, and data center average utilization affect the resulting thermal attacks. We also study the impact

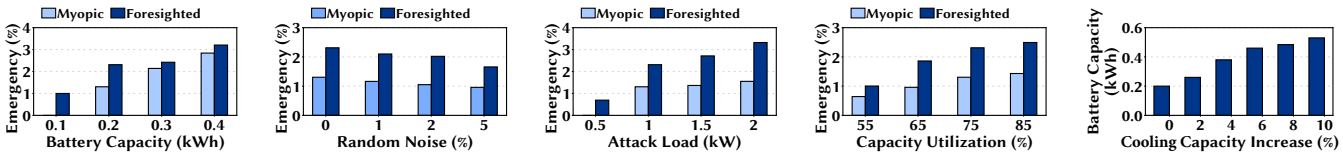


Fig. 12. Sensitivity of Foresighted. (a) Battery capacity. (b) Load estimation due to random noise in side channel. (c) Attack load. (d) Average utilization of data center capacity. (e) Required battery capacity for extra cooling capacity.

of additional cooling capacity on attacker's battery capacity requirement. We exclude Random from our study here since it fails to create any thermal emergency.

Battery capacity. Considering repeated attacks, we vary the battery capacity from 0.1 kWh to 0.4 kWh, and show the annual duration of thermal emergencies due to the attacks in Fig. 12(a). Naturally, a larger battery provides greater flexibility in launching thermal attacks. Hence, we see the annual thermal emergency time increases with battery capacity. We also see the difference between Myopic and Foresighted decreases with a larger battery as the battery is more likely to be available whenever Myopic needs it, like in Foresighted.

Load estimation accuracy. To test robustness against voltage side channel errors, we add varying degrees of random errors to the estimated loads of benign tenants and show our results in Fig. 12(b). As expected, the thermal emergency time decreases for both Myopic and Foresighted when there is more noise in the side channel. Nonetheless, Foresighted can still create a significant amount of thermal emergency, even using a noisy voltage side channel.

Attack load. The attack load determines how much additional cooling load is injected during each attack. We show the results in Fig. 12(c) where we keep the attacker's subscribed capacity at 0.8kW and scale the thermal attack load from 0.5kW to 2kW. We see that the annual emergency time greatly increase with a higher attack load and that Foresighted consistently outperforms Myopic by a great margin.

Capacity utilization. We study the impact of average data center utilization on the thermal attack by scaling the power trace of all the servers while maintaining the peak power at 8kW. Fig. 12(d) shows that the total thermal emergency time increases with increased capacity utilization. This is intuitive since an increased utilization means the data center more frequently operates close to its capacity, thus leading to more thermal attack opportunities.

Extra cooling capacity. We study the impact of the operator's extra cooling capacity on Foresighted's battery requirement to maintain similar impact (i.e., 2.3% emergency). In Fig. 12(e), we see that the extra cooling capacity mandates higher battery capacity. Specifically, the increase in battery capacity for 10% extra cooling capacity is about ∼0.3kWh, which can still be feasible given today's battery energy density. Note, however, that upgrading an existing data center cooling system to add extra cooling capacity is non-trivial due to constraints such as space limitation, data center uptime, etc.

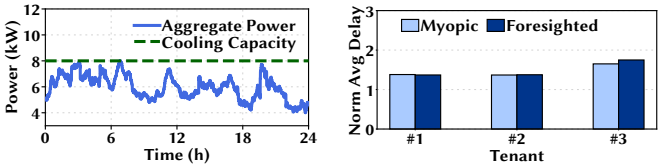


Fig. 13. Results with an alternate power trace. (a) A 24-hour snapshot of the alternate power trace. (b) Tenants' performance during emergencies.

Thus, as discussed in Section VII, other defenses are more effective and cost-efficient, especially for an existing data center that has limited cooling capacity.

F. Results with an Alternate Power Trace

We conduct our year long evaluation with an alternate power trace to demonstrate that Foresighted is effective regardless of the benign tenants' load patterns. We use the Google cluster trace from [40] as the *alternate* total power trace. We show a 24-hour snapshot of the alternate power trace in Fig. 13(a). Like in the default setting, we scale the power trace to have a 75% average utilization in our 8kW edge colocation. We keep the same default settings as in Section V for Myopic and Foresighted. Fig. 13(b) shows that, with the alternate power trace, benign tenants suffer from similar performance degradation as in our earlier results. While we omit detailed discussion for space limitation, these findings are consistent with our earlier results.

VII. Defense Mechanism

Tenants generally expect reliable power and cooling supplies (subject to contractual terms) from the colocation operator which manages non-IT systems. Thus, we offer possible defenses from the operator's perspective. We first discuss defenses that aim at preventing potential thermal attacks, followed by defenses that detect thermal attacks.

A. Prevention

The following defense strategies are proactive measures to inhibit potential thermal attacks.

Infrastructure resilience. A straightforward defense against thermal attacks is to reinforce an edge colocation's physical infrastructure for handling thermal overloads. For this, the operator can deploy a cooling system with additional redundancies. This approach, however, can increase the capital

cost [24], [65] and be particularly challenging for existing systems. Alternatively, the operator can lower its server inlet temperature set point (to 20°C instead of the recommended 27°C) to have more margins for triggering thermal emergencies. The drawback is the increased cooling energy cost [15], [31]. Thus, while oversubscribing data center cooling capacity [15], [16], [24] and increasing temperature set point [31] have been suggested for cost efficiency, they should be carefully exercised, balancing the benefit versus risk to potential thermal attacks.

Rigorous move-in inspection. The colocation operator can employ a more rigorous background check and move-in inspection process for all tenants' servers to detect and remove integrated batteries. Note that without built-in batteries, the attacker cannot have additional power sources to support thermal attacks behind the meter or overload the shared cooling capacity, unless the data center cooling capacity is oversubscribed as suggested by recent studies [15], [16], [24]. Besides, the operator can also enforce on-site power load tests to ensure that the server power is consistent with the tenant's data center capacity subscription. The operator should be particularly careful about the servers' peak power.

Degrading physical side channels. The colocation operator may increase the attacker's uncertainties about timing attacks by degrading/eliminating the physical side channel. For example, it can add jamming noise signals into the colocation power networks and/or use power line noise filters. Additionally, the operator may also prohibit unusual sensors (e.g., microphones) on tenants' servers in order to prevent an attacker from exploiting other possible but unknown side channels.

B. Detection

Detection strategies can be implemented to catch an attacker that may circumvent prevention approaches.

Detecting behind-the-meter cooling loads. The same power reading can result in different cooling loads and server inlet/outlet temperature, depending on whether malicious thermal attacks are launched or not. Thus, by using anomaly detection algorithms (e.g., cross-checking readings by temperature sensors and power meters), the operator can detect an irregular thermal environment possibly due to thermal attacks.

Identifying attacks from impacts. One-shot attacks can be easily identified through a thorough inspection if a system outage occurs. By contrast, repeated-attacks that inject milder loads to trigger more frequent thermal emergencies can require more efforts. Since precise temperature management is difficult with open airflow cooling, there can be occasional thermal emergencies in colocations even without thermal attacks; colocation operators often offer a long-term temperature SLA (e.g., the inlet temperature is conditioned below 27°C for 99% or more of the time) [66], [67]. This may potentially allow an attacker to hide behind the statistics for a longer time. Thus, advanced algorithms can be implemented to monitor SLA metrics to early detect the presence of thermal attacks.

Improved data center monitoring. While the aforementioned approaches can detect thermal attacks, pin-pointing the attacker's servers — the source of the injected cooling load — is still needed to hold the attacker accountable. Thus, to monitor the servers' actual cooling loads, the operator can measure each server's outlet temperature as well as the hot air flows. Alternatively, thermal cameras may be employed to identify the servers that are running extra hot. Likewise, microphone arrays can be used along with the thermal camera to pinpoint servers with fans spinning at a high speed (needed by servers that have higher cooling loads) [7]. While these monitoring apparatuses are not used in all data centers, they are readily available and can be easily installed by data centers to identify malicious cooling loads.

To sum up, there exist readily-available defenses, such as move-in inspection to disallow built-in batteries, advanced anomaly detection, and installation of monitoring apparatuses to locate the attacker. Given the potential threat of thermal attacks that are currently neglected, the edge colocation operator can implement one or more of the suggested defenses to safeguard its thermal environment for tenants.

VIII. Related Works

Power and thermal management. The common practice of aggressive capacity oversubscription can create occasional capacity overloads when the demand peaks [12]–[14], [18], [19], [21]. To safely ride through power emergencies, numerous graceful power capping techniques have been proposed, such as throttling CPU frequencies [13], migrating/deferring workloads [14], [45], and discharging batteries to boost power supply [18], [19], [21]. Likewise, managing server loads to handle thermal emergencies are equally crucial [16], [20], [43]. These studies, however, are not applicable for colocations whose operators have no control over tenants' servers. Moreover, they do not consider an adversarial setting. More recent works [12], [68] propose market approaches to coordinate tenants' power demand in colocations, but they assume that tenants are all benign without any malicious intentions.

Data center security and thermal fault attacks. Securing data centers against cyber attacks, such as network DDoS [69] and data/privacy breach [70], has been extensively investigated. Prior studies have also considered malicious thermal load attacks on a single device [71]. More recently, data center power and cooling system security has been emerging as a crucial concern [5]–[9], [51], [72]. However, these works focus on overloading the power infrastructure (i.e., power attack) of large data centers with multi-level redundancy or creating hotspots (i.e., thermal attack) in Amazon-type cloud with frequent VM shuffling. In contrast, we focus on novel battery-assisted thermal attacks in a shared edge colocation. Moreover, our repeated battery-assisted thermal attacks are *stateful* whereas prior attacks are stateless as the current attack does not depend on any past/future attacks.

Battery management and others. The prior studies have exploited batteries for various purposes, such as better energy capacity [63], concealing a household's electricity usage information from the utility for better privacy [73], smoothing data center power demand [18], [19], [21], among many others.

To our knowledge, however, our study is the first to leverage batteries for a malicious purpose — one-shot or repeated thermal attacks in edge colocations — which highlights the need of attention to the potential threat.

IX. Conclusion

In this paper, we discovered that the sharing of cooling systems may expose edge colocations' potential vulnerabilities to both one-shot and repeated thermal attacks assisted with built-in batteries. For repeated attacks, we presented a foresighted attack policy which, using reinforcement learning, learns on the fly a good timing for thermal attacks. We also ran simulations to validate our attacks and showed that, for an 8kW edge colocation, an attacker can cause performance degradation for affected tenants. Finally, we suggested effective countermeasures against potential thermal attacks that are currently neglected in many data centers.

References

[1] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet of Things Journal*, vol. 3, pp. 637–646, Oct 2016.

[2] DatacenterKnowledge, "NTT plans global data center network for connected cars," http://www.datacenterknowledge.com/archives/2017/03/27/ntt-plans-global-data-center-network-for-connected-cars.

[3] Vapor IO, "The edge data center," https://www.vapor.io/.

[4] Uptime Institute, "Data center industry survey," 2018, https://uptimeinstitute.com/2018-data-center-industry-survey-results.

[5] M. A. Islam and S. Ren, "Ohm's law in data centers: A voltage side channel for timing power attacks," in *CCS*, 2018.

[6] M. A. Islam, S. Ren, and A. Wierman, "Exploiting a thermal side channel for power attacks in multi-tenant data centers," in *CCS*, 2017.

[7] M. A. Islam, L. Yang, K. Ranganath, and S. Ren, "Why some like it loud: Timing power attacks in multi-tenant data centers using an acoustic side channel," in *SIGMETRICS*, 2018.

[8] Z. Xu, H. Wang, Z. Xu, and X. Wang, "Power attack: An increasing threat to data centers," in *NDSS*, 2014.

[9] X. Gao, Z. Xu, H. Wang, L. Li, and X. Wang, "Reduced cooling redundancy: A new security vulnerability in a hot data center," in *NDSS*, 2018.

[10] Ponemon Institute, "2016 cost of data center outages," 2016, https://www.ponemon.org/blog/2016-cost-of-data-center-outages.

[11] P. Jones, "Overheating brings down microsoft data center," *Datacenter Dynamics*, 2013, https://www.datacenterdynamics.com/news/overheating-brings-down-microsoft-data-center/.

[12] 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.

[13] 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.

[14] G. Wang, S. Wang, B. Luo, W. Shi, Y. Zhu, W. Yang, D. Hu, L. Huang, X. Jin, and W. Xu, "Increasing large-scale data center capacity by statistical power control," in *EuroSys*, 2016.

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

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

[17] Supermicro, "Battery backup power - evolutionary design to replace UPS," http://www.supermicro.com/products/nfo/files/bbp/f_bbp.pdf.

[18] B. Aksanli, T. Rosing, and E. Pettis, "Distributed battery control for peak power shaving in datacenters," in *IGCC*, 2013.

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

[20] Y. Kim, J. Choi, S. Gurumurthi, and A. Sivasubramaniam, "Managing thermal emergencies in disk-based storage systems," Dec 2008.

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

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

[23] DatacenterKnowledge, "Vapor IO to sell data center colocation services at cell towers," http://www.datacenterknowledge.com/archives/2017/06/21/vapor-io-to-sell-data-center-colocation-services-at-cell-towers.

[24] M. Skach, M. Arora, D. Tullsen, L. Tang, and J. Mars, "Virtual melting temperature: Managing server load to minimize cooling overhead with phase change materials," in *ISCA*, 2018.

[25] S. Malla, Q. Deng, Z. Ebrahimzadeh, J. Gasperetti, S. Jain, P. Kondety, T. Ortiz, and D. Vieira, "Coordinated priority-aware charging of distributed batteries in oversubscribed data centers,"

[26] V. Sakalkar, V. Kontorinis, D. Landhuis, S. Li, D. De Ronde, T. Blooming, A. Ramesh, J. Kennedy, C. Malone, J. Clidaras, and P. Ranganathan, "Data center power oversubscription with a medium voltage power plane and priority-aware capping," in *ASPLOS*, 2020.

[27] A. Kumbhare, R. Azimi, I. Manousakis, A. Bonde, F. Frujeri, N. Mahalingam, P. Misra, S. A. Javadi, B. Schroeder, M. Fontoura, and R. Bianchini, "Prediction-based power oversubscription in cloud platforms," 2020.

[28] T. Evans, "The different technologies for cooling data centers," http://www.apcmedia.com/salestools/VAVR-5UDTU5/VAVR-5UDTU5_R2_EN.pdf.

[29] Google, "Heat containment," http://www.google.com/about/datacenters/efficiency/external/.

[30] D. L. Moss, "Dynamic control optimizes facility airflow delivery," *Dell White Paper*, March 2012.

[31] Q. Tang, S. K. S. Gupta, and G. Varsamopoulos, "Thermal-aware task scheduling for data centers through minimizing heat recirculation," in *CLUSTER*, 2007.

[32] S. V. Patankar, "Airflow and cooling in a data center," *Journal of Heat Transfer*, vol. 132, p. 073001, July 2010.

[33] R. A. Steinbrecher and R. Schmidt, "Data center environments: Ashrae's evolving thermal guidelines," *ASHRAE Technical Feature*, pp. 42–49, December 2011.

[34] Dell, "Integrated dell remote access controller 9 (iDRAC9) version 3.00.00.00."

[35] 365DataCenters, "Master services agreement," http://www.365datacenters.com/master-services-agreement/.

[36] R. A. Bridges, N. Imam, and T. M. Mintz, "Understanding gpu power: A survey of profiling, modeling, and simulation methods," *ACM Comput. Surv.*, vol. 49, pp. 41:1–41:27, Sept. 2016.

[37] Keysight Technology, "Learn to connect power supplies in parallel for higher current output," https://www.keysight.com/main/editorial.jspx?cc=US&lc=eng&ckey=520808&nid=-11143.0.00&id=520808.

[38] S. Govindan, D. Wang, A. Sivasubramaniam, and B. Urgaonkar, "Aggressive datacenter power provisioning with batteries," *ACM Trans. Comput. Syst.*, vol. 31, pp. 2:1–2:31, Feb. 2013.

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

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

[41] P. Lin, S. Zhang, and J. VanGilder, "Data center temperature rise during a cooling system outage," *APC White Paper 179*, 2014.

[42] Intel, "Intel cloud builders guide to power management in cloud design and deployment using Supermicro platforms and NMView management software," 2013.

[43] L. Ramos and R. Bianchini, "C-Oracle: Predictive thermal management for data centers," in *HPCA*, 2008.

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

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

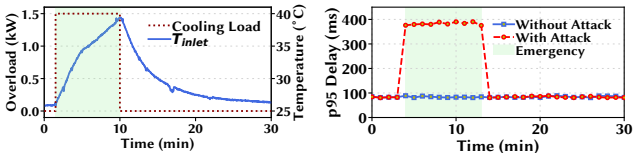
[46] S. Baidya, Y.-J. Ku, H. Zhao, J. Zhao, and S. Dey, "Vehicular and edge computing for emerging connected and autonomous vehicle applications," in *DAC*, 2020.

[47] "Calb 100 ah se series lithium iron phosphate battery," https://www.evwest.com/catalog/product_info.php?products_id=51.

[48] "Poweredge r740xd rack server," https://www.dell.com/en-us/work/shop/povw/poweredge-r740xd.
[49] L. Brochard, V. Kamath, J. Corbalán, S. Holland, W. Mittelbach, and M. Ott, *Energy-Efficient Computing and Data Centers*. John Wiley & Sons, 2019.
[50] K. Lee, N. Klingensmith, S. Banerjee, and Y. Kim, "Voltkey: Continuous secret key generation based on power line noise for zero-involvement pairing and authentication," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 3, Sept. 2019.
[51] X. Gao, Z. Gu, M. Kayaalp, D. Pendarakis, and H. Wang, "ContainerLeaks: Emerging security threats of information leakages in container clouds," in *DSN*, 2017.
[52] J. N. Tsitsiklis, "Asynchronous stochastic approximation and q-learning," *Machine learning*, vol. 16, no. 3, pp. 185–202, 1994.
[53] J. Xu, L. Chen, and S. Ren, "Online learning for offloading and autoscaling in energy harvesting mobile edge computing," *IEEE Transactions on Cognitive Communications and Networking*, vol. 3, no. 3, pp. 361–373, 2017.
[54] Google, "Google's Data Center Efficiency," http://www.google.com/about/datacenters/.
[55] Vertiv, "Smartmod modula data center infrastructure."
[56] X. Wang, X. Wang, G. Xing, and C. xian Lin, "Leveraging thermal dynamics in sensor placement for overheating server component detection," in *IGCC*, 2012.
[57] NVidia, "https://www.nvidia.com/en-us/geforce/graphics-cards/30-series/rtx-3080/."
[58] D. G. Feitelson, D. Tsafrir, and D. Krakov, "Experience with using the parallel workloads archive," *Journal of Parallel and Distributed Computing*, vol. 74, no. 10, pp. 2967–2982, 2014.
[59] Parallel Workloads Archive, http://www.cs.huji.ac.il/labs/parallel/workload/.
[60] X. Fan, W.-D. Weber, and L. A. Barroso, "Power provisioning for a warehouse-sized computer," in *ISCA*, 2007.
[61] M. E. Haque, Y. h. Eom, Y. He, S. Elnikety, R. Bianchini, and K. S. McKinley, "Few-to-many: Incremental parallelism for reducing tail latency in interactive services," in *ASPLOS*, 2015.
[62] E. Even-Dar and Y. Mansour, "Learning rates for q-learning," *Journal of Machine Learning Research*, vol. 5, no. Dec, pp. 1–25, 2003.
[63] L. He, E. Kim, and K. G. Shin, "*aware charging of lithium-ion battery cells," in *ICCPS*, 2016.
[64] P. X. Gao, A. R. Curtis, B. Wong, and S. Keshav, "It's not easy being green," *SIGCOMM Comput. Commun. Rev.*, 2012.
[65] I. Goiri, R. Bianchini, S. Nagarakatte, and T. D. Nguyen, "Approxhadoop: Bringing approximations to mapreduce frameworks," in *ASPLOS*, 2015.
[66] Internap, "Colocation services and SLA," http://www.internap.com/internap/wp-content/uploads/2014/06/Attachment-3-Colocation-Services-SLA.pdf.
[67] Equinix, "Colocation services and SLA," https://enterprise.verizon.com/service_guide/reg/cp_colocation_equinix_data_centers_sla.pdf.
[68] M. A. Islam, H. Mahmud, S. Ren, and X. Wang, "Paying to save: Reducing cost of colocation data center via rewards," in *HPCA*, 2015.
[69] S. Yu, Y. Tian, S. Guo, and D. O. Wu, "Can we beat ddos attacks in clouds?," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, pp. 2245–2254, September 2014.
[70] Y. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Cross-vm side channels and their use to extract private keys," in *CCS*, 2012.
[71] S. Skorobogatov, "Local heating attacks on flash memory devices," in *Workshop on Hardware-Oriented Security and Trust*, 2009.
[72] C. Li, Z. Wang, X. Hou, H. Chen, X. Liang, and M. Guo, "Power attack defense: Securing battery-backed data centers," in *ISCA*, 2016.
[73] L. Yang, X. Chen, J. Zhang, and H. V. Poor, "Optimal privacy-preserving energy management for smart meters," in *INFOCOM*, 2014.
[74] "CloudSuite - The Search Benchmark," http://cloudsuite.ch/.

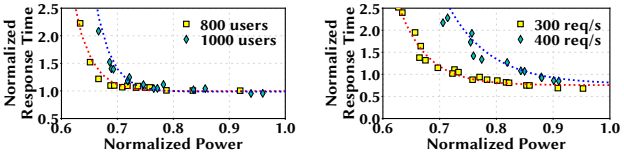
APPENDIX A
PROTOTYPE DEMONSTRATION OF THERMAL ATTACKS

To see the impact of thermal attacks, we run experiments on a rack of 14 Dell PowerEdge servers in a scaled environment with hot-cold aisles to mimic an edge colocation. The cooling system can support up to a cooling load of 3kW. We inject an additional 1.5kW load to overload the cooling system and measure the server inlet temperature. As shown in Fig. 14(a), the inlet temperature rises to nearly 40° C within minutes. Our experiment, albeit on a small scale, demonstrates the rapid increase of server inlet temperature due to a overloaded cooling system. This is also corroborated by other studies that demonstrate rapid temperature rises in data centers due to cooling malfunction [41]. We follow the ASHRAE safety limit and do not further overload our system [33].



(a) Temperature (b) Performance

Fig. 14. Experiment in our server rack. (a) Server inlet temperature increases due to a cooling capacity overload by 1.5kW. (b) Latency performance is compromised due to server power capping for handling an emergency.



(a) Web Service (b) Web Search

Fig. 15. Performance degradation due to power capping.

We implement the ClouSuite Web Service benchmark [74] in a set of 4 servers with a workload of 600 requests/s and show the impact of power capping on the 95-percentile response time, which is the key performance metric [61]. An x-percentile response time means that $x\%$ of the requests have a latency less than this response time. For illustration, we throttle the CPU speed to cap the total server power to 60% of the peak power. We see from Fig. 14(b) that during the emergency, the response time jumps nearly four times to 400ms.

We also extend our experiments to Web Search implementation from CloudSuite [74]. We show the 95-th percentile response time normalized to the service level agreement (100ms) for two different numbers of users for Web Service in Fig. 15(a) and two different request rates for Web Search in Fig. 15(b), respectively. The server power consumption is normalized to the peak. We see that when the server power consumption decreases, the response times for both applications increase for any given workload level. This reveals the degree of performance degradation faced by tenants when they reduce their power consumption while the workload remains unchanged.