Reduced Cooling Redundancy: A New Security Vulnerability in a Hot Data Center

Xing Gao*† Zhang Xu† Haining Wang* Li Li‡ Xiaorui Wang‡

*University of Delaware {xgao,hnw}@udel.edu
†College of William and Mary zxu@cs.wm.edu
‡Ohio State University {li.2251,wang.3596}@osu.edu

***Abstract*—Data centers have been growing rapidly in recent years to meet the surging demand of cloud services. However, the expanding scale and powerful servers generate a great amount of heat, resulting in significant cooling costs. A trend in modern data centers is to raise the temperature and maintain all servers in a relatively hot environment. While this can save on cooling costs given benign workloads running in servers, the hot environment increases the risk of cooling failure. In this paper, we unveil a new vulnerability of existing data centers with aggressive cooling energy saving policies. Such a vulnerability might be exploited to launch thermal attacks that could severely worsen the thermal conditions in a data center. Specifically, we conduct thermal measurements and uncover effective thermal attack vectors at the server, rack, and data center levels. We also present damage assessments of thermal attacks. Our results demonstrate that thermal attacks can (1) largely increase the temperature of victim servers degrading their performance and reliability, (2) negatively impact on thermal conditions of neighboring servers causing local hotspots, (3) raise the cooling cost, and (4) even lead to cooling failures. Finally, we propose effective defenses to prevent thermal attacks from becoming a serious security threat to data centers.**

I. Introduction

As cloud computing has become the mainstream of providing IT services, data centers have expanded their scales and are equipped with more powerful servers to meet the ever-increasing service demands. Correspondingly, the amount of heat emitted by those servers is also surging, which requires the cooling system to more efficiently dissipate the increased heat. Otherwise, the overheating would potentially lead to serious hardware failures and even server shutdown for self-protection. Unfortunately, in recent years, the online service interruptions due to cooling failures have not been rare in cloud vendors and enterprises, including Microsoft [10], Rackspace [11], Wikipedia [13], iiNet [9], and University of Pennsylvania [12].

To regulate the temperature in computer rooms, a significant portion of the power consumption of data centers is used for cooling. The cooling cost has reached 24% of a data center's budget [7]. The key factor affecting the cooling cost of CRAC (Computer Room Air Conditioning) systems is the supply air temperature. Data centers have deployed a variety of methods to control the CRAC supply air temperature. For example, some data centers set the supply air temperature automatically based on the workload. More importantly, a recent study shows that increasing the supply air temperature by merely 1°C can save approximately 2-5 percent of the cooling power [19]. Thus, there is a trend in data centers to raise the highest set temperature from 75°F to 85°F or even higher. It is reported that Google has raised the temperature of the cold aisle to 80°F [5]. Those aggressive cooling energy saving policies achieve a low PUE (Power Usage Effectiveness)[1] in a data center. However, the temperature increment also forces the servers running in a hotter environment than before.

Furthermore, advanced techniques like power oversubscription [20], [42] have been widely adopted to accommodate more servers in data centers without upgrading existing power and cooling infrastructures. While the infrastructures were initially well designed with sufficient cooling redundancies, those redundancies have been excessively consumed by power oversubscription. Under the reduced cooling redundancies, an accidental synchronization of running intensive workloads in a set of adjacent servers could result in a local hotspot and even a cascade effect further deteriorating the thermal conditions.

In this paper, we systematically investigate the security risk posed by those aggressive policies applied on data centers. We introduce the concept of thermal attack, which can be easily and remotely launched to seriously worsen the thermal conditions at a server level, a rack level, or even a data center level, without requiring any privileges of a hypervisor. Thermal attacks simply run thermal-intensive workloads in victim servers or VMs (Virtual Machines) to rapidly generate a large amount of heat, forcing the victim servers into a high temperature. The overheated servers suffer performance and reliability degradation. Even worse, the accumulated heat can further exacerbate the thermal condition of the peripheral atmosphere, raising the inlet temperature of other servers. The increase of the inlet temperature then increases other servers' outlet temperatures, leading to a vicious cycle. The consequence could be the great increase of hardware failures of many servers in a data center, the significant waste of the cooling costs, and even thermal accidents that force some servers to shut down.

Network and Distributed Systems Security (NDSS) Symposium 2018
18-21 February 2018, San Diego, CA, USA
ISBN 1-891562-49-5
http://dx.doi.org/10.14722/ndss.2018.23165
www.ndss-symposium.org

[1] PUE is the ratio of the total energy consumption of a data center to the energy consumed by computing equipments in the data center.

To form the basis for mounting a thermal attack, we measure how thermal-related factors are exhibited in a real server using different HPC (High Performance Computing) benchmarks. We observe that CPU-intensive workloads can generate more heat and cause a higher temperature than other types of workloads, even if the system utilization is at the same level; thus it can be used as thermal-intensive workload to rapidly raise the temperature of a server (i.e., within a few minutes), despite the fact that the temperature increase requires the accumulation of heat. Then, based on our thermal measurements on the real server, we propose to mount thermal attacks in both non-virtualized and virtualized environments, as well as a pulsation thermal attack. As expected, those attacks can largely raise the temperature of the hosting server within a short period time. We further conduct a damage analysis in terms of electromigration, time dependent dielectric breakdown, thermal cycling intrinsic hard failure mechanisms, and disk failure.

To evaluate the impact of thermal attacks on the entire data center, we conduct thermal attacks at the data center level using a computational fluid dynamics based, trace-driven simulation, with a special consideration of the air recirculation condition in the data center. We observe that launching thermal attacks on less than 2 percent of servers in a data center can seriously affect the thermal conditions of the whole data center and raise the cooling costs significantly. Even worse, in some severe attack scenarios, thermal attacks can lead to cooling failures.

Finally, we discuss the attack costs and propose effective defenses against thermal attacks. Although some prior studies have proposed temperature-aware load balancing (e.g., [40]), their approaches are mainly designed based on a *static* profile of the data center thermodynamics, which is profiled using *normal* server workloads. As a result, such solutions cannot effectively handle hotspots that are generated rapidly by malicious workloads at runtime, because the thermal conditions can become significantly different and even completely opposite to the static pre-calculated profile.

The remainder of the paper is organized as follows. We provide the background information on system cooling and its inherent vulnerability to a thermal attack in data centers in Section 2. We present our thermal measurement study on a real server in Section 3. We detail the thermal attacks in non-virtualized/virtualized environments and the damage analysis in Section 4. We evaluate the attacks and their impacts at both the rack and data center levels in Section 5. We propose an effective defense in Section 6. We survey the related work in Section 7, and finally we conclude the paper in Section 8.

II. Background

A. Server Cooling

Computers have to be operated within a specific range of environmental temperatures. Otherwise, electronic devices may fail to have their normal characteristics and may even malfunction. Most hardware failures cause permanent damage and cannot be recovered. For a server, all its components, like CPU, memory, cache, and bus, produce heat. The mega-scale and complicated design of integrated circuits and chips in modern computer systems further exacerbates the heat generation problem. As a result, cooling techniques or thermal management

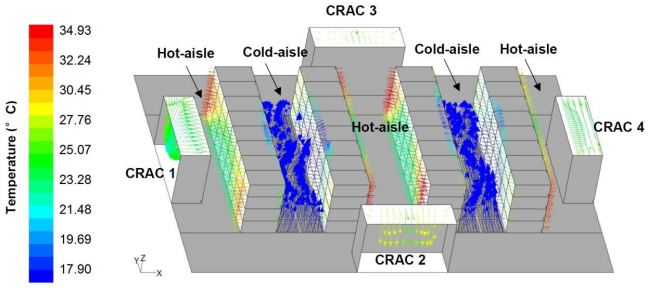


Fig. 1: Layout of a typical data center.

methods are critical for both hardware and software design. For instance, the chip design must consider heat generation and emission. The positions of different chip components must be carefully chosen to avoid hotspots. A fan is also often used to accelerate the heat emission. In software, various dynamic thermal management mechanisms are proposed: If the temperature violates a carefully selected threshold, the component would have to degrade its performance by reducing the operating frequency, or it may even be forced to shut down for protection from physical damages [41]. Therefore, a redline temperature is often set for a server. If the inlet temperature exceeds the redline temperature, the server would be shut down because the fan itself is not sufficient to cool down the server.

B. Data Center Cooling

In a data center, tens to hundreds of thousands of high-density (e.g., blade) servers are placed in one closed space, running simultaneously with a high workload utilization. A large amount of heat is generated every second by those servers. Moreover, the fans push the generated heat into the room. Due to possible air recirculation, the hot air may influence the environmental temperature and further impact other servers. Therefore, data center cooling is even more critical due to the high server density.

To cool down servers, existing data centers are equipped with various cooling technologies, including the traditional CRAC air cooling, free air cooling, and liquid cooling. Restrained by geographic limitations and facility costs, CRAC-based air cooling remains the most widely used cooling solution. To enhance cooling efficiency, computer rooms are divided into hot aisles and cold aisles, as shown in Figure 1. In a cold aisle, cold air is supplied from the raised floor and flows through the back side of the server racks to absorb the heat generated by the servers. The resulting hot air, with the help of server fans, then enters the hot aisle from the front side of the server racks, and is returned (through the ceiling vents) to the CRAC system and cooled down by the chillers again.

The inlet temperatures of servers must be strictly controlled in a data center in order to avoid overheating, which increases the possibility of causing permanent hardware damage. If the inlet temperature exceeds a threshold, parts of servers or even racks would be forced to shut down. Ideally, the data center should have the cold air and hot air perfectly isolated for high cooling efficiency. However, air recirculation can be common in a data center, in which hot air enters the cold aisles through

small gaps between insulation material and the server racks. Thus, cold and hot air can get mixed to some extent as a result. The air density, air flow between servers, supply air temperature of the cooling facilities, power consumption, and outlet temperature of each server can get intertwined in a complicated way to impact the temperature of the whole server room. Therefore, the hot air generated by the servers could potentially cause the inlet temperature to violate the threshold.

To keep the inlet temperature below the safe threshold, the cooling facility's supply air temperature selection is critical. For safety, a low supply air temperature is desired to reduce the possibility of any thermal emergencies. However, a lower supply air temperature can result in a higher cooling cost. As a result, many data centers (e.g., Google) are trying to raise the supply air temperature for a lower cooling cost. Currently, data centers can remain safe even with a higher supply air temperature due to the existence of the cooling redundancies. However, with the deployment of more powerful servers and the current trend of power oversubscription [21], [54], which allows managers to deploy more servers in a room, data centers will soon have almost no redundancies in the near future. As a result, the possibilities of thermal emergencies can significantly increase with a higher supply air temperature, making data centers vulnerable to thermal-related attacks.

Currently in data centers, the common cause of a cooling failure is the breakdown of cooling facilities [13], [11], [9], which significantly reduces cooling capacity and slow down heat dissipation. Conversely, by intentionally running thermal-intensive workloads, the vast heat generation can also exceed the cooling capacity, resulting in a cooling failure in a data center.

C. Threat of Thermal Attacks

The security threat posed by thermal attacks to a data center is real and difficult to address, mainly due to the following five reasons.

I. The root cause of the threat lies in the wide adoption of aggressive cooling and power management policies, such as raising the supply air temperature and power oversubscription, which allow more servers being deployed in a data center with less cooling cost. Although data centers were initially designed with sufficient cooling redundancies, those aggressive policies significantly reduce the cooling redundancies, making data centers themselves vulnerable to abnormal thermal conditions.

II. Although modern servers are equipped with various chip-level sensors, such as temperature sensor for each core, those chip-level sensors cannot provide server-level information (e.g., server inlet and outlet temperature). Core temperature does not equal to inlet temperature and outlet temperature. A core's temperature varies much faster. An effective thermal attack does not need to stress the CPU all the times but just keep the outlet temperature at a high level. As a result, Core sensors cannot provide server-level or DC-level thermal monitoring.

III. While thermal sensors definitely help to monitor the thermal conditions in a computer room, most of today's production data centers have just a few thermal sensors or probes for the entire data center. Deploying sensors on the server- or rack-level would be very costly. Even with rack-level sensors (only 2-4 sensors per rack), it is impossible for them to cover hundreds of servers in a rack. Thus, the occurrence of a local hotspot is inevitable. Without the global knowledge of overall thermal conditions in a computer room, chip-level protections, like the Intel RAPL (Running Average Power Limit) providing and limiting the power consumption for each CPU package, cannot prevent the occurrence of local hotspots.

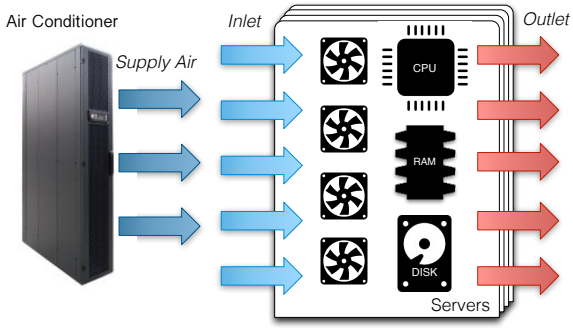


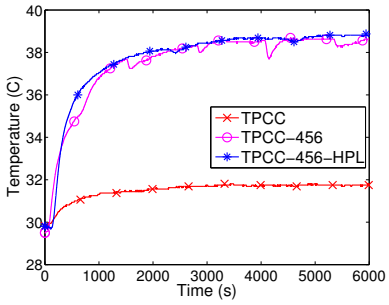
Fig. 2: Inlet and outlet temperature of servers.

IV. To defend against thermal attacks on a data center, the thermodynamics of a data center is important to consider. The temperature-based feedback control mechanism would help to limit the temperature for local hotspots. However, such a feedback control mechanism does not exist in the current data centers. At the chip-level, there are some feedback control mechanisms used for overheating protection; however, at the data-center-level, without the global knowledge on thermal conditions, it is very challenging to deploy feedback control mechanisms for temperature management of the whole data center.

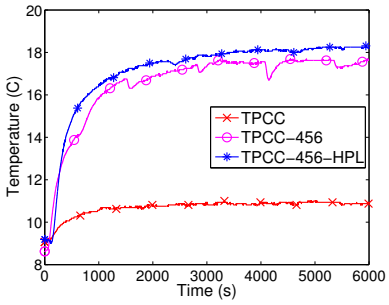
V. Since thermal-intensive workloads themselves are benign and do not exploit any system vulnerabilities, it is difficult to distinguish attackers from normal users. Moreover, process-level power/thermal profiling also cannot defend against data center level thermal attacks. Attackers are not limited to use just one process to generate much more heat. They can use many accounts to run different workloads simultaneously to generate a significant amount of heats.

III. Real Server Measurement

The thermal conditions of a server are affected by various factors. Unlike power that can be generated and terminated instantaneously, the change of temperature is a process of heat accumulation and dissipation. Different components of a server run simultaneously and generate heat. The accumulated heat then causes the temperature to raise. To understand the thermal characteristics of a physical server, we first perform a measurement study running in our small testbed. We carefully design a set of experiments to explore the thermal characteristics of a server in the following four aspects: (1) the impact of different workloads, (2) the thermal condition variations under the same system utilizations, (3) the relationship between the thermal condition and power consumption, and (4) the speed of heat accumulation and dissipation.



(a) Outlet temperature.



(b) Margin temperature.

Fig. 3: Temperature with HPL, SPECCPU and TPCC-UVa.

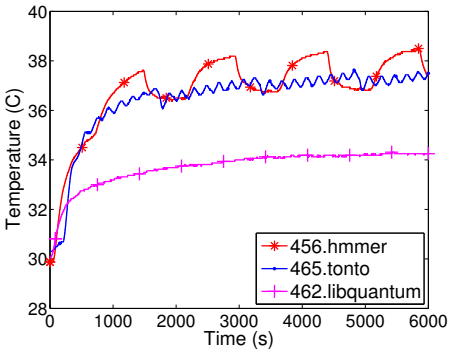


Fig. 4: Outlet temperature with different SPECCPU benchmarks.

Our testbed is a mainstream Supermicro server, equipped with Intel Xeon 2.27GHz CPU with 16 cores, 32 GB of RAM and running Ubuntu Linux 12.04 with kernel version 3.13.0. The testbed is placed in a sealed environment without any window to the outside world. The ambient temperature is about 21°C cooled by the central air conditioner in our building. We measure the inlet and outlet temperature of the server using "Go!Temp" temperature probe, which has a resolution of 0.07°C [8], and monitor the power consumption using "Watts up? .Net" digital power meter. The inlet and outlet temperature of our server is illustrated in Figure 2.

A. Workload Impacts

We first explore how different workloads affect the thermal conditions of a server. The attack workload should be designed to fully utilize all components, generate a large amount of heat and raise the temperature in a fast and significant manner. The "man-made" hot status would reduce the reliability of a server. To evaluate the impact, we pay special attention to the outlet temperature of the server for the following two reasons. First, the output air is a direct sign of the temperature of the server. The inner temperature can only be hotter than the outlet temperature. Second, the output air is exported to the hot aisle of the computer room and further impacts the whole atmosphere due to air recirculation. Thus, a high outlet temperature can negatively affect adjacent servers in a computer room.

We measure the outlet temperature of our server under three different scenarios. First, we use TPCC-UVa database benchmark [36] as the workload to generate a large amount of I/O operations. TPCC-UVa benchmark is an open source TPC-C benchmark, which is an online transaction processing benchmark, written in C language and using the PostgreSQL database engine. We set the warehouse number 50 to ensure a sufficiently large workload, but keep the CPU utilization less than 10% to limit the intensity of CPU activities. In the second scenario, we simultaneously run the TPCC-UVa and the 456.hmmer benchmark from SPECCPU2006, which is a widely used HPC benchmark. We run the 456.hmmer with 16 copies to fully utilize the CPU resources in our server. In the third scenario, we add another HPC benchmark, the High Performance Linpack (HPL). The Linpack benchmarks measure the capability of solving random matrix productions. There are multiple configurable parameters that could affect the workload of the benchmark. We set the number of processors to 16 since our server is a 16-core machine. We set the problem size N, which is the size of the input matrix, to 40,000. For the block size, we choose 100 in this experiment. This configuration promises a heavy computing workload on our server. Since temperature increase is a relatively long process of heat generation, we run the experiment for 100 minutes.

The experimental results are illustrated in Figure 3. Figure 3(a) shows the outlet temperature under the three different scenarios. Although the air conditioning is set with a fixed supply air temperature, the inlet temperature still varies in a range of 1°C due to various surrounding factors. We define the margin temperature as the difference between the outlet and inlet temperature, which is shown in Figure 3(b). The margin temperature clearly indicates the temperature increase because of running specific workloads on our server.

While the outlet temperature of the server at idle is about 30°C, after heating by intensive workloads on CPU, memory, and disk, the outlet temperature can reach up to 39°C, as shown in Figure 3(a). Specifically, I/O intensive workloads create heat and raise the outlet temperature to some extent; more obviously, after running with CPU-intensive workloads like 456.hmmer, the outlet temperature increases significantly. Then, additional workloads like HPL can further generate more heat on this basis. Given that the inlet temperature is about 21°C set by the cold air, the thermal-intensive workloads achieve a more than 18-degree temperature difference. Also, a nearly 40°C outlet temperature indicates an even hotter temperature inside the server. Thus, our experimental results demonstrate the feasibility of mounting a thermal attack against a server using thermal-intensive workloads, especially CPU-intensive workloads.

B. System Utilization

To explore the outlet temperature variation under the same system utilization, we use a set of SPECCPU 2006 benchmarks to represent different types of workloads running in the server. We carefully choose the set of benchmarks in which the system resources are consumed at the same level. To ensure exactly the same CPU utilization, we repeatedly and simultaneously run each benchmark with 16 copies to fully utilize

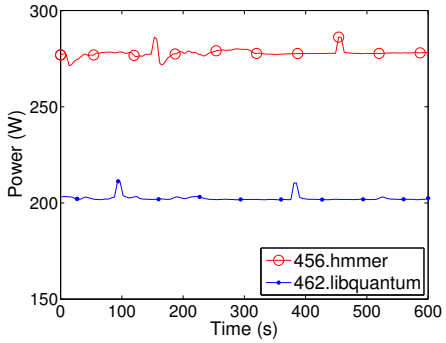


Fig. 5: Power consumption of two benchmarks.

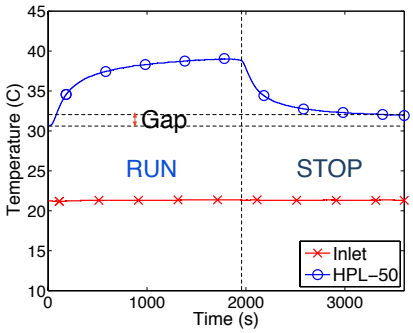


Fig. 6: Temperatures with HPL running and stopping.

all cores. Also, proven by [53], the memory consumption of those benchmarks are quite similar. For benchmark 456.hmmer and 462.libquantum, the average memory utilization is about 24% and 25%, respectively. Moreover, as SPECCPU involves limited I/O activities, the system resources consumed by the chosen benchmarks are very close to one another. The experimental results are illustrated in Figure 4.

We observe that under the same system utilization, different types of workloads could lead to different thermal conditions. An almost 6°C temperature difference is generated by different workloads with the same CPU and memory utilizations. According to [53], 462.libquantum consumes a relatively high memory consumption but produces the minimum outlet temperature increment. By contrast, 456.hmmer can cause a much higher temperature increment than 462.libquantum, and 465.tonto raises the outlet temperature more than 7°C while consuming the least amount of memory. The main reason could be that the types and ratios of instructions composing the benchmarks are different. Although the system utilization is the same, the underlying pipeline flows are actually very different. The ratio of different types of instructions, the probability of branch prediction, and the data dependence could be very different. Those differences further cause CPU halt and leave functional units idle, resulting in generating different amounts of heat. We also observe the zigzag shape of the temperature dynamics of 456.hmmer. The reason is that multiple copies repeatedly run together to keep generating heat; however, they do not finish at the same time. In the gap between the first end of one copy and the start of next round, the system utilization is reduced, resulting in less heat generation.

C. Power Consumption

Heat is generated when currents flow through resistors, obeying the Joule-Lenz law. Consequently, power consumption, which represents the rate of energy transformation, is closely related to heat generation. The heat in joules can be given by:

$$H = I^2Rt = Pt, \tag{1}$$

where I is the current, R is the resistance, and t is the time. From the equation, we can see that the heat generation is linearly increased to the power consumption. Using the results obtained from the same set of SPECCPU 2006 experiments conducted above, we present the power consumptions of 456.hmmer and 462.libquantum in Figure 5. As the benchmark that raises to almost 7°C higher than 462.libquantum, on average 456.hmmer also consumes almost 70W more power than 462.libquantum. Both theory analysis and experimental results indicate that running a power-intensive workload ensures more heat, which could be exploited for mounting a thermal attack. This result (i.e., 70W difference in power consumption) also forms the basis for the configuration of one parameter in our large-scale experimental evaluation.

D. Heating Speed

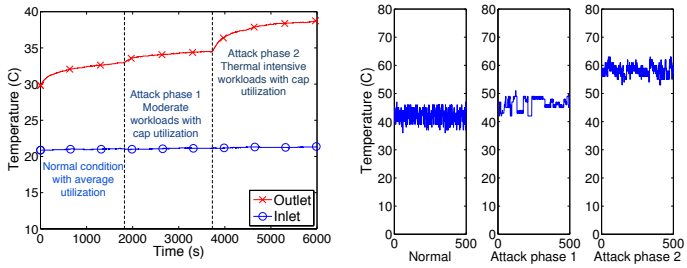
We further measure the speed to heat a server. Unlike a power stimulus that surges instantaneously, the temperature dynamics is a process of heat accumulation. We choose HPL with a block size of 50 as the thermal-intensive workload. We run the workload for 30 minutes and then stop it. In the first 10 minutes, the temperature starts to increase quickly. Then, although the temperature is still raising, the speed drops. The dynamics of the server's outlet temperature is illustrated in Figure 6. The temperature starts to increase quickly in the first 10 minutes. The temperature can increase about 6.8°C higher than that of the idle state. Then, although the temperature is still raising, the speed drops. After we stop the workload, the cooling system can quickly decrease the temperature in the first several minutes. The temperature drops almost 4.5°C within the first five minutes and 6°C within the first 10 minutes. However, after 10 minutes, the speed to dissipate the heat slows down considerably. Even after half an hour, the outlet temperature still cannot return to the original value at the idle state, which implies that the full dissipation of heat requires quite a long time. Overall, we have two major observations: (1) the temperature of a server varies quickly when a thermal-intensive workload starts/stops; and (2) the dynamics of the server temperature is non-linear.

IV. Thermal Attack

In this section, we first describe our threat model. We then mount thermal attacks on both virtualized and non-virtualized environments, as well as a pulsation attack. Based on the attack results, we further conduct damage assessment.

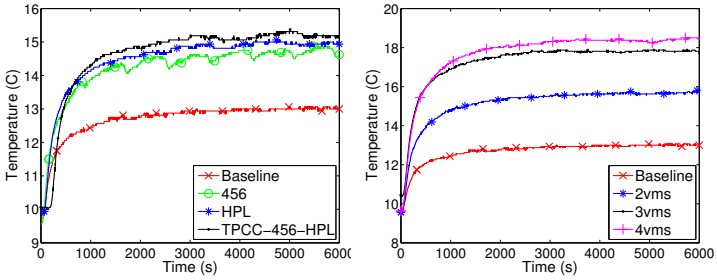
A. Threat Model

A thermal attack can be launched at the server level, rack level, or data center level. We assume that the target data



(a) Outlet temperature (b) Temperature of the 1st core

Fig. 7: Thermal attack on a non-virtualized environment.



(a) One VM. (b) Multiple VMs.

Fig. 8: Margin temperature of thermal attacks on a virtualized environment.

center has the following features. (1) It is cooled by traditional CRAC cooling systems with optimal cooling policies deployed to maximize cooling efficiency, i.e., the supply air temperature is set as high as possible to save energy while keeping the inlet temperature below a redline threshold. (2) It provides utility-based computing services that are accessible over the Internet. (3) Thermal sensors are equipped in the data center, and temperature monitoring is conducted at the rack level. Note that most current data centers have just a few thermal probes for the entire data center, and only some experimental data centers (like HP Labs) have about 2-4 thermal sensors per rack. (4) Like most current data centers, the target data center also deploys power oversubscription.

The attacker could be an individual, a competitor of a cloud service provider, or a cybercrime organization. The attacker does not require more privileges than a regular user to access the target cloud service, and no compromised hypervisor is required. This is mainly due to the fact that in a cloud environment, especially IaaS (Infrastructure as a Service) and PaaS (Platform as a Service), a tenant has the privilege to run any workloads/applications at the guest level, including the benchmarks used in our measurement study. While those workloads are developed to assess the performance of specific scenarios, different combinations and configurations of them compose thermal-intensive workloads.

However, the attacker might utilize the publicly available information or network probing to figure out the network topology inside a cloud [43], and exploit more advanced probing techniques to achieve tenant co-residence in the same physical server or rack [52]. Moreover, the attacker could run advanced thermal-intensive programs (e.g., power virus [22], [23]), instead of regular thermal-intensive workloads, to further exaggerate the heat generation.

Note that a more tangible goal of a thermal attack is not to shut down an entire data center, but to cause a cooling failure in a data center, in which some victim servers are forced to shut down. Under a thermal attack, much more heat will be generated than normal. Once the heat released into the hot aisle surpasses the recyclability of a cooling machine, the inlet temperature increases. The raising of the inlet temperature will further increase the outlet temperature and generate more heat. Such a vicious cycle will finally lead to a cooling failure. Moreover, the overheat caused by thermal attacks will reduce the performance and reliability of victim servers, increase the possibility of hardware failures, and force the data center to reset its cooling configuration, resulting in a much higher cooling cost.

B. Non-virtualized Environments

In this scenario, we assume that an attacker owns a dedicated host (e.g., a dedicated instance in Amazon EC2) in a data center. As the attacker can choose any workload to run at will, the attack vector is straightforward, running thermal-intensive workloads for a relatively long time.

As reported in [15], the average system utilization of most servers in a data center is between 20 - 30%. We also assume that the victim server is running moderate workloads with 25% system utilization. We choose benchmark 462.libquantum to represent a moderate workload. Although 462.libquantum belongs to the CPU-intensive benchmark suite, it is highly vectorizable. The high-dimensional matrix computation requires more I/O operations and makes 462.libquantum consume less power consumption than other SPECCPU benchmarks like 456.hmmer, as shown in Figure 5.

The thermal attack starts after the victim server running the moderate workload for 30 minutes. The attacker first pushes the system utilization to 100%. The first attack phase lasts 30 minutes. After the utilization reaches its cap, the attacker replaces the moderate workload with the thermal-intensive workload to further heat the server. In this attack, we use a combination of SPECCPU 456.hmmer and HPL with a block size set of 50 to represent the thermal-intensive workload. Again, the attack lasts about 30 minutes.

The dynamics of the server's outlet temperature is shown in Figure 7(a). At the beginning, the moderate workload runs at an average of 25% utilization in the server. The outlet temperature of this phase is 32°C. On the next phase, the moderate workload pushes the server into the cap utilization. With fully utilized system sources, the outlet temperature reaches 34°C. Since the difference is less than 2°C, the cooling control system can handle it without any trouble. In the final phase, under the thermal-intensive workloads, the outlet temperature is rapidly raised to more than 38°C, which is 6°C higher than the temperature under the moderate workload. Even under the same utilization, it is evident that malicious thermal-intensive workloads can generate a significant temperature rise and emit a large amount of heat to the computer room. We also show the temperature of the first core in our testbed under the thermal

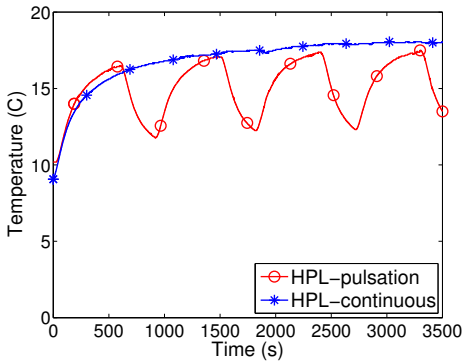


Fig. 9: Margin temperature of pulsation and continuous attacks.

attack on a non-virtualized environment in Figure 7(b), which clearly demonstrates the significant temperature increase of the core under the thermal attack.

C. Virtualized Environments

Most computing services provided by clouds run in virtualized environments, except for dedicated hosting services. A simple approach to mounting a thermal attack is to rent a cloud instance and run thermal-intensive workloads in a VM, regardless of cloud service models (e.g., SaaS (Software as a Service), PaaS, and IaaS). However, only in IaaS can attackers achieve tenant co-residence and fully control the running workloads. Thus, the effective attack vector in a virtualized environment is to subscribe one or multiple VM(s) in one physical host and then run thermal-intensive workloads at the same time.

We emulate the IaaS services by running four VMs on our testbed, each VM allocated with 4GB memory and 4 vCPUs, under Xen hypervisor, which is the same hypervisor running in Amazon EC2. We run all VMs with 25% utilization as our baseline for representing a normal case. Then we select one or more VM(s) with designed workloads running to exhaust the host resources. In our controlled VM(s), we run the following thermal-intensive workloads: (1) SPECCPU 456.hmmer, (2) HPL with the block size set to 200 and N set to 10,000, and (3) a combination of 456.hmmer, HPL and TPCC-UVa. Figure 8(a) shows the attack results.

While the outlet temperature is about 33°C under the baseline condition, some attacks can raise the temperature to almost 37°C. The attacks running HPL, 456.hmmer and TPCC-UVa, which ensure both CPU and I/O intensive workloads, can lead to about a 16°C margin temperature. With just one malicious VM controlled, an attacker can raise the temperature by almost 4°C higher than that of the baseline within 10 minutes.

Recent research shows that it is still relatively easy to achieve tenant co-residence in public clouds. When an attacker can run multiple VMs on the same physical machine, it can mount a more powerful thermal attack. We measure the temperature of the host when different numbers of VMs run thermal-intensive workloads. The results in Figure 8(b) clearly demonstrate that with more VMs controlled, a thermal attack can heat the server to a higher temperature. Also, a thermal attack on three VMs can lead to about a 5°C higher temperature than the baseline case. However, once all CPU resources are already fully utilized in a physical machine (under four VMs in our experiment), running even more VMs in the same machine cannot achieve a significant difference in heating generation.

Attack	CoV
Continuous	0.0071
Pulsation	0.0440

TABLE I: CoV.

In our experiment, we assume that VM live migration is not adopted in the target data center, and the attackers can run thermal-intensive workloads on the target server/VMs to reach full system utilization. This assumption is valid because VM live migration has not yet been widely adopted in real clouds, and live migration suffers non-negligible downtime on the migrated VMs running with intensive workloads. Note that if VM live migration is enabled, the process of live migration can cause a sharp rise of power consumption on both source and destination servers [53]. Such a power spike may trip a circuit breaker and result in a power outage.

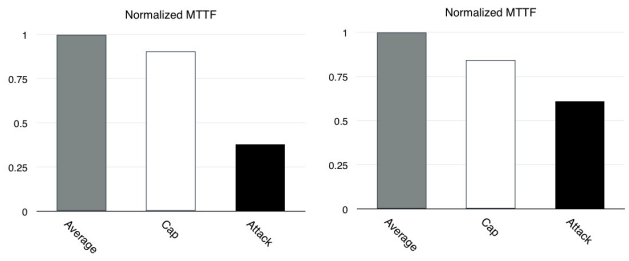
D. Pulsation Thermal Attack

Based on our measurement results, the heat cannot be emitted immediately, and the temperature dynamics is non-linear. Thus, attackers can heat the targeted server within a certain amount of time and then pause for a while, and repeat this on/off heat pattern for many times. We call such an on/off heat strategy a pulsation thermal attack. Using the same HPL benchmark with the block size of 50 and the problem size of 40,000, we mount both a pulsation thermal attack and a continuous thermal attack. The results are shown in Figure 9.

In comparison with the continuous attack, on one hand, the pulsation attack can reach almost the same maximum temperature, with about 0.56°C lost; but over the entire attack duration, the difference of average temperature between these two attacks is about 1.8°C, which means that less heat is generated by the pulsation attack.

On the other hand, in the pulsation attack, the temperature variation can be as large as 6°C in 15 minutes, which is much higher than that of the continuous attack. A recent work [19] reports that a high variability in temperature has a stronger effect on hardware reliability. Based on their observations, the probability of hardware failure is almost double when the CoV (coefficient of variation) in temperature is bigger than 0.0074. We list the CoV in temperature of these two attacks in Table I. It clearly shows that the CoV of the pulsation attack is much larger than both 0.0074 and that of the continuous attack. Thus, a pulsation attack will have a less chance to cause local hotspots and cooling failures but degrade its end-host's hardware reliability more seriously than its continuous counterpart.

Moreover, the attack cost of a pulsation attack is lower than that of a continuous attack, especially for those attacks on SaaS platforms such as web servers. Besides the periodic style,



(a) CPU damage. (b) Disk damage.

Fig. 10: Attack on a Non-Virtualized Environment.

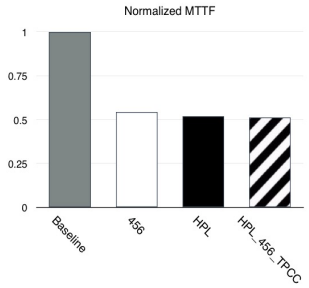


Fig. 11: Attack on a Virtualized Environment.

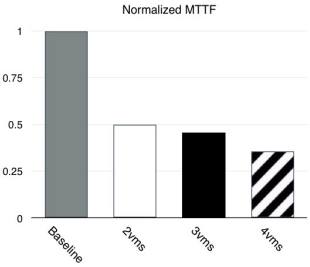


Fig. 12: Attack with various numbers of VMs.

a pulsation attack can also be launched with random intervals to avoid specific traffic patterns, making the pulsation attack harder to detect.

While our current attack vector is to simply run thermal-intensive workloads on IaaS platforms, there are many other potential attack vectors on different platforms (e.g, PaaS and SaaS). For instance, as shown in the previous work [51], random webpage requests can cause significant cache misses and then consume much more energy. Such a feature could also be exploited in thermal attacks. We will explore more effective attack vectors in our future work.

E. Damage Analysis

We use several most common temperature induced intrinsic hard failure mechanisms, including electromigration, time dependent dielectric breakdown, thermal cycling and disk failure [48], [18], [45], to analyze the reliability degradation caused by a thermal attack. We use λ to represent the failure rate of each failure mechanism.

Electromigration (EM). EM is induced by the gradual movement of the ions in a conductor as a result of the momentum transfer between electrons and the diffusing metal atoms [6]. Hardware failures including the opening of metal lines/contacts, shorts between adjacent metal lines, or metal levels or junctions could be caused by EM. Equation 2 gives the EM failure rate (λ_{EM}), which is commonly based on Black's model.

$$\lambda_{EM} = A_0(J - J_{crit})^{-n} e^{(-E_a/kT)} \tag{2}$$

where A_0 is a constant that is empirically determined, J represents the current density, J_{crit} is the threshold current density, E_a is a material dependent constant representing the activation energy, and k is Boltzmann's constant.

Time dependent dielectric breakdown (TDDB). TDDB occurs when the gate oxide breaks down due to low electric fields. It causes the formation of conductive paths through dielectrics. The model is defined in Equation 3. The parameters are similar to those of EM.

$$\lambda_{TDDB} = A_0 e^{\gamma E_{ox}} e^{(-E_a/kT)} \tag{3}$$

Thermal cycling (TC). The large difference in thermal expansion coefficients of silicon substrate and other materials, which causes thermal cycling (TC), eventually leads to permanent failures, such as the creation of cracks, fractures, short circuits, die interface, and more. The effects of low frequency cycling can be modeled via the Coffin-Mason equation [29].

$$\lambda_{TC} = C_0(\Delta T - \Delta T_o)^{-q}, \tag{4}$$

where ΔT is the temperature difference, q is a material dependent Coffin Manson exponent with a common value ranging from 1 to 9. ΔT_o is the portion of the temperature cycle range in the elastic region and could be dropped since it is usually much less than the temperature cycling range.

Disk failure. For damage analysis on a disk due to elevated temperatures, we use the model build in [45], which is derived from the Arrhenius equation. Since our testbed does not contain an embedded sensor for monitoring disk temperature, we analyze the thermal impact on the lifetime of a disk based on the outlet temperature, which should be lower than the actual disk temperature.

$$\lambda_{disk} = Ae^{(-E_a/kT)} \tag{5}$$

We use the reliability model in RAMP (Reliability-Aware MicroProcessors) [48], which assumes that all individual failures are independent. We add up all the individual failure rates and compute the MTTF (Mean-Time-To-Failure) as $1/\lambda$.

Our model is calibrated so that the default MTTF is set to 10 years under normal conditions as previous hardware reliability research such as [30]. While different hardware with different technology could possess different lifetimes, the trend remains similar. Therefore, we normalize all results to minimize the calibration error.

We use the temperature of all CPU cores to estimate the CPU reliability. The normalized results of impacts upon reliability are illustrated in Figure 10, where Figure 10(a) indicates the core damages while Figure 10(b) shows the disk failure. The average bar represents the first phase where the testbed is running moderate workloads with average utilization. The second phase is represented as "cap", which means the utilization has reached the cap utilization. The "attack" bar represents the last phase where thermal-intensive workloads are running. We can see that when the system utilization reaches a certain cap, the increased temperature affects the hardware reliabilities. However, under thermal attacks, the hardware reliabilities are seriously degraded.

Figure 11 shows the impact of thermal attacks on a virtualized environment. Even with just one VM controlled, the reliability of the host suffers much degradation, with

almost a half reduction of the baseline case. Figure 12 shows the dynamics of MTTF under thermal attacks with different numbers of VMs, indicating that the reliability would suffer more degradation with more VMs under malicious control.

Soft errors. High temperature could also negatively affect the soft errors, which will happen if the collected charge at a junction is equal to the critical charge. The critical charge represents the minimum charge to flip the bit in the cell. It was observed that the increase of temperature decreases the value of the critical charge [17], [31], making the occurrence of soft errors more probable.

V. Attacks on Data Centers

The damage of thermal attacks is not limited to victim servers. Thermal attacks also change the surrounding environmental temperature of these victim servers, and thus impacting the thermal conditions of the entire computer room, as well as the thermal performance of other servers and the cooling costs of the data center. To evaluate the impact of thermal attack at the data center level, we conduct a trace-driven simulation based on computational fluid dynamics (CFD), a powerful mechanical fluid dynamic analysis approach that can be used to simulate the air recirculation conditions in the data center. We assume that the simulated data center is equipped with air-cooling CRACs. Also, the data center has adopted a smart workload scheduling policy (e.g., [34]), so that the supply air temperature set point can be set higher for minimizing cooling costs, as suggested in [19].

For the data center layout, we use a standard layout with alternating hot and cold aisles, as illustrated in Figure 1. The cold air goes under the raised floor and enters the cold aisle through perforated tiles to cool down the servers. Note that the standard data center layout has been designed to minimize the air recirculation effect. Other data center structures could result in worse damages under a thermal attack. We assume that the targeted data center has one computer room, which contains four rows of servers, with eight racks in each row. Each rack contains 40 servers, totaling 1,280 servers. The volumetric flow rate of the intake air is set to $0.0068 m^3/s$ for each server. The rate for a CRAC unit to push chilled air into a raised floor plenum is $9{,}000 ft^3/min$. We use a widely adopted CFD package, Fluent [3], to simulate the thermal environment under different workload distributions, and then get the percentage of heat flow recirculated among the servers.

Based on the CFD analysis [49], the air recirculation effect among different servers can be modeled as follows:

$$K_i T^i_{out} = \sum_{j=1}^{N} h_{ji} K_j T^j_{out} + (K_i - \sum_{j=1}^{N} h_{ji} K_j) T_{sup} + P_i, \quad (6)$$

$$T^i_{in} = \sum_{j=1}^{N} h_{ji} * (T^j_{out} - T_{sup}) + T_{sup}, \quad (7)$$

where T^i_{out} represents the outlet temperature of server i. K_i is a multiplicative factor representing the air density and the air flow rate. The air recirculation is described with h. T^i_{in} and T_{sup} represent the inlet temperature of server i and the supply air temperature of CRAC cooling units, respectively. P_i is the power consumption of server i. The outlet temperature of a server is impacted by the air recirculation from server j to server i, the cooling effect of the supplied cooling air, and the power consumption of server i. Equation 7 indicates that the supply air temperature and the recirculation heat, together, determine the inlet temperature. The power consumption also contributes to the thermal condition in the data center.

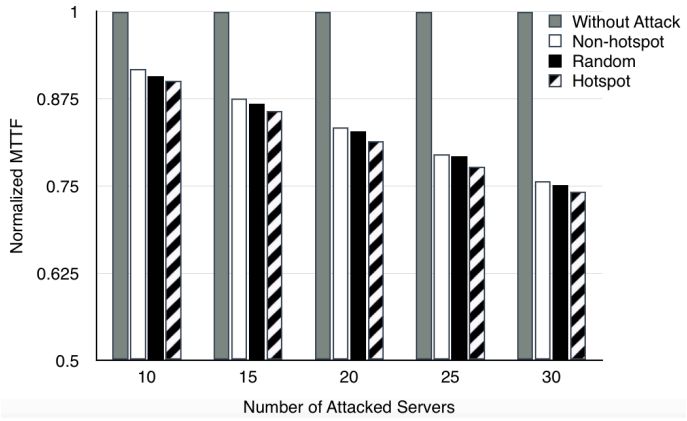
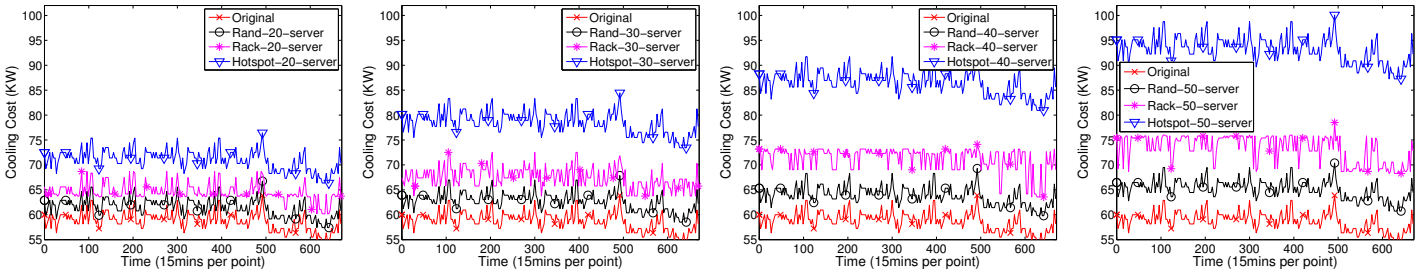


Fig. 13: Thermal attacks on a rack significantly reduces the server MTTF. A hotspot attack is shown to be more effective because of its consideration of air recirculation.

We use a server trace file from one of the largest cloud service vendors in our simulation, which contains the average CPU utilization of 5,415 servers in every 15 minutes for one week. We conduct the thermal attack based on the results from our server-level experiments. We assume the server consumes 100W of idle power and 300W when fully utilized with moderate workloads. As illustrated in Figure 5, by running some thermal-extensive workloads, the power consumption can become higher than that of running moderate workloads, even when the utilization is the same. The difference could be as large as 70W. In our attacks, we conservatively assume that thermal-intensive workload consumes 60W more than the normal cases.

A. Rack-level Attack

We have shown that a thermal attack can significantly raise the temperature of the targeted server. Here we first evaluate the impact of a thermal attack on a server rack. In the rack environment, air recirculation causes servers to affect the temperatures of one another. As a result of heat flows, different rack locations lead to different temperature impacts. For example, servers at the bottom of the rack might have smaller impacts because cold air is supplied from the raised floor and flows upward to the ceiling vents. As result, the bottom servers normally have lower temperatures. To consider the thermal effects of air recirculation, we compare three attack strategies for a rack: (1) hotspot attack; (2) non-hotspot attack; and (3) random attack. In a hotspot attack, we mainly attack servers that have larger impacts on the rack (e.g., at the top of the rack) and so it is easier to create hotspots. In a non-hotspot attack, we focus on servers located in the position with small thermal effects (e.g., at the rack bottom). In a random attack, we randomly choose servers to attack. We attack different



(a) 20 servers under attack. (b) 30 servers under attack. (c) 40 servers under attack. (d) 50 servers under attack.

Fig. 14: Cooling costs in the data center under different types of thermal attacks.

numbers of servers in one rack and leave other servers in the idle state. As introduced in Section IV-E, the temperature rise affects the hardware failure rate. We thus conduct a hardware failure analysis using EM, TDDB, and disk failure to assess the thermal attack on a rack. We use the same parameters and reliability model as Section IV-E.

Figure 13 illustrates the results (normalized to those of conducting no attacks). We can see that the average server MTTF is reduced when more servers are under attack. For example, with about 10 servers under attack, thermal attacks can enlarge the failure rate by $\sim$10% for all servers in a rack, resulting in a normalized MTTF of $\sim$90%. When 30 servers are attacked, the normalized MTTF drops to 75%. Among the three types of attack, the hotspot attack causes the most damage as those positions can maximize the thermal effects. It indicates that air recirculation indeed affects the server thermal conditions even in a single rack.

B. Datacenter-level Attack

1) Attack scenarios: Since the knowledge of air recirculation can make a difference, we now consider three types of attackers. In the first case, the attacker does not have any knowledge about the target data center, such as the position of the servers or the layout of the data center. Such attackers just randomly choose a certain number of servers to launch thermal attacks. To evaluate this kind of attacks, we randomly attack different numbers of victim servers in the target data center, and call them *random* thermal attacks.

Unlike the first type, the second type of attacker owns some advanced networking intrusion knowledge, but still does not know the data center layout. Those advanced attackers could launch more powerful thermal attacks by exploiting side-channel techniques. For example, by repeatedly creating instances and checking IP addresses or using networking probe techniques, an advanced attacker can achieve co-residence on one rack. We assume that the selection of the racks for mounting a rack-level thermal attack is random. To evaluate those advanced attack scenarios, we conduct rack-level thermal attacks on one or a few randomly selected racks with massive thermal-intensive workloads. We call such attacks *rack-level* thermal attacks.

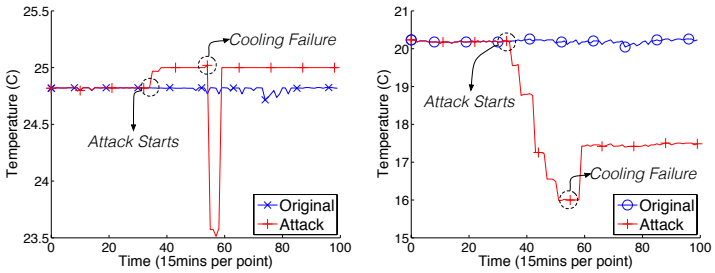
In the third case, attackers may gain much knowledge of the target data center via various approaches. For example, they could know the overall physical layout of the target data center through either public documents or network probing. Also, via an elaborately long-term observation of servers in the target data center, the attackers may roughly infer where the hotspots could be and whether a server is located in a hotspot, as well as the load balancing policy possibly adopted by the data center. Though somewhat costly to acquire, such knowledge can largely expand the damage of a sophisticated thermal attack and be devastating to the data center. We evaluate these well-planned attack scenarios to understand the consequences. In such attacks, we assume that an attacker can roughly pinpoint the positions of the target servers, and call them *hotspot* thermal attacks.

2) Impacts on cooling costs: Thermal attacks not only damage the hardware, but can also generate local hotspots in the computer room. Those servers located in a hotspot suffer higher environmental temperatures than others. A local hotspot can further affect the entire data center. To eliminate any local hotspots, the data center has to decrease the supply air temperature set point. Otherwise, the hotspot might cause cooling system failures. We first investigate the impacts of a thermal attack on the cooling costs. As mentioned above, existing data centers normally deploy aggressive cooling energy saving strategies to reduce cooling costs. The temperature raised by a thermal attack can force the targeted data center to change the cooling conditions, thus increasing the cooling electricity bills. We assume that the targeted data center is equipped with intelligent cooling adjustment mechanisms, where the supply air temperature set point is automatically adjusted as high as possible, in order to keep the inlet temperatures of all servers below the redline threshold. Note that such an intelligent cooling system is not yet adopted in most of today's production data centers due to their incapability of conducting thermal monitoring for individual servers. We assume such a system to investigate the impacts of thermal attack on even a data center with advanced cooling systems.

We estimate the cooling cost based on the power consumption of traditional CRACs, which relies on the cooling coefficient of performance (COP) and the power consumption of all servers. It can be calculated as follows,

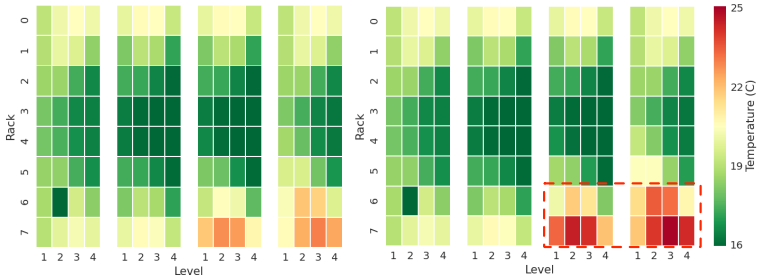
$$P_{CRAC} = \frac{P_{server}}{COP}, \tag{8}$$

where P_{server} is the power consumption of all the servers.



(a) Inlet temperature. (b) Supply air temperature.

Fig. 15: A snapshot of a thermal attack on one block of servers.



(a) Normal condition. (b) Thermal attacks.

Fig. 16: The global views of thermal conditions in a computer room.

COP characterizes the cooling efficiency of a CRAC system. According to [14], the COP could be further acquired based on the supply air temperature, T_{sup},

$$COP=0.0068 \times T_{sup}^2 + 0.0008 \times T_{sup} + 0.458. \quad (9)$$

We conduct three types of thermal attacks (i.e, random, rack-level, and hotspot thermal attacks) on the targeted data center with different numbers of servers under attack. In thermal attacks, the position of the attacked servers plays an important role because of the air recirculation in the computer room. However, only in a hotspot thermal attack, an attacker can have control on the selection of racks and servers for mounting the attack. Due to the random selection of servers or racks in the random or rack-level thermal attacks, we conduct both attacks for 10 times and average the results. For the hotspot thermal attacks, we choose those servers located in hotspots, which have the largest thermal effects on the entire room.

Figure 14 shows the cooling costs under the three types of thermal attacks with different numbers of attacked servers, indicating that thermal attacks can significantly increase the cooling costs. Even with only 20 servers under a thermal attack, which is less than 2% of all servers, the cooling costs increase by about 5∼20% on average. The attack becomes more powerful with more servers under attack. For example, when 50 servers are under attack, the data center has to pay 58% more cooling costs compared to the normal case.

Among those three types of attacks, the hotspot attack again consumes the highest cooling costs compared to the other two attacks. It indicates that the locations of the attacked servers indeed impact the effectiveness of thermal attacks. Due to air recirculation, servers in specific locations can be used to magnify the heat generated by the attack. We can also see that the rack-level attack consumes more cooling costs than the random attack. The reason lies in the fact that, by attacking adjacent servers, the heat can be easily accumulated to form a local hotspot, forcing the data center to lower the temperature set point of the cooling system, resulting in higher cooling costs.

3) Cooling failures: **A snapshot view.** We simulate thermal attacks using the server trace obtained from one of the largest cloud service vendors. The targeted data center is supposed to have a smart scheduler that performs load balancing and adjusts the supply air temperature of CRACs in the room, with the minimum supply air temperature of 16°C [44]. The redline threshold for the inlet temperature is set to 25°C, which is higher than the set points used in most existing data centers. We simulate thermal attacks with thermal-intensive workloads. As shown in Figure 6 in Section III, the temperature can be raised within 10 minutes. In our attack simulation, we gradually increase the number of attacked servers with a period of 15 minutes. The simulation results are presented in Figure 15, where Figure 15(a) shows the dynamics of the inlet temperature of our server block, which contains 10 servers. Figure 15(b) shows the dynamics of the supply air temperature of the CRACs. Before the attack starts, the supply air temperature set point is configured to 20°C. Once the thermal attack starts, the heat generated by the thermal attack forces the CRACs to gradually lower the temperature set point. As more and more servers go under attack, the temperature set point is decreased to 16°C. Finally, the cooling system fails to cool down the hotspot generated by the thermal attack. To avoid hardware damage, 10 servers are forced to shut down. A few minutes later, since no more heat is being generated, the scheduler adjusts the supply air temperature set point back to 17.3°C.

Global views. Figure 16 demonstrates the global views of thermal conditions in the targeted data center. Each block contains 10 servers and four blocks stack up as a rack. The level indicates the position of a block in a rack. For example, the level 4 means the block locates at the top of the rack. The supply air temperature is fixed as 16°C. The attacker first increases the utilization of controlled servers to the capping limit by running moderate workloads. Under this scenario as shown in Figure 16(a), the targeted data center still stays in a health thermal condition: the block with the highest temperature is about 22°C. After that, the attacker switches all moderate workloads to thermal-intensive workloads. As Figure 16(b) shows, although all servers' utilizations remain unchanged, the thermal condition seriously deteriorates. The inlet temperature of multiple blocks at the right corner is approaching the redline threshold, and one of them already surpasses the redline temperature (25°C). Such results indicate that existing utilization-based load balance cannot defend against thermal attacks.

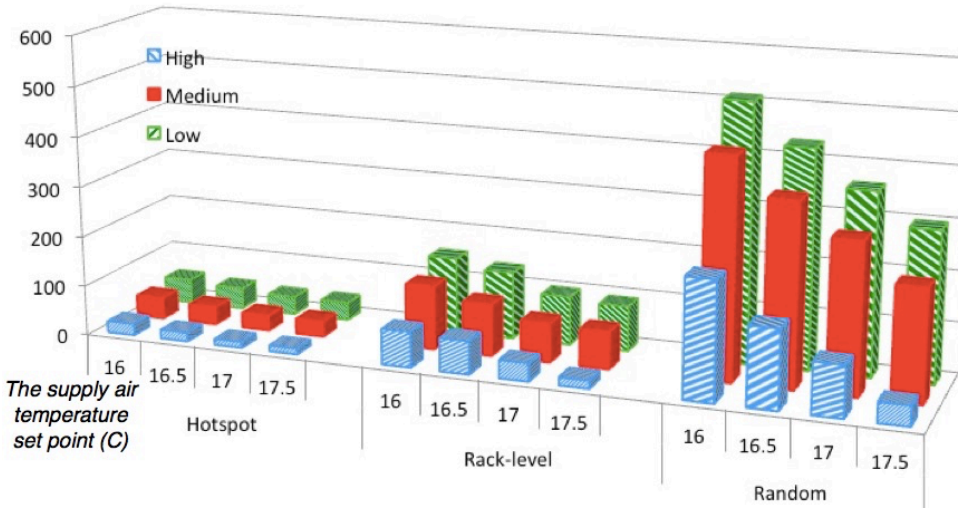


Fig. 17: Number of servers to cause a cooling failure under different supply air temperature.

	Low	Medium	High
Rack-level	608	556	292

TABLE II: Attack efficiency on a large data center.

Attack efficiency. We further explore the efficiency of different attack scenarios. We check the number of servers needed to successfully cause a local cooling system failure. We consider three scenarios: high, medium, and low, representing background workload utilizations around 60%, 40% and 20%, respectively, in the data center. We also consider the impacts of different supply air temperature. While the result for the hotspot attack is deterministic, for the random and rack-level attacks, due to the randomness in the or rack selection, we conduct each type of attack ten times and take the average. All the results are shown in Figure 17. It is clear that, with the same supply air temperature (e.g., 16°C), the hotspot attack requires the smallest number of attacked servers (e.g., 30-50) to cause a local cooling failure. Also, we are not surprised to see that the rack-level attack achieves much higher attack efficiency than the random attack, even though the rack selection is random. For a random attack, a large number of servers are required to cause a local cooling failure. That is reasonable as the cooling facility in a data center is designed to support all servers even if some of them may experience high power consumption at the same time. However, under extreme conditions, it is still possible to cause local hotspots even when the CRAC cooling system is working properly.

The results further suggest that a thermal attack is more effective when the data center workload is in the high scenario, where the data center is already in a hot environment. Thus, by running thermal intensive workloads in the high scenario, it is easier to exhaust the remainder cooling redundancy and cause a cooling failure. As the usage of data centers normally follows specific patterns and might be leaked by side-channels, it is not difficult for a well-motivated attacker to detect the occurrence of the high scenario.

Figure 17 illustrates the number of servers needed to cause a cooling failure under different supply air temperature. It also clearly indicates that the higher the supply air temperature set by data centers, the smaller efforts are needed for attackers to cause cooling failures. More specifically, with the supply air temperature raising by 1°C (e.g., from 16°C to 17°C), the number of required servers is reduced to half. These results imply that although a higher temperature set point of the cooling system is able to significantly lower the cooling costs, as suggested in [19], this management strategy also makes a data center more vulnerable to thermal attacks. While cloud vendors might like a data center hot to reduce cooling cost, malicious attackers will also like the data center hot to much more easily launch a thermal attack. Thus, if today's data centers should continue raising their temperature, thermal attacks would become a serious threat to these future more cooling-efficient data centers.

Extension to a larger scale. We extend the data center model to include 10,240 servers to evaluate the impacts of thermal attacks in a larger scale. For hotspot attacks, we achieve quite similar results as before: with the knowledge of the locations of controlled servers, just tens of servers are enough to cause a local hotspot. We list the results of rack-level attacks in Table II. As we can see, if the attacker can achieve rack-level co-residence, controlling 5% of servers is sufficient to cause a cooling failure. Note that the rack selection is still random in this rack-level attack model. Thus, it is feasible for an advanced attacker to cause a local hotspot and even a cooling failure in a large data center. Moreover, such a potential threat would be more serious in the future given the fact that cloud vendors have been continuing to deploy more servers in their data centers to meet the rapidly increasing cloud service demands.

Financial losses caused by cooling failures. Cooling failures are devastating to cloud vendors and various services hosted in their data centers. Since one physical server could be shared by dozens of VMs in a cloud environment, the shutdown of even a single server would lead to the disruption

of many services and significant financial losses. It is reported that the four-hour outage of Amazon Web Services S3 in February 2017 caused an incredible loss of $150 million for S&P 500 enterprises, plus with a loss of $160 million for U.S. financial services [2]. In 2016, Delta Airline also suffered a loss of $150 million in a five-hour disruption of their data center [4].

C. Attack Cost Analysis

The cost of mounting a thermal attack against a data center mainly lies in subscribing the computing services offered in the data center, in which the dedicated server hosting is the most expensive service and its usage sets the upper-limit of the attack cost. A dedicated server service enables a client to own an entire physical server without any resource sharing with others, allowing a thermal attack to be mounted by fully exploiting its computing sources for heat generation.

The price of an Amazon EC2 dedicated server with 20 Cores is $1.84 per hour [1]. Assume that each thermal attack uses 50/100/400 dedicated servers and lasts 60 minutes as our simulation results show the cooling failure occurs within one hour of the thermal attack being launched. Therefore, to own 50/100/400 dedicated servers for one hour, the attacker just needs to pay $92/$184/$736, for mounting a hotspot attack, a rack-level attack, or a random attack, respectively. However, based on our simulation results, such thermal attacks are powerful enough to cause a cooling failure in a computer room with 1,280 servers.

Compared with service subscription costs, the other costs for mounting a thermal attack is minor. Although it is technically challenging, there is no extra financial cost for conducting advanced network probing to explore the layout of a data center. Even if VPC (Virtual Private Cloud) has been deployed to lower the risk of tenant co-residence, a recent work proposes cost-effective attack strategies to achieve co-residence without much cost. For example, it just costs 14 cents to have a more than 90 percent chance of achieving co-residence on Google Computing Engine [50].

VI. Defense Approaches

The root cause of thermal attacks is the adoption of aggressive cooling energy saving policies in the data centers, which results in heavily reduced cooling redundancies. Although this thermal vulnerability cannot be completely fixed without the increase of cooling redundancies, a well-designed load balancing strategy could mitigate the thermal attacks. Obviously, the traditional utilization-based load balancing strategy cannot effectively defend against a thermal attack, as shown in Section III. The previous temperature-based load balancing approaches (e.g., [40]) relying on a *static* thermal profile running with normal workloads cannot ward off the threat either. This is because the thermal conditions could be dramatically changed due to thermal attacks, even completely opposite to the static pre-calculated profile. The absence of fine-grained sensors further limits the effectiveness of those approaches.

Note that a high power consumption process or high temperature core not only affects itself, but also impacts its neighboring servers. Thus, without a global knowledge of the thermodynamics in a computer room, any single server based capping mechanism (e.g., power capping) alone cannot fully address the security threat posed by thermal attacks.

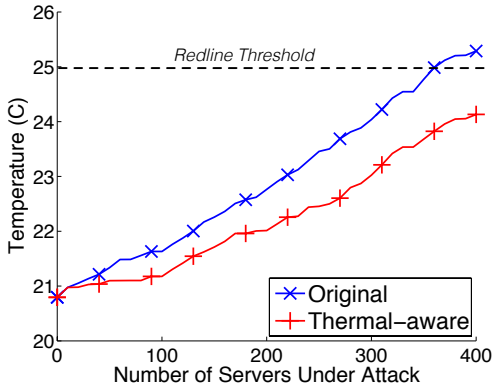


Fig. 18: Inlet temperature of a local hotspot.

To mitigate the thermal attacks, we propose a dynamic thermal-aware load balancing approach at the rack and data center levels. Our thermal-aware load balancing allocates workloads with the consideration of the thermal conditions and physical locations of servers. Based on the thermal condition in a rack or the entire data center, thermal-intensive workloads would be dynamically distributed to servers to limit their thermal impacts. Such a mechanism can effectively mitigate the negative impacts of a local hotspot, allowing further actions like dynamic voltage and frequency scaling to be taken inside the local hotspot for temperature reduction.

We implement a simple prototype of the proposed thermal-aware load balancing mechanism and conduct one experiment to demonstrate its effectiveness. In our prototype, we maintain a list of spots that have large air recirculation effects, especially under thermal attacks, due to their physical locations. The VMs of these spots are less likely to be allocated to users. The configuration of the targeted data center is the same as Section V. The initial system utilization is 25%.

In our experiment, we conduct a thermal attack using thermal-intensive workloads that consume 20% more power consumption than regular workloads. The attacker gradually increases the number of servers under attack. We fix the supply air temperature as 16°C. We measure the inlet temperature of the local hotspot that has the highest temperature under different number of attack servers. The result is illustrated in Figure 18, in which the "+" line indicates the results using thermal-aware load balancing and the "x" line represents the results under the original utilization-based load balancing policy. Whereas the cooling system cannot keep the hotspot below the redline threshold using the original policy, our simple thermal-aware load balancing is able to lower the temperature by more than 1°C and keep the hotspot safe under the thermal attack.

Overall, the proposed thermal-aware load balancing is simple and straightforward to implement, and it can significantly improve the robustness of a data center against thermal attacks. However, in an oversubscribed data center, such a reactive thermal management will still fail to handle much more severe

attacks with much more servers running malicious thermal-intensive workloads simultaneously. In our future work, we plan to pursue more advanced proactive defense strategies to detect the occurrence of thermal anomalies in real-time and prevent potential hotspots from overheating themselves and surrounding servers. The more effective defense systems are expected to have the following capabilities: (1) robust anomaly detection with an emphasis on thermal-intensive workloads at chip and server levels; (2) cost-effective sensing solutions to profile the thermal dynamics by deploying sensors with the consideration of costs, networking traffics, locations and sense ranges of those additional sensors; and (3) proactive thermal management on rack and data center levels by exploiting energy storage such as phase change materials [46].

VII. Related Work

Thermal/power threats. There are some previous works focus on thermal attacks on individual components of machines. Kong et al. [32] studied thermal attacks on instruction caches. They run malicious code to heat other places such as the back end of the cache, instead of those traditional hotspots, and thus avoid the temperature regulation by the dynamic thermal management (DTM). A heating attack on flash memory devices is studied in [47]. A memory cell inside a memory array is locally heated up by an inexpensive laser-diode module. As a result, the contents of the memory can be altered and compromised. Paul et al. [41] launched a thermal attack on disk storages. They used intensive hot seeks that maximize the power consumption of the disk arm to rapidly and repeatedly increase the temperature. The disk is throttled due to DTM. Hasan et al. [26] conducted a heat based attack on Simultaneous Multithreading (SMT) by repeatedly accessing a shared source to create a hotspot in one malicious thread. Thermal covert channel attacks based on CPU cores' temperature variations have been presented in [39], [16]. While these works provide a guidance on thermal attacks on individual components, they do not study the impacts of thermal attacks on servers and data centers.

Wu et al. [51] conducted energy attacks on servers. They manipulated malicious requests that can cause a large number of cache misses and achieved up to 42.3% power consumption increment. Xu et al. [53] demonstrated the vulnerability of cloud services to power attack due to power oversubscription at data centers. An attacker can force victim servers to reach their power peaks at the same time and then trip the circuit breaker. Islam et al. [28] further proposed using a hot air recirculation based thermal side channel to infer the power usage of benign users and then launch power attacks when the aggregated power consumption of benign users is high. Even battery-backed data centers are vulnerable to such attacks [33]. Power attack [24] differs from thermal attack in that power attack attempts to generate a power spike within a very short period and cause a power outage. In comparison with power attack, thermal attack is more stealthy but its damage could be as serious as that of power attack. The overheat induced by thermal attack leads to hardware failure and even shutdown for self-protection.

Cooling in data center. Different cooling strategies in data centers have been proposed in the past [25], [27], [35], [40], [38]. Currently a popular trend is leveraging CFD to simulate the air recirculation in computing rooms [34]. The introduction of CFD helps a data center to place CRAC and servers in an efficient way and thus saves cooling cost. However, those cooling strategies only focus on the optimization of the cooling power in data centers, without considering the risk of a thermal attack.

Impacts of temperature on servers. The impact of temperature on servers has been investigated for years. El-Sayed et al. [19] collected a large amount of field data from different data centers to study the impact of temperature on hardware reliability. They found that temperature variation exhibits strong correlation to hardware failures and server outages. Sankar et al. [45] also used a large collection of data to study the correlation between temperature and hard disk failures. They reported that temperature is strongly correlated with disk failures. Coskun et al. [18] studied the reliability of different job scheduling and power management methods. They proposed a fine grained technique to simulate the thermal behaviors and input the thermal trace to the reliability model. PGCapping [37] integrates various techniques to optimize the multiprocessor performance within a power cap.

VIII. Conclusion

In this paper, we have presented a new security threat called thermal attack on data centers. We first conduct a real server measurement study to systematically investigate the impacts of different factors on thermal conditions of a physical server. Based on the measurement results, we propose effective attack vectors to mount thermal attacks on both virtualized and non-virtualized environments. To evaluate the impacts of thermal attack on a data center, we further simulate rack-level and datacenter-level thermal attacks under three different attack scenarios using a real-world data center trace. We present the damage analysis at both the server and data center levels. Our evaluation and analysis results demonstrate that thermal attack can degrade the performance and reliability of victim servers, cause local hotspots, increase the cooling cost, and even worse lead to cooling failures, in which some servers are forced to shut down for overheat protection. Finally, we present a simple thermal-aware load balancing mechanism to help data centers defend against thermal attacks, which paves the way for more sophisticated defenses.

Acknowledgment

We are very grateful to the anonymous reviewers for their insightful and detailed comments, which help us to improve the quality of this work.

References

[1] Amazon EC2 Dedicated Hosts Pricing. http://aws.amazon.com/ec2/dedicated-hosts/pricing/.

[2] Amazon outage cost S&P 500 companies $150M. https://www.axios.com/amazon-outage-cost-150m-for-s-p-500-companies-2295532812.html.

[3] Computational Fluid Dynamics: ANSYS CFX and FLUENT CFD Software. http://www.caeai.com/cfd-software.php.

[4] Delta Air Lines says the total bill for its devastating computer outage will come to $150 million. http://money.cnn.com/2016/09/07/technology/delta-computer-outage-cost/index.html.

[5] Efficiency: How we do it. https://www.google.com/about/datacenters/efficiency/internal/temperature.

[6] Electromigration. https://en.wikipedia.org/wiki/Electromigration.

[7] Enterprises Still Failing to Cut Data Center Power, Cooling Costs. http://www.eweek.com/it-management/enterprises-still-failing-to-cut-data-center-power-cooling-costs.html.

[8] Go!Temp. http://www.vernier.com/products/sensors/temperature-sensors/go-temp/.

[9] Heatwave, Cooling Failure Bring iiNet Data Center Down in Perth. http://www.datacenterknowledge.com/archives/2015/01/06/heatwave-cooling-failure-bring-iinet-data-center-down-in-perth/.

[10] Overheating brings down Microsoft data center. http://www.datacenterdynamics.com/content-tracks/power-cooling/overheating-brings-down-microsoft-data-center/74543.fullarticle.

[11] Truck Crash Knocks Rackspace Offline. http://www.datacenterknowledge.com/archives/2007/11/13/truck-crash-knocks-rackspace-offline/.

[12] Univesity of Pennsylvania Data Center Overheats. http://www.datacenterknowledge.com/archives/2010/01/20/u-of-penn-data-center-overheats/.

[13] Wikipedias Data Center Overheats. http://www.datacenterknowledge.com/archives/2010/03/25/downtime-for-wikipedia-as-data-center-overheats/.

[14] F. Ahmad and T. Vijaykumar. Joint Optimization of Idle and Cooling Power in Data Centers While Maintaining Response Time. In *ACM Sigplan Notices*, 2010.

[15] L. Barroso and U. Hölzle. The Case for Energy-Proportional Computing. *IEEE Computer*, 2007.

[16] D. Bartolini, P. Miedl, and L. Thiele. On the Capacity of Thermal Covert Channels in Multicores. In *ACM EuroSys*, 2016.

[17] V. Chandra and R. Aitken. Impact of Technology and Voltage Scaling on the Soft Error Susceptibility in Nanoscale CMOS. In *IEEE DFTVS*, 2008.

[18] A. Coskun, R. Strong, D. Tullsen, and T. Rosing. Evaluating the Impact of Job Scheduling and Power Management on Processor Lifetime for Chip Multiprocessors. In *ACM SIGMETRICS*, 2009.

[19] N. El-Sayed, I. Stefanovici, G. Amvrosiadis, A. Hwang, and B. Schroeder. Temperature Management in Data Centers: Why Some (Might) Like it Hot. *ACM SIGMETRICS*, 2012.

[20] X. Fan, W. Weber, and L. Barroso. Power Provisioning for a Warehouse-Sized Computer. In *IEEE ISCA*, 2007.

[21] X. Fu, X. Wang, and C. Lefurgy. How Much Power Oversubscription is Safe and Allowed in Data Centers. In *ACM ICAC*, 2011.

[22] K. Ganesan, J. Jo, W. Bircher, D. Kaseridis, Z. Yu, and L. John. System-Level Max Power (SYMPO): A Systematic Approach for Escalating System-Level Power Consumption using Synthetic Benchmarks. In *ACM PACT*, 2010.

[23] K. Ganesan and L. John. MAximum Multicore POwer (MAMPO): An Automatic Multithreaded Synthetic Power Virus Generation Framework for Multicore Systems. In *ACM SC*, 2011.

[24] X. Gao, Z. Gu, M. Kayaalp, D. Pendarakis, and H. Wang. ContainerLeaks: Emerging Security Threats of Information Leakages in Container Clouds. In *IEEE/IFIP DSN*, 2017.

[25] Í. Goiri, T. Nguyen, R. Bianchini, and Í. Presa. CoolAir: Temperature- and Variation-Aware Management for Free-Cooled Datacenters. In *ACM ASPLOS*, 2015.

[26] J. Hasan, A. Jalote, T. Vijaykumar, and C. Brodley. Heat Stroke: Power-Density-Based Denial of Service in SMT. In *IEEE HPCA*, 2005.

[27] W. Huang, M. Ware, J. Carter, E. Elnozahy, H. Hamann, T. Keller, A. Lefurgy, J. Li, K. Rajamani, and J. Rubio. Tapo: Thermal-Aware Power Optimization Techniques for Servers and Data Centers. In *IEEE IGCC*, 2011.

[28] M. Islam, S. Ren, and A. Wierman. Exploiting a Thermal Side Channel for Power Attacks in Multi-Tenant Data Centers. In *ACM CCS*, 2017.

[29] JEDEC Solid State Technology Association. Failure Mechanisms and Models for Semiconductor Devices. *JEDEC Publication JEP122-B*, 2003.

[30] A. Kahng, S. Nath, and T. Rosing. On Potential Design Impacts of Electromigration Awareness. In *IEEE ASP-DAC*, 2013.

[31] A. Kleinosowski, P. Oldiges, R. Williams, and P. Solomon. Modeling Single-Event Upsets in 65-nm Silicon-on-Insulator Semiconductor Devices. *IEEE Transactions on Nuclear Science*, 53(6), 2006.

[32] J. Kong, J. John, E. Chung, S. Chung, and J. Hu. On the Thermal Attack in Instruction Caches. *IEEE Transactions on Dependable and Secure Computing*, 2010.

[33] C. Li, Z. Wang, X. Hou, H. Chen, X. Liang, and M. Guo. Power Attack Defense: Securing Battery-Backed Data Centers. In *IEEE ISCA*, 2016.

[34] L. Li, W. Zheng, X. Wang, and X. Wang. Coordinating Liquid and Free Air Cooling with Workload Allocation for Data Center Power Minimization. In *USENIX ICAC*, 2014.

[35] Z. Liu, Y. Chen, C. Bash, A. Wierman, D. Gmach, Z. Wang, M. Marwah, and C. Hyser. Renewable and Cooling Aware Workload Management for Sustainable Data Centers. In *ACM SIGMETRICS*, 2012.

[36] D. Llanos and B. Palop. TPCC-UVa: An Open-Source TPC-C Implementation for Parallel and Distributed Systems. In *IEEE IPDPS*, 2006.

[37] K. Ma and X. Wang. PGCapping: Exploiting Power Gating for Power Capping and Core Lifetime Balancing in CMPs. In *ACM PACT*, 2012.

[38] I. Manousakis, Í. Goiri, S. Sankar, T. Nguyen, and R. Bianchini. CoolProvision: Underprovisioning Datacenter Cooling. In *ACM SoCC*, 2015.

[39] R. Masti, D. Rai, A. Ranganathan, C. Müller, L. Thiele, and S. Capkun. Thermal Covert Channels on Multi-Core Platforms. In *USENIX Security*, 2015.

[40] J. Moore, J. Chase, P. Ranganathan, and R. Sharma. Making Scheduling "Cool": Temperature-Aware Workload Placement in Data Centers. In *USENIX ATC*, 2005.

[41] N. Paul, S. Gurumurthi, and E. David. Thermal Attacks on Storage Systems. In *IEEE MSST*, 2006.

[42] R. Raghavendra, P. Ranganathan, V. Talwar, Z. Wang, and X. Zhu. No Power Struggles: Coordinated Multi-Level Power Management for the Data Center. In *ACM ASPLOS*, 2008.

[43] T. Ristenpart, E. Tromer, H. Shacham, and S. Savage. Hey, You, Get Off of My Cloud: Exploring Information Leakage in Third-Party Compute Clouds. In *ACM CCS*, 2009.

[44] P. Sacco. Data Center Cooling Best Practices.

[45] S. Sankar, M. Shaw, and K. Vaid. Impact of Temperature on Hard Disk Drive Reliability in Large Datacenters. In *IEEE/IFIP DSN*, 2011.

[46] 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 *IEEE ISCA*, 2015.

[47] S. Skorobogatov. Local Heating Attacks on Flash Memory Devices. In *IEEE HOST*, 2009.

[48] J. Srinivasan, S. V. Adve, P. Bose, and J. A. Rivers. The Case for Lifetime Reliability-Aware Microprocessors. In *IEEE ISCA*, 2004.

[49] Q. Tang, T. Mukherjee, S. K. Gupta, and P. Cayton. Sensor-Based Fast Thermal Evaluation Model for Energy Efficient High-Performance Datacenters. In *IEEE ISSNIP*, 2006.

[50] V. Varadarajan, Y. Zhang, T. Ristenpart, and M. Swift. A Placement Vulnerability Study in Multi-Tenant Public Clouds. In *USENIX Security*, 2015.

[51] Z. Wu, M. Xie, and H. Wang. On Energy Security of Server Systems. *IEEE Transactions on Dependable and Secure Computing*, 2012.

[52] Z. Xu, H. Wang, and Z. Wu. A Measurement Study on Co-Residence Threat Inside the Cloud. In *USENIX Security*, 2015.

[53] Z. Xu, H. Wang, Z. Xu, and X. Wang. Power Attack: An Increasing Threat to Data Centers. In *NDSS*, 2014.

[54] W. Zheng, K. Ma, and X. Wang. Exploiting Thermal Energy Storage to Reduce Data Center Capital and Operating Expenses. In *IEEE HPCA*, 2014.