ATAC: Ambient Temperature-Aware Capping for Power Efficient Datacenters

Sungkap Yeo Mohammad M. Hossain Jen-Cheng Huang Hsien-Hsin S. Lee

Georgia Institute of Technology

{sungkap, mhossain7, tommy24, leehs}@gatech.edu

Abstract

The emergence of cloud computing has created a demand for more datacenters, which in turn, has led to the substantial consumption of electricity by computing systems and cooling units. Although recently built warehouse-scale datacenters can nearly completely eliminate cooling overhead, small to medium datacenters, which still spend nearly half of their power on cooling, still labor under heavy cooling overhead. Often overlooked by the cloud computing community, these types of datacenters are not in the minority: They are responsible for more than 70% of the entire electrical power used by datacenters. Thus, to tackle the cooling inefficiencies of these datacenters, we propose ambient temperature-aware capping (ATAC), which maximizes power efficiency while minimizing overheating. ATAC senses the ambient temperature of each server and triggers a new performance capping mechanism to achieve 38% savings in cooling power and 7% savings in total power with less than 1% degradation in performance.

Categories and Subject Descriptors C.4 [*Performance of Systems*]: Reliability, availability, and serviceability

General Terms Performance, Reliability

Keywords Cloud Computing, Energy Efficient Data Center, Cooling Power, Performance Capping

1. Introduction

Cloud computing has emerged as a cost-effective way of providing and managing enormous computing power as well as centralizing and synchronizing personal data [6, 31, 37]. Aside from the cost of building the infrastructure of a datacenter, operating a datacenter can be very costly due to the considerable amount of electrical energy required. Accumulated electricity bills over time can easily surpass the cost of the hardware acquisition of the datacenter. Generally speaking, the consumption of electrical power in datacenters is used primarily for computing and cooling. Previous studies in 2003 [19] and 2005 [35] revealed that some datacenters spend more than 50% of their entire power budget on cooling their facilities. Ever since the datacenter community identified cooling as a major source of power inefficiency, the industry has devoted a tremendous effort to reducing cooling overhead in datacenters. For example, Google announced in 2012 that its new datacenter in Finland consumes only about 11% of its entire power budget on cooling; the Yahoo datacenter has reported 7% in 2011; and the Facebook datacenter in 2013 reported 8%. In other words, for these recently built, warehouse-scale datacenters, cooling power no longer poses a major problem.

Although cooling power has ceased to be a problem for new datacenters, which consume electrical power on the order of tens of megawatts, small and medium datacenters still consume enormous amounts of power. In 2011, such datacenters, consuming nearly half of their power budget for cooling, accounted collectively for 72% of total datacenter power [16]. Unfortunately, they cannot adopt cooling techniques commonly used in large datacenters. For example, the Google datacenter in Finland uses cold seawater for cooling, and the Yahoo datacenter uses a warehouse-scale building resembling a chicken coop. Unlike large-scale datacenters, small and medium-scale datacenters are not suitable to take advantage of those geographical resources or economy of scale. Accordingly, researchers focused on developing free-cooling techniques by leveraging on-site evaporative cooling systems. For example, Endo *et al.* [10] or Goiri *et al.* [13, 14] reported significantly improved energy efficiency of small to medium-scale datacenters. With these state-of-the-art technologies, small and medium datacenters can substantially reduce their cooling overhead.

We note that those datacenters still have unsolved problems in accommodating sufficient cooling capacity for the worst-case scenario, which narrow down to (1) power-delivery infrastructures, (2) CRAC units, and (3) reserved

SoCC '14, 3-5 Nov. 2014, Seattle, Washington, USA.

ACM 978-1-4503-3252-1...$15.00.

http://dx.doi.org/10.1145/2670979.2670996

peak-power allowance from power companies. The worst-case scenario assumes a contingency situation under the combination of a high user demand, a high ambient temperature, a high humidity, and near-zero renewable energy production. Without such a contingency cooling infrastructure, the datacenter must compromise reliability and availability. However, we note that such a scenario rarely happens; therefore, if we can safely cope with the extremely rare event without compromising the reliability of computing machines, we will significantly reduce the burden on small to medium datacenters in preparing their cooling infrastructures. To this end, we propose a technique that guarantees reliability of computing machines in the case of temperature emergency by trading off their performance.

As we show in later sections, we also find that our technique, which trades off performance for reliability, greatly improves the cooling efficiency of legacy datacenters where state-of-the-art fresh-air-cooling techniques are not applicable. For example, fresh-air-cooling datacenters are typically located in a intermodal container or a steel box in open places where access to fresh air and immediate heat-exchange are easy. On the contrary, typical legacy small datacenters or server rooms are located inside a building with limited access to fresh air. In addition, some geographical locations may have a higher average ambient temperature than the locations tested in the research papers. Because the average ambient temperature is one of the most important parameters for fresh-air-cooling datacenters, free-cooling options cannot be available depending on the environment where the datacenters are located. In summary, fresh-air-cooling techniques often require radical infrastructural changes and strict environmental requirements. As a result, legacy and future datacenters unsuitable for fresh-air-cooling still carry heavy cooling overhead. Therefore, the purpose of this paper is to identify methods of improving the cooling efficiency, or the ratio between total facility power and productive IT equipment power (tPUE [17]), of those types of datacenters.

Before tackling the cooling inefficiency of datacenters, we first summarize the fundamentals of datacenter cooling. Moore *et al.* [25] found that in removing a given amount of heat, cooling power exponentially decreases as computing room air conditioning (CRAC) units increase the discharge air temperature. In general, a datacenter operating at a higher room temperature achieves a better cooling efficiency. However, increasing the room temperature of a datacenter will violate the thermal guidelines for computing devices, which typically have an upper bound of operating temperature, termed often as *emergency temperature*. Operating above the emergency temperature will affect reliability of computing devices. Therefore, to adhere to thermal guidelines, we must carefully examine increases in room temperature.

Because the operating temperature of a computing device must remain under the emergency temperature, we must monitor every server location because each server maintains a unique temperature, which depends on the regional properties of the locations of the servers. Figure 1 illustrates inlet air temperature variations in the server chassis[1]. The five clusters of bars (from left to right), representing the average server temperature of the ten chassis, correspond to increasing distances from the bottom of the rack. The figure shows that the inlet air temperature of the server chassis at the highest position of the racks is higher than that of the other chassis. Meanwhile, CRAC units must guarantee that all the servers are under the emergency temperature; therefore, they must cool down the datacenter until these servers meet thermal guidelines. In general, by raising the room temperature, datacenters can save on cooling power as long as all the servers are operating under the emergency temperature. To meet the guidelines, typically servers located in lower position in the racks are over-cooled at an inlet temperature of about $10 \sim 20^{\circ}C$ below the emergency temperature.

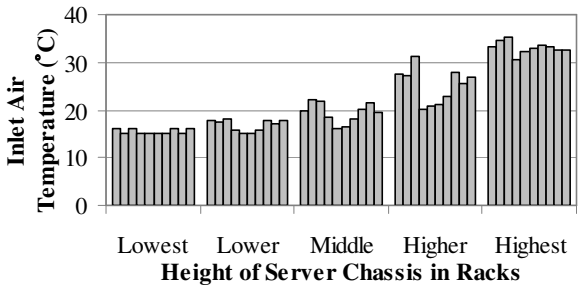


Figure 1: Inlet temperature distribution when CRAC units supply $15^{\circ}C$ air + full server load

In addition, another fundamental piece of datacenter cooling is in CRAC control algorithms. CRAC units without control algorithms must supply the maximum amount of cool air regardless of how busy the datacenters are. However, because the utilization levels of datacenters change dynamically according to the level of human activity, this strategy will usually over-cool the datacenter. Such shortcomings can be alleviated by a dynamic CRAC control algorithm [5, 25]. With a dynamic CRAC control algorithm, CRAC units lower supply air temperature when a datacenter is under heavy workload and raise supply air temperature when it is under light workload. More specifically, CRAC units monitor the inlet air temperature of all the servers. If all of the servers operate under the emergency temperature, CRAC units start to raise the supply air temperature (*i.e.*, they consume less power). Meanwhile, if any server hits the emergency temperature, CRAC units now lower the supply air temperature (*i.e.*, they consume more power). Dynamic CRAC control saves a significant amount of cooling power; however, this

[1] In generating Figure 1, we use the same simulation setup as detailed in Section 4.

control algorithm introduces another source of cooling inefficiency.

The strategy employed in dynamic CRAC control is intuitive and reasonable. However, a recent study [44] found that the simple algorithm discussed above raises the emergency temperature of the servers an average of 1% of the time. The failure scenarios of such cases are as follows. When any server reaches the emergency temperature, CRAC units start to lower the supply air temperature (*i.e.*, they consume more power). However, because delivering cool air takes time, the CRAC units are not able to immediately lower the inlet air temperature of the server. That is, while cool air travels from the CRAC units to the server, the server remains above the emergency temperature. The study also found that to avoid such thermal failures, the CRAC units must maintain a margin of safety by lowering the supply air temperature well before the servers reach the emergency air temperature, consuming 73% more energy.

After a careful review of the fundamentals of datacenter cooling, we determined that a novel method of efficient cooling must have the following features. First, it must perform locally inside a server and not wait for assistance by other means. By doing so, it can effectively eliminate the need for safety margins that dynamic CRAC control algorithms introduce. Second, it must recognize that the inlet air temperatures of datacenter servers vary depending on their locations and that only a partial number of servers suffer from high temperatures. More specifically, it must be triggered only for the servers at *hot spots*, leaving other servers unaffected. Lastly, it must be applicable to both future and already-built small to medium datacenters, the latter of which consume more than half of all the datacenter power today.

With all these critical concepts in mind, we propose a novel system-level approach, ambient temperature-aware capping (ATAC), on a per-server level for datacenters. Implemented in a simple USB-connected sensor with simple fan control software, ATAC can be employed in both already-built and future small to medium datacenters. To prevent thermal emergencies, it allows datacenter servers to run at a higher ambient temperature and applies local dynamic voltage and frequency scaling (DVFS) using its sensed inlet air temperature as input. With such dynamic regulation, the power of CRAC units can be turned down, thereby reducing supply of cool air.

The rest of the paper is organized as follows. Section 2 presents the motivation of the proposed scheme by showing the thermal impact on server and fan power. Section 3 discusses ATAC, and Section 4 describes the simulation platform and specifies the parameters for the modeled datacenter. Section 5 evaluates and analyzes the results. Section 6 highlights the distinction of our paper by discussing other relevant research, and Section 7 concludes.

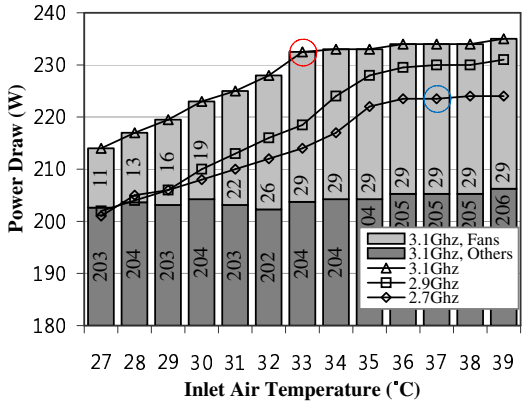


Figure 2: Inlet Temperature vs. Power

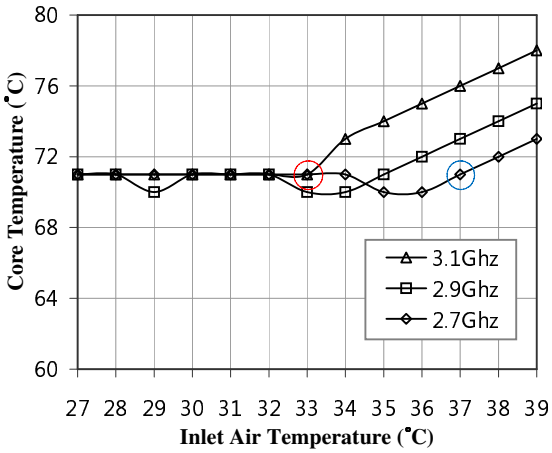


Figure 3: Inlet Temperature vs. Core Temperature

2. Motivation

Before delving into the technical details of the proposed techniques, we present a simple experiment that provides a more thorough understanding of the complex interaction among inlet air temperature ($T_{inlet\ air}$), core temperature (T_{core}), server-level power consumption, and fan speed. For this experiment, we set up a server running at a maximum load enclosed in a controlled area with a thermocouple. During the experiment, we measure the system level power, T_{core} and the fan speed at various inlet air temperatures, as depicted in Figure 2 through Figure 4. Moreover, we repeat the experiment at three core frequencies: 3.1GHz, 2.9GHz, and 2.7GHz.

2.1 Thermal Impact on Server Power

Figure 2 shows the power consumption of the server at various $T_{inlet\ air}$. The three solid lines show system-level power consumption at operating frequencies of 3.1GHz, 2.9GHz, and 2.7GHz while the stacked bars show the power breakdown for only the 3.1GHz run. We focus on the 3.1GHz run for the following analysis. For data points of $T_{inlet\ air} \leq 33°C$ shown in Figure 2, server-level power increases mostly due

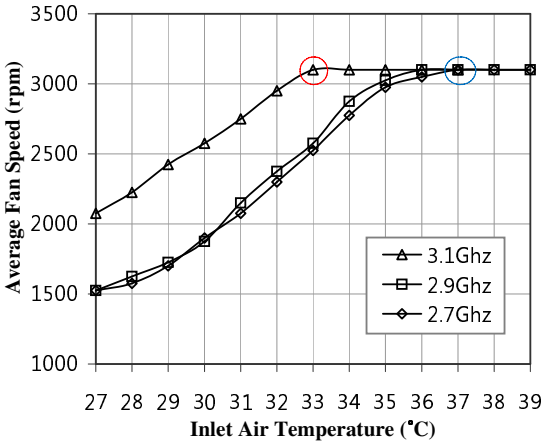


Figure 4: Inlet Temperature vs. Fan Speed

to the increase in the fan power; therefore, these data points follow the trend of the increase in the fan speed, shown in Figure 4. As a result, Figure 3 shows that the core temperature remains unchanged around 71°*C*. A previous study [44] also pointed that fan power is not constant and that ignoring increased fan power could dramatically affect the effectiveness of energy-saving strategies. Once the fan speed reaches the maximum (3100 rpm), the core temperature begins to rise, and the upward trend of the power in Figure 2 slows down. The slight power increase in this region ($T_{inlet\ air} > 33°C$) is likely considered as a result of increased leakage current caused by higher core temperature.

We now compare the results of the runs at different frequencies. First, when running at lower frequencies (e.g., 2.7GHz and 2.9GHz), the system does not attempt to cool down the core temperature as shown in Figure 3. Instead, to reduce the fan power consumption, it lowers the fan speed (Figure 4). As shown in Figure 4, the system can save 13.5W at 2.9 GHz and 18W at 2.7GHz from the power rating of 232W for 3.1GHz at 33°*C*.

According to Newton's law of cooling, the rate of heat loss (in watts) is proportional to the temperature difference between an object and its surroundings. We now apply this theory to the measured power numbers in Figure 2. To eliminate the effect of the fan power and the difference in core temperatures, we pick the data points of two systems when the fan reaches its maximum speed with the same core temperature. As indicated by the circles in Figure 2, the runs of 3.1GHz and 2.7GHz reach that state when $T_{inlet\ air} = 33°C$ and $T_{inlet\ air} = 37°C$, respectively. As Figure 3 shows, the core temperature of 71°*C* is the same in both scenarios. Then the temperature differences between the core and its surroundings (*i.e.,* $T_{core} - T_{inlet\ air}$) are $38°C(= 71-33)$ and $34°C(= 71-37)$ for the 3.1GHz and 2.7GHz core, respectively. The 3.1GHz core has an advertised thermal design power (TDP) of 80W; in other words, rotating the fan at a maximum speed, the cooling system can remove heat generated by an 80W core when the delta temperature is 38°*C*. Based on the law of cooling, the 2.7GHz system will remove heat generated by a 71.6W ($= 80W \times \frac{34°C}{38°C}$) core. The results of the measurements of these two systems, shown in Figure 2 (*i.e.*, the power difference between two dashed circles), reveal a 9W difference, which closely conforms to the theoretical deduction of 8.4W.

By using the relationship discussed above, we now illustrate how to keep the core temperature under control while the inlet temperature exceeds the emergency temperature ($T_{emergency}$). Initially, we assume a server whose temperature difference between the core ($T_{core} = 70°C$) and the ambience ($T_{inlet\ air} = 30°C$) is 40°*C* when the inlet temperature is 30°*C*. Now we tune down the cool air supply from the CRAC unit, and the server subsequently senses that the $T_{inlet\ air}$ has risen to 35°*C*, which is 5°*C* above $T_{emergency}$. In other words, the temperature difference (ΔT) between the core and the ambience declines to 35°*C*. According to our previous discussion, as the fan has reached its maximum rotation speed, the server must increase its core temperature by 5°*C* to 75°*C* to achieve equilibrium, which affects reliability. Another option for the server is to reduce its power consumption to maintain a core temperature of 70°*C*. Based on our prior deduction, to achieve this goal, the power draw must decrease proportionally. Therefore, to keep the core temperature from rising, the server has to reduce power to $\frac{35}{40}$th of its original power using a technique such as DVFS.

2.2 Thermal Impact on Fan Power

To build a link from $T_{inlet\ air}$ to fan power, we adopt an approach similar to that in prior studies [28, 44]. First, we use *Fan Affinity Laws*, which indicate that (1) the fan power is in cubic growth of the rotational speed; and (2) the volume capacity (the amount of air) of a fan is proportional to the rotational speed. Thus, the following relationships hold.

$$\begin{aligned} Fan\ Power &\propto (RPM)^3 \\ Volume &\propto RPM \\ Fan\ Power &\propto (Volume)^3 \end{aligned} \tag{1}$$

Second, we use the *Laws of Convective Heat Transfer*, which indicate that heat transfer or power (in watts) is proportional to (1) the volume capacity of air, and (2) the temperature difference between T_{core} and $T_{inlet\ air}$, or ΔT.

$$\begin{aligned} Heat\ Removal\ (\text{Power, in watts}) &\propto Volume \\ Heat\ Removal\ (\text{Power, in watts}) &\propto \Delta T \end{aligned} \tag{2}$$

Therefore, when the temperature difference ($\Delta T = T_{core} - T_{inlet\ air}$) becomes half of what it was, the volume capacity must double to maintain the cooling capacity.

$$\begin{aligned} \frac{Heat\ Removal\ Per\ Volume_{before}}{Heat\ Removal\ Per\ Volume_{after}} &= \frac{\Delta T_{before}}{\Delta T_{after}} = 2 \\ \text{To make } Heat\ Removal_{before} &= Heat\ Removal_{after} \\ \therefore Volume_{after} &= 2 \times Volume_{before} \end{aligned} \tag{3}$$

Since the volume capacity of a fan is proportional to the rotational speed, a halved ΔT will result in double the rotation speed. The fan now rotates twice as fast and consumes 8x more power.

$$\begin{aligned} \frac{Volume_{after}}{Volume_{before}} &= \frac{RPM_{after}}{RPM_{before}} = 2 \\ \therefore \frac{Fan\ Power_{after}}{Fan\ Power_{before}} &= (\frac{RPM_{after}}{RPM_{before}})^3 = 8 \end{aligned} \quad (4)$$

In summary, a higher $T_{inlet\ air}$ results in a smaller ΔT and increases the fan power.

3. ATAC: Ambient Temperature Aware Capping

In this section, we propose ATAC (ambient temperature-aware capping), a system-level technique that guarantees the reliability of operations when we tune down the cooling units to improve energy efficiency. Our proposed scheme enables the inlet air supply to furnish less cooling air to save cooling energy while applying ATAC, which enables each server to dynamically scale down its frequency and voltage (*i.e.*, capping the performance). Local to each server, the ATAC mechanism collects information that includes the temperature of the core and the inlet air, the rotational speed of fans, and the thermal design power (TDP) of the CPU, and then decides to initiate performance capping by checking whether the inlet air temperature ($T_{inlet\ air}$) is above the emergency temperature ($T_{emergency}$).

In dynamic CRAC control, CRAC units raise the discharge air temperature until the highest inlet air temperature of a server reaches $T_{emergency}$. When any of the servers experience $T_{emergency}$, CRAC units begin to lower the supply air temperature. However, the server should remain over $T_{emergency}$ until cool air from the CRAC units reaches the server; such a thermal failure has been identified as one of the most important sources of cooling inefficiency. In this scenario, ATAC continuously monitors the inlet air temperatures from each server, obtained using the thermal sensor embedded in the servers. If $T_{inlet\ air}$ remains below $T_{emergency}$, the triggering event does not occur. Otherwise, to reduce the power consumption, ATAC of the violating server will cap its own performance by scaling down its frequency/voltage. To keep T_{core} under control, ATAC assures that the power proportionally decreases with the temperature difference ($\Delta T = T_{core} - T_{inlet\ air}$) based on the discussion in Section 2.1.

We now discuss the relationship among the operating frequency, performance, and power for the design of an effective performance-capping mechanism. In general, performance is not proportionally degraded by a reduction in the operating frequency. To illustrate this point, we evaluate CloudSuite 2.0 [12] at various CPU frequencies including 3.1Ghz, 2.9Ghz, and 2.7Ghz, and measure the performance of each cloud-computing application.

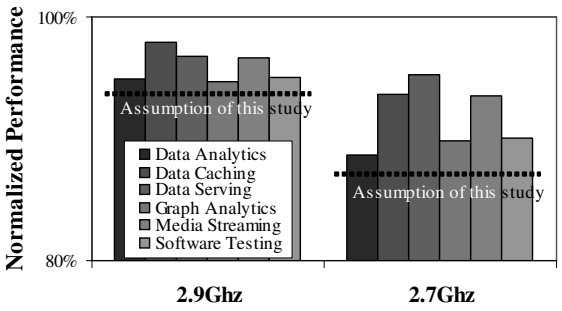


Figure 5: Performance of 2.9Ghz and 2.7Ghz run

Figure 5 illustrates the performance of the 2.9Ghz and 2.7Ghz runs normalized to that of the 3.1Ghz run. In the figure, the dashed lines at 93.5% and 87.1% depict simple ratios between the operating frequency (2.9Ghz and 2.7Ghz) and the baseline (3.1Ghz). All CloudSuite applications we present in Figure 5 outperform the dashed lines, or the simple ratio between the frequencies. However, in this study, we conservatively assume that the performance of a server at a lower frequency proportionally degrades, as the dashed lines in Figure 5 show. For the relationship between operating frequency and power, we adopt a general power and performance model from other studies [30, 43], in which the power reduction is equal to the square of the performance reduction. Based on these assumptions, if the operating frequency of the core decreases by 90%, the performance will degrade by 90%, and its power consumption will decrease to 81% ($= (0.90)^2$). For the following evaluation of ATAC, we use this conservative assumption.

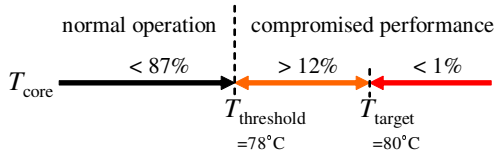
All in all, Algorithm 1 summarizes the capping mechanism of ATAC. Firstly, ATAC assumes that it knows two static parameters, $T_{emergency}$, and ΔT, the temperature difference between T_{core} and $T_{inlet\ air}$ when the CPU fan is rotating at maximum. Now ATAC takes a dynamically measured parameter, $T_{inlet\ air}$ into account. If we use the same parameters explained at the end of Section 2.1, for example, then $T_{emergency} = 30°C$ and $\Delta T = 40°C$. Secondly, ATAC triggers power capping when $T_{inlet\ air}$ exceeds $T_{emergency}$. In the previous example, $T_{inlet\ air}$ became $35°C$ or $5°C$ above $T_{emergency}$ (*i.e.*, $\alpha = 5°C$ in Algorithm 1). Now ATAC caps the CPU power by the amount of $\frac{\Delta T - \alpha}{\Delta T}$ according to Equation (2). Therefore, the previous example reduced power to $\frac{35}{40}$th ($= \frac{40°C - 5°C}{40°C}$) of its original power. The relative performance of the CPU under ATAC becomes $\sqrt{(\Delta T - \alpha)/\Delta T}$, as discussed in Figure 5.

Although ATAC is designed to eliminate thermal failure caused by timing delays in delivering cool air, we also argue that more aggressive ATAC policies could further reduce datacenter power consumption with negligible performance degradation. In previous sections, we found that servers have different $T_{inlet\ air}$ depending on their locations, and few servers suffer from high $T_{inlet\ air}$. Therefore, by aggressively raising the room temperature of a datacenter to intentionally

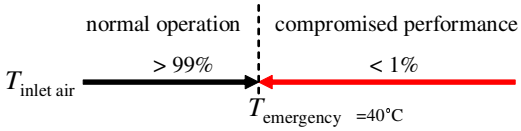
Algorithm 1 ATAC Algorithm

$T_{inlet\ air} \Leftarrow$ *Measured inlet air temperature*
loop
  if $T_{inlet\ air} > T_{emergency}$ **then**
    $\alpha \Leftarrow (T_{inlet\ air} - T_{emergency})$
    CPU Power capping $\Leftarrow max$(CPU Power) $\times \frac{\Delta T - \alpha}{\Delta T}$
    Relative Performance $\Leftarrow \sqrt{\frac{\Delta T - \alpha}{\Delta T}}$
  else
    CPU Power capping $\Leftarrow max$(CPU Power)
  end if
end loop

let some servers operate above $T_{emergency}$, the datacenter can save a significant amount of cooling power by sacrificing the performance of only a small number of servers. The servers operating over $T_{emergency}$ will trigger ATAC, and the T_{core} of such servers will remain under control. Details of more aggressive ATAC will be discussed in Section 5.2.



(a) A case of Dynamic Thermal Management (DTM)



(b) A case of the proposed technique (ATAC)

Figure 6: Chances of compromising performace of servers

Because it can be implemented by a simple thermocouple connected through USB and OS-level software support, ATAC is a practical, viable solution to future and already-built small to medium datacenters. Moreover, ATAC outperforms and differs from dynamic thermal management (DTM) [9], which does not take $T_{inlet\ air}$ into its control loop. More specifically, when DTM targets T_{core} under T_{target} in Figure 6a, it must secure a temperature margin, $T_{threshold}$ ($< T_{target}$), and start to lower the DVFS level when T_{core} reaches $T_{threshold}$. However, at high ambient temperature (HTA) datacenters, the temperature of T_{core} is more likely to fall in-between $T_{threshold}$ and T_{target}. For example, when we secure $2°C$ temperature margin for DTM then $T_{threshold} = 78°C$ and $T_{target} = 80°C$. In such a case, T_{core} in our baseline experiment, which we will discuss in the following sections, spends more than 12% of the time in this safety margin where the performance of servers must be degraded. In contrast, as depicted in Figure 6b, ATAC degrades the performance of servers only when $T_{inlet\ air}$ exceeds $T_{emergency}$, which has significantly less chances than $T_{core} \geq T_{threshold}$.

Nonetheless, CPUs must employ DTM support even with ATAC since ATAC is not useful in abnormal emergencies such as failure of cooling fans or accidental removal of heat sinks. DTM is indispensable in securing the most robust reliability preventing chips from being burned and melted down; however, the ATAC mechanism manages system-level power and performance more effective in setting a preferred temperature range of microprocessors for power efficient datacenters.

4. Simulation Setup

4.1 The Simulation Setup

In this paper, we use modified SimWare [44] as an evaluating platform. SimWare implements a variety of critical components of a datacenter in a holistic way [34] including detailed server power models, cooling power models [25], the effect of heat recirculation [40], and the effect of the timing delay of cool air delivery from the CRAC to the front plate of servers. The evaluating platform, SimWare, has been modified and configured as follows.

[Dynamic CRAC control algorithm] CRAC begins to supply cool air at the lowest possible temperature and raises the temperature until the inlet air temperature of any server reaches a triggering temperature, $T_{trigger}$. Upon such an event, CRAC begins to lower the supply air temperature to cool down the room temperature. In general, the inlet air temperature of a server is computed as [26]

$$T_{inlet\ air} = T_{supply\ air} + T_{recirculated\ heat}. \quad (5)$$

Here, $T_{recirculated\ heat}$ represents the thermal impact caused by the heat recirculation of the other servers. Note that this heat recirculation effect is the primary reason why $T_{inlet\ air}$ varies according to the location of the servers. In addition, the goal of the dynamic CRAC control can be expressed as

$$\forall T_{inlet\ air} < T_{trigger}. \quad (6)$$

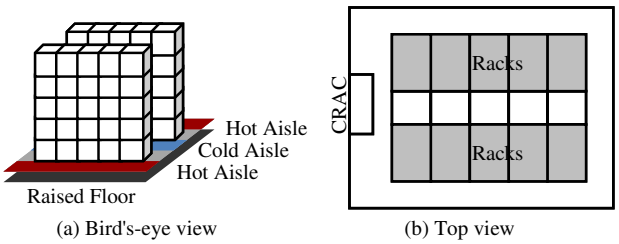
Throughout this paper, we use this dynamic CRAC control with only one configurable variable, $T_{trigger}$, for all simulations. The remaining configurable parameters for SimWare are discussed in the following sections.

[Cooling power model] A prior research [25] found that the efficiency of CRAC units change by their supply air temperature (T_{supply_air}). More specifically, the study showed that the efficiency of CRAC units improves when they operate at higher T_{supply_air}. This relationship between CRAC power and T_{supply_air} is expressed as

$$\frac{\text{Amount of heat removed (W)}}{\text{Power draw from CRAC units (W)}} = 0.0068T^2_{supply\_air} + 0.0008T_{supply\_air} + 0.458. \quad (7)$$

From Equation (7), if CRAC units operate at a higher discharging temperature (T_{supply_air}), they consume less power

while removing the same amount of heat. We employ the same CRAC power model in the evaluation.



(a) Bird's-eye view (b) Top view

Figure 7: Simulated Datacenter Layout

[Datacenter layout] The simulated datacenter hosts 50 blade server chassis, and a rack holds five blade server chassis; therefore, the datacenter has ten racks as shown in Figure 7. Five racks form a row of racks, and we have two rows of racks in the datacenter. In addition, the simulation assumes a raised floor and hot/cold aisle layout where the cold aisle is located in-between the rows of racks. Hot and cold aisles are not partitioned; therefore, heat generated by the servers may recirculate to the cold aisle or front plate of other servers. In modeling such heat recirculation effect, SimWare employs *heat distribution matrix* (HDM) [40], which converts heat dissipation from a blade server chassis to the inlet air temperature increase of all other chassis. Because we assume fifty blade server chassis, HDM is a 50 by 50 matrix.

4.2 Specifications for Blade Servers

A blade server chassis in our simulation hosts 16 blade servers. Therefore, the simulated datacenter has 800 (=50 blade server chassis $\times$ 16 servers/chassis) blade servers, and we assume that each blade server houses a single 16-core AMD Opteron 6386SE. This configuration simulates a total of 12,800 cores (800 blade servers $\times$ 16 cores/server). Table 1 summarizes the specifications of the blade servers in our simulated datacenter.

We also elaborate on the detailed specifications of the fan attached to each blade server. First of all, the fan consumes a maximum of 15W and removes heat generated by 140W when the fan rotates at the maximum speed of 3000 rpm We also assume that when the fan removes the maximum power, 140W, the minimum temperature difference (ΔT) between the die and the inlet air is $40°C$, generated from our experiment discussed in Figure 3 in which the core is at $71°C$ and the inlet air temperature is measured at $33°C$ when the fan rotates at full speed. For simplicity, instead of $38°C (= 71°C - 33°C)$, we use $40°C$, a number particularly important for performing ATAC. As discussed in Section 2.1 and Section 2.2, we use the temperature difference to calculate the desired power level to be achieved by DVFS. An example of how to reach the desired power level appeared at the end of Section 2.1.

Two other fans with the same specifications are used in the server. One is located at the front panel of the server and the second one at the back. The rotational speeds of these fans are directly proportional to the power consumption of the server and the temperature of the inlet air. For simplicity, the boundary condition is that the fans are rotating at 3,000 rpm (maximum) when the server is fully loaded at $30°C$. In addition, assuming that the goal of fan control is to save fan power, we set the die temperature lower than $80°C$ for reliability. In terms of the peak power of the blade server, we add the power of the idle, peak CPU, and all three fans. We first assume that the blade server, when idle, consumes half of the peak power, 140W [4]. Then the peak power becomes 140W (idle power) + 140W (peak CPU power) + $3\times$15W (three fans) = 325W.

4.3 Google Cluster Data as an input

We use Google cluster data (GCD) [32, 33] as the input to SimWare. Google released GCD, one of the most detailed utilization traces, to the public in 2011. It comprises 178GB of text files containing detailed information collected from jobs submitted to one of the company's datacenters. The overall computing cluster has about 12,500 heterogeneous computing nodes in ten groups. Although the groups have disparate hardware specifications, we regroup the nodes into three groups based on the CPU performance metric, for current SimWare models only the power consumption by CPU utilization of a server, not by memory or disk utilization. In terms of normalized CPU performance, servers in GCD can be categorized into three different types: 0.25, 0.5, and 1. By contrast, our simulated datacenter contains homogeneous servers (*i.e.*, the servers share the same computing capacity). Since more than 92% of the servers in GCD have a normalized CPU scale of 0.5, we assume that a CPU scale of 0.5 matches one core in the simulated datacenter. For servers with a CPU scale of 0.25 or 1, we assume linearly decreased or increased execution time, respectively. For example, one second in a machine with a CPU scale of 0.25 corresponds to a half second in a machine with a CPU scale of 0.5.

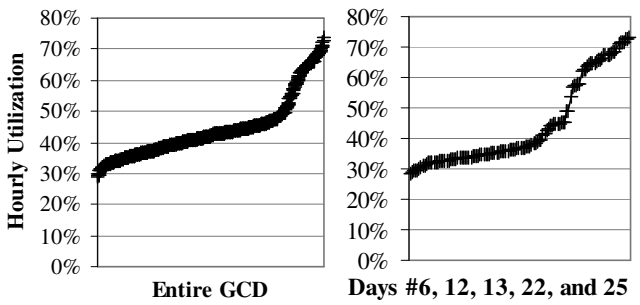


Figure 8: Distribution of Hourly Utilization

For a particular configuration, a single simulation run takes about ten days because of the massive volume of GCD

Table 1: Specification of our Simulated Blade Server.

Component name	*Specification*
CPU	16 cores AMD Opteron 6386 SE. TDP=140W. Heat capacity is assumed to be 22.5 $m^2 \cdot kg \cdot s^{-2} \cdot K^{-1}$
CPU Cooling Capacity	CPU fan removes heat generated by 140W when the fan rotates at the maximum speed
$\Delta T = T_{core} - T_{inlet\ air}$	When the fan rotates at its maximum speed and the CPU is at full load, the temperature difference between the processor die's temperature and the ambient air is $40°C$.
CPU Fan	Maximum speed = 3000 rpm; power = 15W.
Other Fans	Two more fans with the same specification are located at the front and back of each server.
Fan Control	When $T_{core} < 80°C$ the priority of the fan control is in saving fan power. Otherwise, when $T_{core} \geq 80°C$, the priority is in lowering T_{core}. The CPU fan cannot be turned off and runs at 500 rpm when the server is idle. Case fans increase their rotational speed proportional to the power consumption of the server and the inlet air temperature.
Idle Power	The blade server consumes 140W plus corresponding fan power when idle.
Peak Power	The blade server consumes 140W + 140W + 3×15W = 325W in maximum.

trace and the complex simulation methodology incorporated in the SimWare framework. To expedite the simulation time and to explore a wide spectrum of various configuration parameters, we sample data from a 29-day trace. More specifically, we tune SimWare to the experiment with data from days 6, 12, 13, 22, and 25. Out of these five samples, we select day 12 because it observes the lowest hourly server utilization and day 22 because it experiences the highest hourly server utilization. We select the remaining three days such that the average datacenter-level utilization for sampled trace (44.5%) closely matches that of the entire GCD trace (44.6%). More specifically, Figure 8 shows hourly datacenter-level utilization from the entire GCD trace and from the five days we selected for this study. In Figure 8, each dot represents an hour of operation, and all of the dots are sorted according to the utilization level. These two curves, which have the same minimum value from day 12 and the same maximum value from day 22, exhibit a reasonably similar trend.

5. Evaluation and Analysis

5.1 The Baseline Analysis

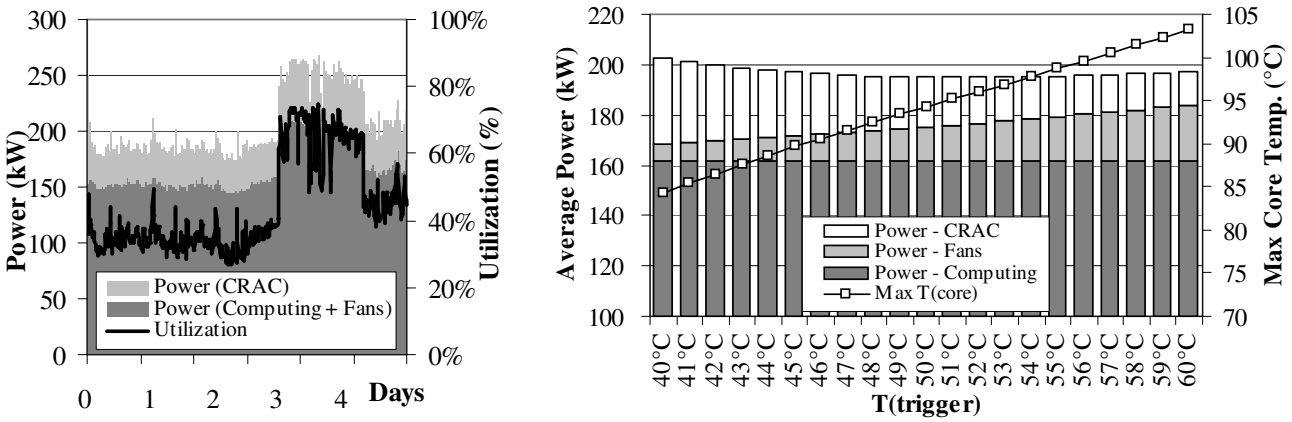
The legacy datacenters that this study targets typically use a $T_{trigger}$ value from $20°C$ to $30°C$ with an average $T_{supply\ air}$ of about or even lower than $15°C$ [1, 27, 36]. However, Intel projects that in the future, high ambient temperature (HTA) datacenters will enable servers to operate at above $40°C$ and even more than $50°C$. Although the baseline of $T_{trigger} = 20°C \sim 30°C$ outshines the potential benefits of ATAC, we assume that the baseline datacenter is HTA-ready and that $T_{trigger} \geq 40°C$.

Figure 9a shows the overall utilization level of the simulated datacenter when $T_{trigger} = 40°C$. The X-axis represents the elapsed time while the primary Y-axis (left) and the background area chart shows the power consumption in watts. In addition, the secondary Y-axis (right) and the solid line chart represent the utilization level. As stated before, our truncated GCD contains job traces for five days, and the average daily utilization level ranges from 27% to 75%. In general, the power consumption curve for computing and cooling units track the utilization level. Figure 9b reveals interesting information: It shows that when we increase $T_{trigger}$ from $40°C$ to $60°C$ in the X-axis, the datacenter consumes less power on cooling while expending more power to blow the fans harder. Thus, the net power savings continue to decrease until $T_{trigger}$ reaches $51°C$. Beyond this inflection point, the consumption of fan power offsets the savings in cooling power.

One of the important implications of higher $T_{trigger}$ is higher T_{core}. As illustrated on the secondary Y-axis in Figure 9b, the all-time highest value of T_{core} increases as room temperature increases. When the fan is not at maximum rotational speed, a system can hold T_{core} even at a higher $T_{inlet\ air}$ by increasing the fan power. However, in rare situations, the following three conditions can occur simultaneously. The fans are initially at maximum rotational speed. Then, the CPU is operating at full load. Finally, $T_{inlet\ air}$ exceeds $T_{emergency}$. When all three conditions are met, T_{core} rises to maintain the temperature difference ($= \Delta T$) between T_{core} and $T_{inlet\ air}$ constant. In fact, for such a rare failure scenario, our proposed ATAC breaks the second condition by sensing $T_{inlet\ air}$ and changes the DVFS state, so the CPU cannot be fully utilized. As we show in the next section, ATAC initiates performance capping only for a small fraction of time, so the overall responsiveness of the datacenter remains nearly the same while the maximum T_{core} drastically decreases.

5.2 Evaluating ATAC

In applying the ATAC mechanism to the servers, a datacenter administrator can control the level of aggressiveness of ATAC. For example, in Power Capping [20], the administrator can set a server with a 1,000W name plate to consume 900W or even 800W. When the power consumption of the server is capped at 800W, its performance is lower than when it is capped at 900W. Similarly, an administrator can configure the aggressiveness of ATAC. Aggressive ATAC will activate performance capping more often.



(a) Utilization trace

(b) Average power and core temperature

Figure 9: Google Cluster Data in 2011

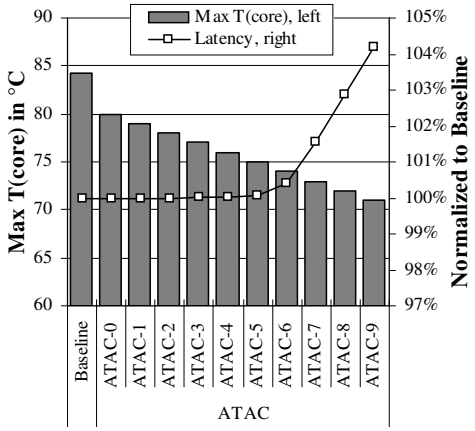


Figure 10: T_{core} and Latency for ATAC when $T_{trigger} = 40°C$

We start with the most basic strategy, ATAC-0, which activates performance capping when $T_{inlet\ air} = T_{trigger}$. For example, we assume that $T_{trigger} = T_{emergency} = 40°C$ and that one of the servers in the datacenter senses that $T_{inlet\ air} = 45°C$. In this case, without ATAC support, T_{core} can be as high as $85°C (= T_{inlet\ air} + \Delta T = 45°C + 40°C)$ according to the ΔT specification discussed in Section 4. However, with ATAC support, after acknowledging that $T_{inlet\ air}$ exceeds $T_{trigger}$ by $5°C$, ATAC reduces the maximum power consumption of the CPU to $\frac{\Delta T - 5°C}{\Delta T}$, thus reducing the required temperature difference between T_{core} and $T_{inlet\ air}$ to $35°C$; then the maximum T_{core} becomes $80°C$.

Figure 10 shows the results of the scenario described above. In Figure 10, the highest value of T_{core} for the baseline configuration is as high as $84°C$ while the worst-case T_{core} of ATAC-0 is $80°C$. We also define a more aggressive ATAC from ATAC-1 to ATAC-9. ATAC-1 activates performance capping activated by ATAC when $T_{inlet\ air} = T_{trigger} - 1$, and ATAC-9 activates it when $T_{inlet\ air} = T_{trigger} - 9$. As a result, the maximum T_{core} declines to $79°C$ for ATAC-1 and $71°C$ for ATAC-9. Even though ATAC-0 lowers the maximum T_{core} about $4°C$ from the baseline, the likelihood of activating performance capping is extremely low. Therefore, the average latency of the jobs submitted to the datacenter shows less than a 1% increase until ATAC-6, raising an interesting question: Which ATAC is the best design choice?

Before answering the above question, we analyze the relationship between ATAC and $T_{trigger}$. We assume that a datacenter employs ATAC-α with $T_{trigger} = \beta$. In this case, ATAC-α triggers performance capping when $T_{inlet\ air} = T_{trigger} - \alpha = \beta - \alpha$. If another datacenter, however, uses ATAC-(α +1) with $T_{trigger} = \beta + 1$, this configuration triggers performance capping when $T_{inlet\ air} = \beta + 1 - (\alpha + 1) = \beta - \alpha$, the same temperature as that of the former datacenter. If we recall that T_{core} can be as high as $T_{inlet\ air} + \Delta T$, these two data centers share the same max T_{core} requirement. As a more aggressive ATAC indicates a stronger performance penalty, the datacenter will perform better in the former (*i.e.*, lower) ATAC setting than in the latter. By fixing the max T_{core} constant, we compare different ATAC and $T_{trigger}$ configurations in Figure 11.

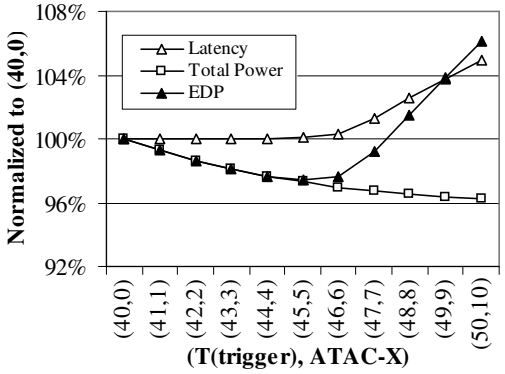


Figure 11: All configurations share the same $max(T_{core})$

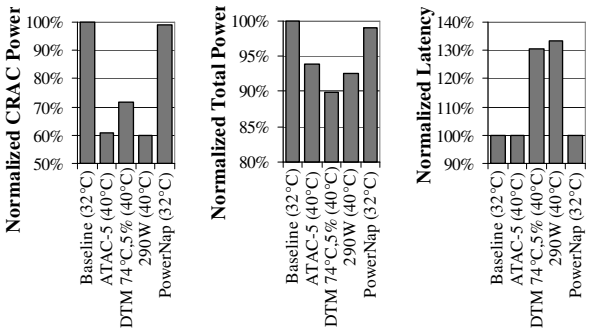
Figure 11 presents a comparison of average latencies, total power consumption, and the energy-delay product (EDP) of various $T_{trigger}$ and ATAC configurations that result in the

same max T_{core}. All of the values are normalized to those of ($T_{trigger}$, ATAC-X) = $(40,0)$. To save more power but compromise performance, we use higher $T_{trigger}$ and aggressive ATAC values (from left to right). The average latency of the datacenter significantly increases from ($T_{trigger}$, ATAC-X) = $(45,5)$ while the power savings begin to saturate from ($T_{trigger}$, ATAC-X) = $(48,8)$, mainly resulting from increased fan power. Therefore, in terms of EDP, ($T_{trigger}$, ATAC-X) = $(45,5)$ is the best design choice. Later sections will discuss configurations of only up to ATAC-5.

5.3 Comparing ATAC against DTM, Power Capping, and PowerNap

ATAC is unique in that it accounts for ambient temperature. Because ATAC activates performance capping from the servers at the highest inlet air temperature, it exploits temperature differences between servers and outperforms the other power management schemes. Figure 12 shows the maximum T_{core} value and the normalized latency of the simulated datacenter for different power management algorithms including DTM [9], power capping [20] and PowerNap [24]. Note that power capping and PowerNap are not designed for thermal management but for other purposes; however, we still compare these schemes against ATAC only for reference. We evaluate four different configurations for DTM: $< 79°C, 10\% >$, $< 79°C, 5\% >$, $< 78°C, 10\% >$, and $< 78°C, 5\% >$. For all configurations, ($X°$, $Y\%$) denotes that DTM degrades performance in a step of $Y\%$ whenever T_{core} exceeds $X°C$. Power capping is a power management technique for datacenters that activates performance capping by sensing system-level power consumption and strictly limits the maximum power consumption under the bar. In our experiment, when power capping is available, server power is capped to 310W, 300W, or 290W. We also implement the ideal PowerNap. Although the original PowerNap has a 300 micro-second performance penalty for waking up from the napping state, we assume a zero penalty to show the upper bound of the effectiveness of the algorithm. In addition, we use the same configurations for the baseline and ATAC-0 $\sim$ ATAC-5, as in the previous section.

First, Figure 12a shows that ATAC, DTM, and power capping are effective at reducing the maximum value of T_{core}. For example, when DTM is set to $< 78°C, 5\% >$, the highest T_{core} is $79°C$, which is close to the T_{core} of ATAC-1. However, as shown in Figure 12b, at this configuration, DTM compromises about 13% of the latency of the datacenter. When power capping set to 290W, the highest T_{core} is $75°C$, which is close to T_{core} of ATAC-5; however, such aggressive power capping results in more than 30% performance degradation while all ATAC configurations shows less than 1%. The reason why ATAC underperforms DTM is as follows. At high ambient temperature (HTA) datacenters, the temperature of T_{core} is more likely to fall between the DTM threshold and emergency core temperatures as in Figure 6. In such a case, DTM always degrades the performance of servers



(a) CRAC Power (b) Total Power (c) Normalized Latency

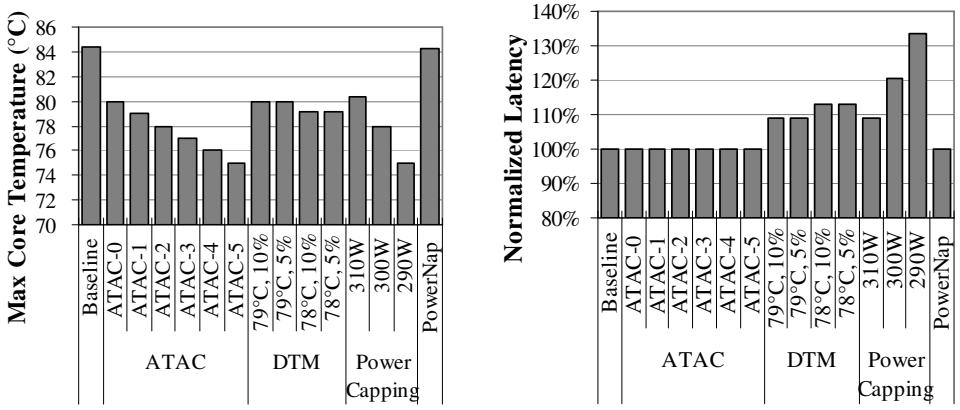
Figure 13: Max(T_{core})-Equivalent Comparison

while ATAC degrades only when $T_{inlet\ air}$ exceeds $40°C$. The same rationale holds for power capping, which lowers the performance of the CPU only by detecting the system-level power consumption. In ATAC, even when the server burns at full power (*i.e.*, peak utilization) no performance capping is triggered as long as $T_{inlet\ air}$ is substantially low. By contrast, PowerNap has little impact on T_{core} since we only show the maximum value of T_{core}. $Max(T_{core})$ occurs when servers are occupied, and PowerNap, which is designed for energy proportional datacenters [4], cannot help occupied servers.

5.4 Max(T_{core})-Equivalent Comparison

As discussed previously, ATAC, DTM, and power capping algorithms effectively lower the upper bound of T_{core}. For example, ATAC-5, which only activates performance capping when $T_{inlet\ air}$ is higher than $T_{trigger} - 5$, lowers the maximum T_{core} value from $84.3°C$ to $75°C$ when it is compared to the baseline, in which $T_{trigger} = 40°C$. Results from additional simulations show that the baseline datacenter without any power management mechanism lowers $T_{trigger}$ from $40°C$ to $32°C$ to achieve the same level of $Max(T_{core})$. Similarly, since PowerNap has no impact on T_{core}, PowerNap also has to lower $T_{trigger}$ to $32°C$ to achieve the maximum T_{core} of $75°C$. However, when DTM is set to $< 74°C, 5\% >$ or power capping is set to 290W, the maximum value of T_{core} is the same as ATAC-5 with $T_{trigger} = 40°C$. Thus, ATAC-5, DTM set to $< 74°C, 5\% >$, and power capping is set to 290W, both achieve the maximum T_{core} of $75.0 \pm 0.1°C$ while the baseline and PowerNap have to lower $T_{trigger}$ to $32°C$.

Figure 13 presents a comparison of the power consumption of all four configurations. The labels on the X-axis show the name of the configurations and corresponding $T_{trigger}$ values in parentheses. Note that all configurations have the same peak T_{core} values of $75.0 \pm 0.1°C$. In terms of cooling power, savings for ATAC-5, DTM, power capping, and PowerNap are 39%, 28%, 40%, and 1%, respectively. These savings are translated to about 6%, 10%, 7%, and 1% savings in terms of the total datacenter power, including all the components such as computing, fan, and cooling power. DTM set to



(a) Max Processor Die's Temperature

(b) Normalized Latency

Figure 12: Comparing ATAC against Other Power Management Algorithms when $T_{trigger} = 40°C$

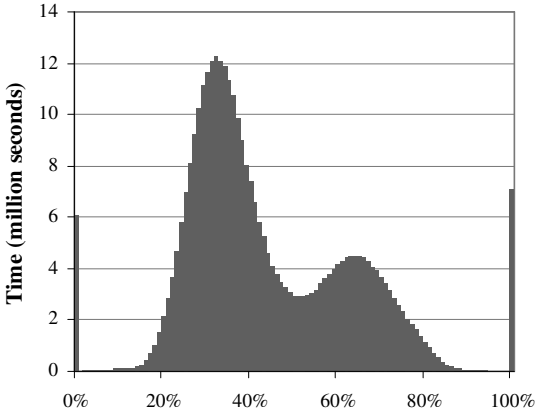


Figure 14: Per Server Utilization Distribution

$< 74°C, 5\% >$ is the most effective power-saving technique; however, it significantly degrades performance.

Figure 13c shows that the datacenter with DTM set to $< 74°C, 5\% >$ exhibits a latency penalty of more than 30%. In contrast, the impact of ATAC-5 on performance is negligible, less than 1%. Even though our implementation assumes the ideal PowerNap, Figure 13c shows that PowerNap has limited impact on the overall power consumption of the datacenter. Figure 14 provides an explanation for such an observation. The figure shows the distribution of the server-level utilization of the baseline configuration ($T_{trigger} = 40°C$ without any power management scheme) in seconds. As shown, servers typically maintain a utilization level of 20% to 80%. Servers in GCD are completely idle only 1.8% of the time. Because PowerNap places servers in a napping state when they are completely idle, PowerNap shows headroom less than 1.8% for this specific utilization trace. However, we also find that PowerNap can be used in conjunction with ATAC to save an additional 1% of the total power consumption.

5.5 Discussions

5.5.1 Parallel Workloads

ATAC is a system-level technique that allows individual servers to operate with their own decisions; each server locally senses $T_{inlet\ air}$ and decides whether or not to cap its performance. In this case, a parallel job, which often utilizes a number of servers and waits for the last server to finish, may experience higher performance overhead than what we have discussed in Section 5.3. Such inter-server dependencies were common in traditional high performance computing (HPC) workloads. However, in modern data centers, HPC jobs are in fact considered to be latency insensitive. In modern data centers, the most latency-sensitive jobs are online, data-intensive (OLDI) workloads, such as social networking and web searching [23]. They typically utilize thousands of servers to distribute key-value pairs. However, since each request to individual server is independent from others, each request can be processed and returned by the server without involvements of other servers. As such, these OLDI workloads do not suffer from server interferences. In summary, server interferences may further degrade the performance of ATAC for HPC workloads; however, these workloads are latency insensitive by nature. For latency-senstivie OLDI workloads, ATAC degrades performance by less than 1% on average.

5.5.2 Server interferences

Because ATAC is a per-system technique without centralized management, decisions made from each server can be suboptimal. For example, only a few servers may need to slow down to reduce the heat being generated and recirculated. However, in our case, ATAC penalizes performance of all servers at hot spots. Although our evaluations show that the impact of suboptimal decisions on the overall performance of datacenters is negligible *on average*, it is also true that latencies of certain jobs can be significantly slow compared to the baseline datacenter without ATAC. In other words, these suboptimal decisions or server interferences may signifi-

cantly compromise latencies at the 99th or 95th percentiles. To prevent servers from triggering unnecessary performance capping, servers now need to communicate with each other so that selective numbers of servers with fewer jobs in their queue can slow themselves down. We leave harmonic effects of ATAC and centralized management as our future work.

5.5.3 Simulation Errors

Using a heat distribution matrix proposed by a previous study [40], SimWare models the heat recirculation effect. However, the method used in the previous study showed simulation errors of $0.38^{\circ}C$ on average when estimating the inlet air temperature of servers. In this section, we show the impact of these simulation errors on the findings of this study. To do so, we extend Equation (5) by adding T_{error} as follows.

$$T_{inlet\ air} = T_{supply\ air} + T_{recirculated\ heat} + T_{error} \qquad (8)$$

However, since the original work did not reveal the details of how errors are distributed, we present and evaluate four different error models, all of which have an average error of $0.38^{\circ}C$. *Plus* and *Minus* models assume that T_{error} is always $+0.38^{\circ}C$ or $-0.38^{\circ}C$, respectively, and *Uniform* and *Normal* models assume that T_{error} is a random variable picked from a uniform distribution $U(-0.76, 0.76)$ or from a normal distribution $N(0, 0.4763^2)$. We select plus and minus models to show the lower and upper bound of the impact of errors on our study and other two models to show the behavior of ATAC under random error. Particularly for the random error models, we evaluate 13 times and show the average of all of the runs.

We compare the power and performance numbers of ATAC(45,5) with and without T_{error}, and find that all of the metrics fall within 0.05% of the range of the baseline without errors (T_{error}=0). Such narrow range is extraordinary if we consider that the errors in temperature, $0.38^{\circ}C$, is already about 0.8%; however, this is because of the following reasons. The errors in temperature impact on CRAC power; but CRAC power for ATAC(45,5) is already less than 10% of the total power. Therefore, 0.8% change in temperature is translated to 0.1% change in the total power. Moreover, because ATAC(45,5) is already optimal in balancing fan power and CRAC power, any effort in decreasing CRAC power results in more fan power. Therefore, possible CRAC power savings have been compensated by increased fan power to result in sub 0.1% changes. In all, we observe here that errors in temperature are not exacerbated in our results.

6. Related Work

Researchers have investigated increasing the supply air temperature without compromising reliability. Moore *et al.* [25] proposed a new job scheduling policy to minimize the heat recirculation effect, and Banerjee *et al.* [3] further improved it. A prior study found that when $T_{inlet\ air}$ increases, the processor cores contribute to the majority of additional power consumption [7]. Atwood *et al.* [2], however, showed that the failure rates of servers have little correlations to temperature, dust, and humidity. These studies motivated us to design system-level support that exploits the cooling inequality among the servers in datacenters.

In this work, we primarily focus on the power consumption of cooling units and servers; nonetheless, other sources of inefficiency were explored in prior research. For example, Wang *et al.* [42] and Pelley *et al.* [29] proposed efficient power delivery and smarter cluster-level power controller, and Li *et al.* [22] proposed power-efficient execution of programs. In addition, Haque *et al.* [18] proposed a new definition of service-level agreements, Green SLAs, for the clients who care about using green energy. Alleviating the peak power consumption is an important issue for datacenters [15] because their electricity bills are based on (1) the amount of energy they use and (2) the peak power that they demand. Use of fresh-air cooling [10] or renewable energy [13, 14, 21] also improves cooling efficiency of datacenters. Although ATAC achieves the same goal (i.e., improving the cooling efficiency), it can be used in parallel with aforementioned techniques. For example, with ATAC support, a datacenter with free-cooling systems can exploit high temperature variations among server locations.

Similar to ATAC, Zephyr [41] discussed blade chassis-level power optimizations including fan and server power, while our study focuses on datacenter-level power optimizations including cooling power. In addition, the novelty of ATAC lies in exploiting location-dependent and regional cooling characteristics inside datacenters.

Advancements of micro-architectures and memory technologies can lead to significant energy savings in datacenters. For example, Razor [11] allows microprocessors to operate at a lower voltage by comparing results from multiple flip-flops operating at different speeds. Razor is in fact conceptually similar to ATAC: Razor lowers a supply voltage and exploits voltage safety margins of microprocessors, while ATAC lowers cooling power and exploits temperature safety margins of datacenters. Emerging memory technologies, such as die-stacked memory [8], would also play a key role in alleviating power concerns in datacenters. Stacked DRAM caches already become practical to be deployed in large-scale servers by alleviating hardware overhead [38] and resiliency concerns [39]. These advancements could greatly reduce computing and memory power in datacenters.

7. Conclusion

Motivated by the knowledge that small- to medium-scale datacenters, which comprise the majority of datacenters, consume nearly half of their power for cooling, we initiated this study to determine an efficient method of cooling.

We began by carefully reviewing the fundamentals of datacenter cooling and found that considerable cooling energy is wasted because of (1) the safety margin that cooling units must ensure and (2) the non-uniform inlet air temperatures across servers. These issues stem from the location of each server relative to the CRAC unit and their height from the floor. To address this drawback, we proposed a system-level approach that first aggressively reduces the cool air supply from the CRAC unit to save power and then uses a new system-level control called ATAC, which is applied to each server. By sensing the inlet temperature to reduce the core temperature, ATAC can dynamically cap the performance of the server using DVFS. Using a modified SimWare framework with the Google production trace, we evaluated ATAC and found that a datacenter can reduce the cool air supply with 38% savings of cooling power, or 7% savings of total power while degrading performance by a negligible sub-1%.

Acknowledgments

This research is supported in part by an NSF grant CNS-0644096. The authors would also like to thank Thu D. Nguyen of the Rutgers University for the constructive comments and technical discussions.

References

[1] F. Ahmad and T. N. Vijaykumar. Joint optimization of idle and cooling power in data centers while maintaining response time. In *Proceedings of the Fifteenth Edition of ASPLOS on Architectural Support for Programming Languages and Operating Systems*, ASPLOS-15, 2010.

[2] D. Atwood and J. Miner. Reducing data center cost with an air economizer. *Intel White Paper, Tech. Rep*, 2008.

[3] A. Banerjee, T. Mukherjee, G. Varsamopoulos, and S. K. S. Gupta. Cooling-aware and thermal-aware workload placement for green hpc data centers. *International Green Computing Conference*, 2010.

[4] L. A. Barroso and U. Holzle. The case for energy-proportional computing. *Computer*, 40(12):33–37, 2007.

[5] C. Bash, C. D. Patel, and R. K. Sharma. Dynamic thermal management of air cooled data centers. In *Proceedings of the Tenth Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronics Systems*, ITHERM, 2006.

[6] T. Benson, A. Akella, A. Shaikh, and S. Sahu. Cloudnaas: a cloud networking platform for enterprise applications. In *Proceedings of the 2nd ACM Symposium on Cloud Computing*, SoCC, 2011.

[7] S. Biswas, M. Tiwari, T. Sherwood, L. Theogarajan, and F. T. Chong. Fighting fire with fire: modeling the datacenter-scale effects of targeted superlattice thermal management. In *Proceeding of the 38th annual international symposium on Computer architecture*, ISCA-38, 2011.

[8] B. Black, M. Annavaram, N. Brekelbaum, J. DeVale, L. Jiang, G. H. Loh, D. McCauley, P. Morrow, D. W. Nelson, D. Pantuso, et al. Die stacking (3d) microarchitecture. In *Proceedings of the 39th Annual International Symposium on Microarchitecture*, MICRO-39, 2006.

[9] D. Brooks and M. Martonosi. Dynamic thermal management for high-performance microprocessors. In *Proceedings of the Seventh Annual Symposium on High Performance Computer Architecture*, HPCA-7, 2001.

[10] H. Endo, H. Kodama, H. Fukuda, T. Sugimoto, T. Horie, and M. Kondo. Effect of climatic conditions on energy consumption in direct fresh-air container data centers. In *International Green Computing Conference*, IGCC, 2013.

[11] D. Ernst, N. Kim, S. Das, S. Pant, R. Rao, T. Pham, C. Ziesler, D. Blaauw, T. Austin, K. Flautner, et al. Razor: A low-power pipeline based on circuit-level timing speculation. In *Proceedings of IEEE/ACM 36th International Symposium on Microarchitecture*, MICRO-36, 2003.

[12] M. Ferdman, A. Adileh, O. Kocberber, S. Volos, M. Alisafaee, D. Jevdjic, C. Kaynak, A. D. Popescu, A. Ailamaki, and B. Falsafi. Clearing the clouds: a study of emerging scale-out workloads on modern hardware. In *Proceedings of the seventeenth international conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS-17, 2012.

[13] Í. Goiri, W. Katsak, K. Le, T. D. Nguyen, and R. Bianchini. Parasol and greenswitch: Managing datacenters powered by renewable energy. In *Proceedings of the 18th Symposium on Architectural Support for Programming Languages and Operating Systems*, ASPLOS-18, 2013.

[14] Í. Goiri, W. Katsak, K. Le, T. D. Nguyen, and R. Bianchini. Designing and managing datacenters powered by renewable energy. *IEEE Micro*, 2014.

[15] S. Govindan, A. Sivasubramaniam, and B. Urgaonkar. Benefits and limitations of tapping into stored energy for datacenters. In *Proceeding of the 38th annual international symposium on Computer architecture*, ISCA-38, 2011.

[16] J. Hamilton. 2011 European Data Center Summit. http://perspectives.mvdirona.com/2011/05/25/2011EuropeanDataCenterSummit.aspx.

[17] J. Hamilton. Where does the power go in high-scale data centers (keynote address). *Proceedings of the International Conference on Measurement and Modeling of Computer Systems (Sigmetrics)*, 2009.

[18] M. E. Haque, K. Le, Í. Goiri, R. Bianchini, and T. D. Nguyen. Providing green slas in high performance computing clouds. In *International Green Computing Conference*, IGCC, 2013.

[19] Lawrence Berkeley National Laboratory. Data Center Energy Benchmarking Case Study — Facility 8, 2003.

[20] C. Lefurgy, X. Wang, and M. Ware. Power capping: a prelude to power shifting. *Cluster Computing*, 11(2):183–195, 2008.

[21] C. Li, A. Qouneh, and T. Li. iswitch: coordinating and optimizing renewable energy powered server clusters. In *Proceedings of the 39th Annual International Symposium on Computer Architecture*, ISCA-39, 2012.

[22] D. Li, B. R. D. Supinski, M. Schulz, K. W. Cameron, and D. S. Nikolopoulos. Hybrid mpi/openmp power-aware computing. In *Proceedings of the 24th IEEE International Parallel and Distributed Processing Symposium*, IPDPS, 2010.

[23] D. Lo, L. Cheng, R. Govindaraju, L. A. Barroso, and C. Kozyrakis. Towards energy proportionality for large-scale latency-critical workloads. In *Proceeding of the 41st Annual International Symposium on Computer Architecuture*, ISCA-41, 2014.

[24] D. Meisner, B. T. Gold, and T. F. Wenisch. PowerNap: Eliminating Server Idle Power. In *Proceeding of the Fourteenth International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS-14, 2009.

[25] J. D. Moore, J. S. Chase, P. Ranganathan, and R. K. Sharma. Making scheduling cool: Temperature-aware resource assignment in data centers. In *2005 Usenix Annual Technical Conference*, pages 61–75, 2005.

[26] T. Mukherjee, A. Banerjee, G. Varsamopoulos, S. K. Gupta, and S. Rungta. Spatio-temporal thermal-aware job scheduling to minimize energy consumption in virtualized heterogeneous data centers. *Computer Networks*, 53(17):2888–2904, 2009.

[27] C. D. Patel, R. Sharma, C. E. Bash, and A. Beitelmal. Thermal considerations in cooling large scale high compute density data centers. In *Proceedings of the Eighth Intersociety Conference on Thermal*

and Thermomechanical Phenomena in Electronic Systems, ITHERM, 2002.

[28] M. K. Patterson. The effect of data center temperature on energy efficiency. In *11th Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*, ITHERM, 2008.

[29] S. Pelley, D. Meisner, P. Zandevakili, T. F. Wenisch, and J. Underwood. Power Routing: Dynamic Power Provisioning in the Data Center. In *Proceeding of the Fifteenth International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS-15, 2010.

[30] F. J. Pollack. New microarchitecture challenges in the coming generations of cmos process technologies (keynote address). In *Proceedings of the 32nd annual international symposium on Microarchitecture*, MICRO-32, 1999.

[31] K. P. Puttaswamy, C. Kruegel, and B. Y. Zhao. Silverline: toward data confidentiality in storage-intensive cloud applications. In *Proceedings of the 2nd ACM Symposium on Cloud Computing*, SoCC, 2011.

[32] C. Reiss, A. Tumanov, G. R. Ganger, R. H. Katz, and M. A. Kozuch. Towards understanding heterogeneous clouds at scale: Google trace analysis. Technical report, 2012.

[33] C. Reiss, A. Tumanov, G. R. Ganger, R. H. Katz, and M. A. Kozuch. Heterogeneity and dynamicity of clouds at scale: Google trace analysis. In *Proceedings of the Third ACM Symposium on Cloud Computing*, SoCC, 2012.

[34] A. Shah, C. Patel, C. Bash, R. Sharma, and R. Shih. Impact of rack-level compaction on the data center cooling ensemble. In *11th Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*, ITHERM, 2008.

[35] G. Shamshoian, M. Blazek, P. Naughton, R. S. Seese, E. Mills, and W. Tschudi. High-tech means high-efficiency: The business case for energy management in high-tech industries. 2005.

[36] R. K. Sharma, C. E. Bash, C. D. Patel, R. J. Friedrich, and J. S. Chase. Balance of power: Dynamic thermal management for internet data centers. *IEEE Internet Computing*, pages 42–49, 2005.

[37] Z. Shen, S. Subbiah, X. Gu, and J. Wilkes. Cloudscale: elastic resource scaling for multi-tenant cloud systems. In *Proceedings of the 2nd ACM Symposium on Cloud Computing*, SoCC, 2011.

[38] J. Sim, G. H. Loh, H. Kim, M. O'Connor, and M. Thottethodi. A mostly-clean dram cache for effective hit speculation and self-balancing dispatch. In *Proceedings of the 2012 45th Annual International Symposium on Microarchitecture*, MICRO-45, 2012.

[39] J. Sim, G. H. Loh, V. Sridharan, and M. O'Connor. Resilient die-stacked dram caches. In *Proceedings of the 40th Annual International Symposium on Computer Architecture*, ISCA-40, 2013.

[40] Q. Tang, T. Mukherjee, S. K. Gupta, and P. Cayton. Sensor-based fast thermal evaluation model for energy efficient high-performance datacenters. In *Proceedings of the Fourth International Conference on Intelligent Sensing and Information Processing*, ICISIP, 2006.

[41] N. Tolia, Z. Wang, P. Ranganathan, C. Bash, M. Marwah, and X. Zhu. Unified thermal and power management in server enclosures. In *the ASME/Pacific Rim Technical Conference and Exhibition on Packaging and Integration of Electronic and Photonic Systems, MEMS, and NEMS*, InterPACK, 2009.

[42] X. Wang and M. Chen. Cluster-level feedback power control for performance optimization. In *Proceedings of the 14th International Symposium on High Performance Computer Architecture*, HPCA-14, 2008.

[43] D. H. Woo and H.-H. S. Lee. Extending amdahl's law for energy-efficient computing in the many-core era. *Computer*, 41(12):24–31, 2008.

[44] S. Yeo and H.-H. S. Lee. Simware: A holistic warehouse-scale computer simulator. *Computer*, 45(9):48–55, 2012.