TAPO: Thermal-Aware Power Optimization Techniques for Servers and Data Centers

Wei Huang, Malcolm Allen-Ware, John B. Carter,
Elmootazbellah Elnozahy, Hendrik Hamann*, Tom Keller,
Charles Lefurgy, Jian Li, Karthick Rajamani, and Juan Rubio
IBM Research – Austin, *IBM Research – T. J. Watson

ABSTRACT

A large portion of the power consumption of data centers can be attributed to cooling. In dynamic thermal management mechanisms for data centers and servers, thermal setpoints are typically chosen statically and conservatively, which leaves significant room for improvement in the form of improved energy efficiency. In this paper, we propose two hierarchical thermal-aware power optimization techniques that are complementary to each other and achieve (i) lower overall system power with no performance penalty or (ii) higher performance within the same power budget.

At the data center level, we trade off facility Heating, Ventilation and Air Conditioning (HVAC) power with server fan power by choosing between two thermal setpoints for the HVAC chiller based on the cooling zone utilization levels. This optimization can reduce total data center total power by as much as 12.4%-17%, with no performance penalty.

At the server level, we trade off fan power and circuit leakage power by dynamically adjusting the server thermal setpoint, allowing the system to heat up when this saves more fan power than it costs in terms of leakage power. We evaluate this optimization on an IBM POWER 750 and find that it reduces total server power by up to 5.4% with no performance penalty for workloads that heavily exercise a server.

1. INTRODUCTION

Power consumption has become a primary concern for enterprise servers and data centers [1]. The cooling subsystem is a large portion of the data center's power consumption and frequently rivals the power spent powering IT equipment [2]. For individual servers, especially high-density server blades with high-performance processors, fan power can account for up to 23% of typical server power [3] and scales super-linearly with server utilization. Therefore, it is critical to reduce cooling power for both data centers and servers using novel thermal-aware power management techniques.

There is substantial prior art on mechanisms to reduce the energy required to cool servers and data centers (e.g., [3, 4, 5, 6, 7, 8]). However, prior work has not examined in depth the relationship between thermal setpoints in the data center and server utilization, and the relationship between thermal setpoints in the data center and the behavior of server fan arrays. We address these issues and find that substantial power savings can be achieved by co-managing power and thermals in both servers and data centers.

1.1 Thermal Setpoint

Thermal management systems for servers and data centers employ a control system that monitors and adjusts certain system parameters, e.g., a Computer Room Air Handler's (CRAH) inlet air temperature, a chiller's supply water temperature, or a server's internal temperature, to maintain a target thermal setpoint. These thermal setpoints typically are chosen empirically from spreadsheet models or conventional wisdom. This approach eases thermal control design, but static thermal setpoints do not consider the dynamic nature of the system being cooled, which leads to over-designed and power-inefficient cooling solutions.

For example, it is a common practice to set the data center Heating, Ventilation and Air Conditioning (HVAC)'s chiller thermal setpoint (e.g., chilled water temperature) to a constant cold temperature around 10°C, regardless of the utilization level of the data center cooling zone. While this very cold set point is appropriate when the cooling zone is heavily utilized, i.e., when the IT equipment in the data center is generating a substantial amount of heat, it is an overkill when the data center is lightly utilized. In practice, most data centers spend a significant amount of time at low utilization levels (5% to 20%), especially when virtualization and workload consolidation are not being employed [9, 10]. As a result, data center HVAC systems typically operate at a power-inefficient setpoint.

Another example of inefficient cooling occurs within individual servers running workloads that heavily exercise the server. In these conditions, the server's fans are operated close to their maximum cooling capacity. The amount of energy consumed by server-level fans increases superlinearly with fan speed, which corresponds to their cooling capacity. Relaxing the operational thermal setpoint of processors inside a server by 2-3°C, i.e., allowing the processors to run a few degrees hotter, can save a significant amount of fan power without reliability concerns, as there is usually a large margin between the operational thermal setpoint and the critical temperature (e.g., 70°C vs. 85°C) for the processor.

1.2 Cooling-Related Power Tradeoffs

Important trade-offs can be made between cooling-related components.

For a data center cooling zone, the power expended by chiller units is the largest component of cooling power. In a recent attempt to reduce cooling costs, data center operators commonly run data centers warmer. This makes the desired inlet temperature closer to the exhaust heat temperature and improves chiller efficiency. As a result, the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) suggests a 90°F (32°C) operating environment [11]. Microsoft allows its chiller-less data center to operate up to 95°F (35°C) [12]. Rackable Systems even allows up to 105°F (40°C) [13].

However, increasing server inlet temperature increases the power consumed by IT equipment fans, which are forced to push higher volumes of (warmer) room air across their components to maintain server-level thermal setpoints. To move these higher air volumes requires running the fans at higher RPMs, which consumes substantial energy. In fact, server-level fan energy is frequently the second highest source of energy consumption in server and storage

978-1-4577-1221-0/11/ $26.00

systems and can consume 23% of total server power [3]. Although recent efforts on more intelligent data center cooling control (e.g. [4, 14]) improve data center cooling power efficiency, the tradeoff between cooling infrastructure power and IT fan power is still an orthogonal approach worth investigating.

There is another opportunity for trading off cooling energy and thermal setpoints within an individual server. It is well known that processor leakage power is strongly dependent on temperature – the higher the processor temperature, the more energy lost due to circuit leakage. Expending more power in the fans leads to lower chip temperatures and hence lower leakage power, and vice versa.

In general, to minimize the overall system power, cooling subsystems must operate at different thermal setpoints under different system conditions. In the case of data centers, the optimal HVAC chiller setpoint minimizes aggregate HVAC and server fan power. In the case of individual servers, the optimal processor thermal setpoint minimizes aggregate fan and leakage power. Another benefit of optimizing system power by tuning thermal setpoints is that it incurs no performance degradation, unlike many other server/DC energy optimizations, because performance is largely independent of thermal setpoints.

1.3 Scope and Contributions

In this paper, we first demonstrate the power saving potentials of adjusting thermal setpoints and explore the tradeoff among cooling-related power components. We then propose simple yet effective thermal-aware power management techniques to search for optimal thermal setpoints during runtime. These power management techniques are orthogonal to existing power management techniques.

1. At the data center cooling zone level, we developed Thermal-Aware Power Optimization for data centers (TAPO-dc), which switches between two distinct HVAC chiller setpoints (high and low) for a cooling zone based its utilization level. TAPO-dc optimizes aggregated HVAC and server fan power, and can achieve up to a 12.4%-17% reduction in total data center power with no performance penalty.

2. At the server level, we developed Thermal-Aware Power Optimization for servers (TAPO-server), which uses runtime measured power to adjust server thermal setpoint and optimize aggregated server fan and leakage power. We built a working prototype on an IBM POWER 750 server and demonstrate 5.4% total server power reduction for a workload that heavily exercise the server processor, with no performance penalty.

The two novel thermal-power management techniques (TAPO-dc, TAPO-server) do not rely on each other, and hence can be implemented independently or in combination. The power savings from TAPO-dc and TAPO-server lower the operational cost of a data center without hurting performance. Alternatively, the power savings can be used to deploy new servers for more revenue.

2. HARDWARE INFRASTRUCTURE

There are numerous flavors and choices of server platforms, which make it difficult to present detailed models and perform detailed analysis without specifying a particular platform. In this paper, we use the state-of-the-art IBM POWER 750 server [15] as the underlying hardware infrastructure. The system has four POWER7 processors [16] and 64GBytes of system DRAM arranged from 16×4GB DIMM devices. We are also able to measure the performance of each processor core in the form of instructions per second (IPS). The power for the entire system is also measured as well as the power consumption and speed of the cooling fans.

Each POWER7 processor contains eight cores, each supporting up to four-way simultaneous multi-threading running at a nominal

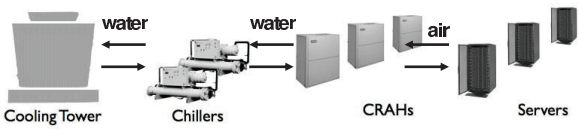

Figure 1: HVAC components and Servers.

frequency. In Turbo mode, cores run up to an 8.7% higher clock frequency with all eight cores active. None of the cores need to be shut off on the POWER7 to achieve this full Turbo performance.

Inside each core are five digital thermal sensors averaged to establish the core average temperature. At the system level, all 32 cores' average temperatures are compared to find the hottest core [17]. The temperature of the hottest core is fed into a dynamic fan control algorithm. The server system has four identical system cooling fans. All fans are controlled synchronously and have identical fan speeds. The dynamic fan control algorithm also receives the following as input: hottest DIMM temperature, hottest memory buffering chip temperature, and the hottest I/O chip temperature. Each unique component type has its own thermal control setpoint, which by default are all fixed. The closed-loop fan control algorithm around the thermal setpoints causes server fan power to track server component power.

Furthermore, the data center cooling zones idealized in this paper are assumed to be homogenously built with racks of POWER 750 servers, and energy proportionality is also of primary consideration in the data center design. This assumption becomes more realistic with the advent of data centers built for cloud computing services with more homogeneous cooling zones for each type of services. For the data center HVAC subsystem, we present the power models based upon data from an experimental data center [8]. Although the data center control and analysis in this paper are not actually implemented and measured, all assumptions are made within typical ranges for contemporary systems across different vendors. Therefore, we believe the insights and conclusions from this work are also applicable to a broad selection of platforms and data centers.

3. TAPO-DC

To reduce data center level cooling power, we consider the tradeoff between the power from the facility HVAC that is used to cool server inlet airflow in a data center cooling zone and the power from cooling fans inside individual servers. The basic logic is that more HVAC power leads to cooler server inlet ambient air temperature, which in turn make the server cooling fans work less hard and thus save server fan power. Similarly, lower HVAC power leads to warmer ambient air temperature, which causes server fans to work harder and consume more IT power.

An HVAC consists of three major parts—the Computer Room Air Handler (CRAH), the refrigeration chiller system, and a cooling tower [7]. In a CRAH, chilled water is pumped through a set of coils, which exchanges heat with incoming air forced into the CRAH by the blowers, thus cooling the air on its way to the raised floor and servers inlet. The warm water is circulated back to a chiller system (usually in a separate room) that exchanges the heat by phase change from a compressor and produces cold chilled water. The chiller expels its heat dissipation through a cooling tower to the outside environment. The chilled water thermal setpoint impacts the power consumption of the entire HVAC—a lower thermal setpoint requires more chiller power. Figure 1 illustrates the interactions among components of an HVAC. Chilled water refers to the water flow from chiller to the CRAH. The amount of required HVAC power is also obviously determined by the total dissipated power of the data center.

Total data center power can be written as

$$\begin{aligned} P_{DC\_total} &= P_{IT} + P_{cooling} + P_{others} \\ &= P_{IT} + (P_{chiller} + P_{tower} + P_{pump} + P_{blower}) \\ &\quad + (P_{lighting} + P_{PDU}) \end{aligned} \quad (1)$$

Since we only consider tradeoff between HVAC power and the IT fan power, we ignore the lighting and Power Distribution Unit (PDU) power, which aggregately account for about 6% of data center power [18].

3.1 Chiller Power

The calculation of power in a chiller system ($P_{chiller}$) can be very complicated. For the purpose of this study, we have established that the average power dissipation for a typical chilled water system can be well described by two parts, based on literature search and example measurements [7, 8, 19]:

$$\begin{aligned} P_{chiller} &= \frac{1}{a(1+b\cdot(T_{s\_chiller} - T_0))} P_{DC\_total} + 0.05 P_{DC\_total} \\ &= \frac{P_{DC\_total}}{COP_{chiller}} \end{aligned} \quad (2)$$

with the first term as the power dissipation associated with the chiller and the second as the power associated with the pump and the cooling tower. Notice we also include CRAH pump power and cooling tower power into Eq. (2) to further simplify the analysis. $T_{s\_chiller}$ is the chilled water thermal setpoint, T_0 is the reference temperature and is set to 48°F (9°C) in this study. a and b are chiller dependent coefficients.

In this paper, we combine the two terms in Eq. (2) into one and define the coefficient of performance (COP) of the chiller system as the ratio of P_{DC_total} to $P_{chiller}$. One important observation from Eq. (2) is that *increasing chiller thermal setpoint ($T_{s\_chiller}$) leads to higher chiller COP and less chiller power.* Different chiller designs have different values for a and b in Eq. (2) and hence different COP ranges across the temperature range of interest. In this study we look at two chillers with different COP characteristics—one with a wider COP range of 3.0 to 6.0, the other with a narrower COP range of 4.1 to 5.5, each across a reasonable range of $T_{s\_chiller}$ from 10 to 30°C. From the COP ranges, we know that at high $T_{s\_chiller}$, the first chiller is more efficient, whereas at low $T_{s\_chiller}$ the second chiller is more efficient. High COP is desirable because a high performance chiller can remove more heat with less chiller power consumption. Additionally, there is a constant temperature difference between chilled water and server inlet temperature as a result of the control of return air temperature to CRAH unit and the blower inside the CRAH. We choose a typical value of 10°C for this.

3.2 CRAH Blower Power

To transport air throughout the CRAH to the raised floor and servers, a rule of thumb based on blowers in an operating data center is that 10k CFM (cubic feet per minute) of air can transport 100kW of IT power. For a particular blower, we also know from the specification its maximum CFM ($maxCFM$) and its associated blower power (P_{blower_max}). With this, we can derive the blower power as follows

$$P_{blower} = P_{blower\_max} \cdot \left(\frac{P_{IT} + P_{chiller}}{100kW} \cdot \frac{10kCFM}{maxCFM} \right)^{\alpha} \quad (3)$$

The term inside the parentheses is the ratio of the required air flow to the blower's maximum air flow, which is equivalent to relative blower speed. α is specific to different blowers, and is usually greater than 2.0 [8]. Due to this superlinear factor, a better practice in data center cooling design is to use larger blowers to transport more IT power, instead of running small blowers at maximum speed. Larger blowers running at much lower speeds consume much less power and have better power efficiency. This makes the term inside the parentheses much less than 1.0, and hence Eq. (3) becomes approximately linear to $P_{IT} + P_{chiller}$. This linear relationship between the blower power and the sum of IT power and chiller power is what we assume in this work.

3.3 Server Fan Power

P_{IT} includes both power consumption for active computing (processors, memories, disks, etc) and the IT fan power (i.e. server cooling fan power). It is the IT fan power that we try to optimize with HVAC power. For an IBM POWER 750 server, the measured server cooling fan power has the characteristics shown in Figure 2.

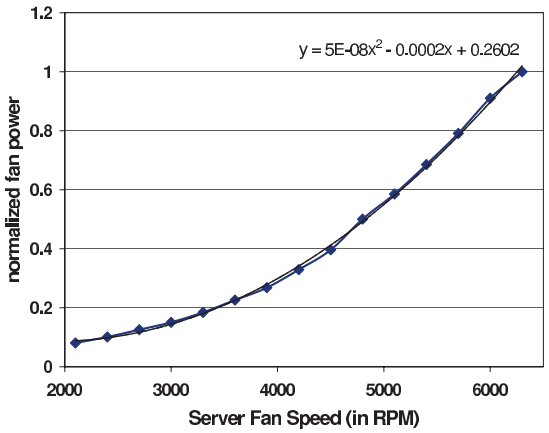

Figure 2: Server cooling fan power is super-linear to its RPM.

Specific fan power for this particular server model is normalized to its maximum power. We assume the processor inside the server has a constant silicon junction temperature of 70°C, and calculate the necessary fan RPM for a server inlet temperature range of 20 to 40°C. This corresponds to an HVAC chiller thermal setpoint range of 10 to 30°C as a result of the 10°C temperature difference between chilled water and server inlet temperatures. After calculating RPM, we calculate the fan power from Figure 2. As we can see, the server fan power has a strong superlinear relationship to RPM (e.g., a 10% fan power increase from 2000 to 3000 RPM, as opposed to a 40% fan power increase from 5000 to 6000 RPM). Therefore, to decrease the server fan power, it is highly desirable to operate the fan at a lower speed.

3.4 Server Power without Fans

We also need to know the IT power (excluding server fan power). In this study, we assume the IT power varies linearly between idle and peak power [20].

$$P_{IT\_without\_fan} = P_{idle} + \mu \cdot (P_{full} - P_{idle}) \quad (4)$$

where μ is the power utilization factor of a data center cooling zone, varying from 0 to 1. We use measured values from our IBM POWER 750 server for P_{full} and P_{idle}. Since we focus on power optimization, this power-based indicator of data center cooling zone utilization level (μ) makes our approach agnostic of detailed workload-specific characteristics.

3.5 Results and Analysis of TAPO-dc

Since all the above power components are to the first order linearly proportional to number of servers, the relative tradeoff results are scalable with the number of servers in a data center cooling zone, as long as it does not exceed chiller and blower capacities.

While different HVAC and server fan designs may lead to different results, it is important to notice the opportunity of power savings by exploiting the tradeoff relationship between HVAC power and server fan power, especially for modern high-performance and energy-proportional servers.

Figs. 3-5 shows the modeled aggregate power of IT, server fans, chiller and blower as server inlet air temperature varies, at three different data center cooling zone utilization levels (80% utilized, 40% utilized and 10% utilized). Two chillers with different efficiency ranges are listed (narrow: COP 4.1-5.5; wide: COP 3.0-6.0). Absolute power is normalized to the maximum power of the two chiller choices across all the temperatures for each utilization level. The lines in each figure show the overall power trend for different thermal setpoint with a zoom-in scale (0.8 to 1.0) on the left Y-axis. The bars show each power component's contribution to the total power on a scale from 0 to 1, on the right Y-axis. The two chiller choices are compared side-by-side in the bar charts for each thermal setpoint.

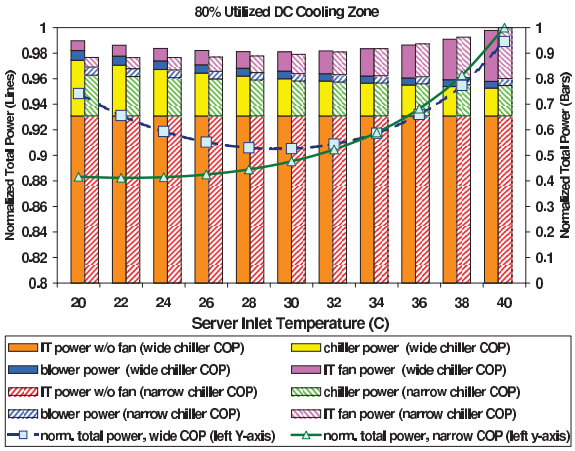

Figure 3: Normalized power for a data center cooling zone with a 80% power utilization level. Lines (left Y-axis): total power trend for two chiller choices. Bars (right Y-axis): the breakup portion of each major power components.

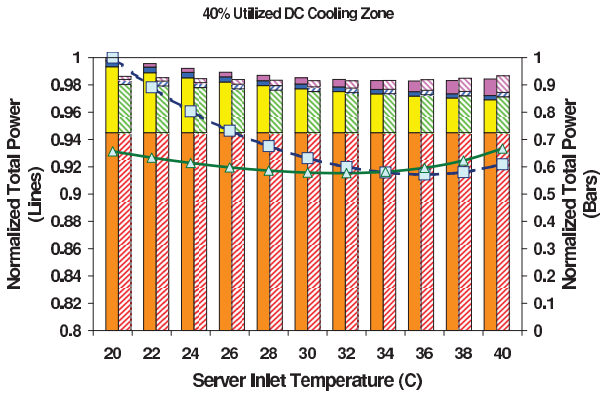

Figure 4: Normalized power for a data center cooling zone with a 40% power utilization level.

The first observation is that at high power utilization levels, total power of a data center with the narrow chiller COP range is more sensitive to a change in thermal setpoint (solid line in Figure 3). This is because at low thermal setpoints, this chiller's power efficiency is higher(4.1 vs 3.0, at 20°C). For example, for a 80%

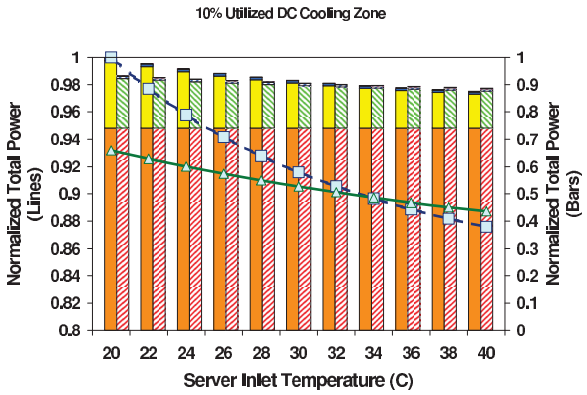

Figure 5: Normalized power for a data center cooling zone with a 10% power utilization level.

utilized cooling zone with a narrow chiller range, moving towards a 20°C thermal setpoint as opposed to 40°C can save 11.8% total power (for a fully utilized cooling zone, this can be as high as 17% total power saving). On the other hand, for low power utilization levels (idle to 30%), total power of a data center with a wide chiller COP range is more sensitive to the change in thermal setpoint (dashed line in Figure 5). This is because at higher thermal setpoint, this type of chiller is more power efficient (6 vs. 5.5, at 40°C). In fact, for a 10% utilized data center with a wide chiller COP range, moving to a 40°C setpoint saves up to 12.4% total power compared to a 20°C thermal setpoint. Even more power savings can be achieved with free air cooling without a chiller [12, 21]. For the middle power utilization levels range (40-60%), the optimal setpoint lies somewhere between the two extremes (Figure 4).

A more important observation from Figs. 3-5 is that there is no single optimal thermal setpoint for all utilization levels. The traditional approach of having a low 20°C thermal setpoint leads to up to 12.4% power being wasted for low-utilized data centers, due to excessive chiller power. Similarly, the new ASHRAE guideline towards a higher thermal setpoint of around 35-40°C causes 8-16% power waste for highly utilized DCs, due to the drastic increase in server fan power to keep a constant processor junction temperature. A constant thermal setpoint around 27-30°C seems to be a much better choice. However, it still does not fully exploit the total power saving potential at low DC cooling zone utilization levels. For example, it still consumes 4-6% more total power than a 35-40°C setpoint for idle data centers with a wide-COP-range chiller.

The key player here is the IT fan power that is very sensitive to the ambient air thermal setpoint as shown in Figure 2. In previous studies, power optimization around IT fan power was not fully considered in the whole picture of data center power management under different thermal setpoints. Instead, server fans are usually assumed to operate within a narrow speed range and consume relatively constant power. This is no longer true as servers are becoming more energy-proportional, requiring fan speed and hence fan power to change significantly according to server utilization level. The bar charts in Figs. 3-5 illustrate the server fans' contribution to overall data center power. It is obvious that server fans consume negligible power (less than 2%) when the cooling zone is idle or under-utilized. The power reduction in chiller power leads to overall power reduction at high chiller thermal setpoints. In this case, it is desirable to operate the data center in a warmer ambient, even with free cooling if humidity and altitude are not an issue. However, at high utilization levels, the server fan power percentage goes up faster than the reduction in chiller power, resulting in a higher percentage of overall cooling power (40% vs. 25%). In other words, the increase in server fan power outweighs the re-

duction in chiller and blower power, leading to an overall power increase at high chiller thermal setpoints. Therefore, it is desirable to use a low chiller thermal setpoint in this case. In such cases, metrics such as Power Usage Effectiveness (PUE) can be misleading, because although PUE is higher, the total power of the data center cooling zone is lower. The blower power is small and relatively constant for each utilization level.

3.6 Control Method for TAPO-dc

From the above analysis, we can naturally reach a control algorithm which monitors utilization level and power for each cooling zone inside a data center, and dynamically and continuously adjusts the thermal setpoint to reach minimum total power consumption for that particular utilization level. However, in reality, such an approach is extremely difficult to implement as the HVAC is such a complicated system that it is hard to set a continuous thermal setpoint and expect an accurate, timely response. In addition, it may take hours to days for such a control algorithm to fully converge to the optimal thermal setpoint given the huge thermal mass and complexity of a data center cooling zone, if it can converge at all, before utilization levels change.

Here, we propose a simple binary dynamic control method that chooses from two thermal setpoints according to utilization. From the results shown in this section, it is proper to select 27°C when the utilization level is greater than 50%, and select 40°C when the utilization is less than 50% (or 35°C if the server vendor does not guarantee proper operation at a higher ambient temperature). Notice that 27°C is also close to what ASHRAE suggests for Class 1 mission-critical data centers' operating environment. The monitor and actuation of the control method can be performed at a rather long time interval (e.g. every few hours), taking into account the large thermal time constant of a cooling zone. This binary control is easy to implement and achieves almost the same power savings as the aforementioned continuous control method. The control flowchart is shown in Figure 6.

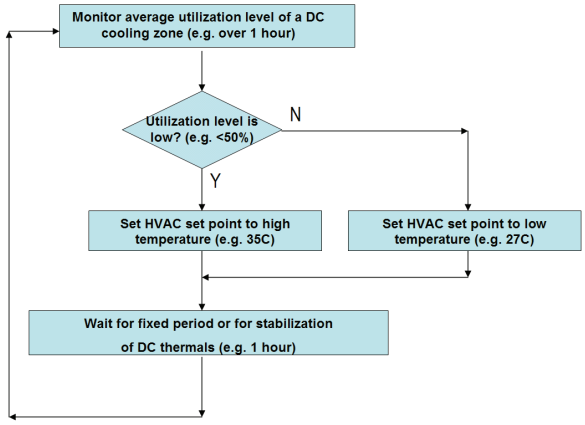

Figure 6: Simple binary dynamic control method for TAPO-dc.

Figure 7 shows the efficiency of the binary control method compared to the ideal control method with continuous adjustment of the chiller thermal setpoint. As we can see, the binary control achieves almost the same amount of power saving as the ideal control does, especially for the low-utilization levels (5-20%) that are typical for data centers. The biggest difference occurs at the 40% data center utilization level for the chiller with narrow COP range.

In the case of heterogeneous servers or heterogeneous utilization levels inside a cooling zone, we can calculate a weighted average for the utilization level of the cooling zone. Further analysis shows that the binary control method still works well in such a heterogeneous environment, although the amount of power savings is less significant (not shown due to page limit). This is because the fan power increase in the "hot" servers outweighs the fan power reduction in the "cool" servers. Nevertheless, the "hot" servers are guaranteed to operate safely at 35°C inlet temperature as per most vendors. In a rare case of local hot spots inside a cooling zone caused by highly imbalance workload distribution among racks and/or worse air recirculation and/or the "hot" server is far away from the CRAH, if the server fan cannot cool server to a safe operating temperature, the chiller setpoint decided by TAPO-dc can be override to a lower value.

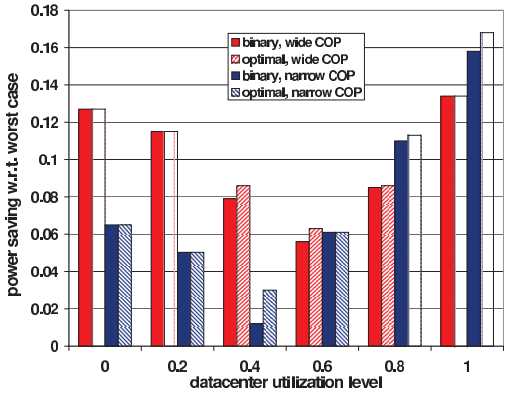

Figure 7: Comparison between a binary control method (27°C or 40°C) and an ideal optimal control method that is not practical to implement.

4. TAPO-SERVER

In this section, we demonstrate a run-time optimization technique that reduces the aggregate server fan power and processor leakage power of a server system that operates at Turbo frequencies without compromising performance. In Section 3, for the server fan power calculation, we assume the thermal setpoints for the server components are static, as is the practice today. For example, among all processor cores, the hottest core's junction temperature is maintained at 70°C. This processor thermal setpoint (T_{s_proc}) is usually determined empirically and conservatively, leaving enough margin to prevent excess leakage, reliability degradation and timing errors at a higher critical temperature (T_{crit}, e.g., 85°C or higher). At run time, it is the server fan's job to bring them back to T_{s_proc}. Thermal setpoints for other server components are set similarly.

Given the large margin between T_{s_proc} and T_{crit}, and the fact that server fan power is a strong superlinear function to T_{s_proc} at high fan speeds, there is a potential power saving at the server level by increasing T_{s_proc}. The rationale is that higher T_{s_proc} relaxes the cooling needs from the server fans, allowing fans to operate at lower speeds, resulting in superlinear reduction in fan power. On the other hand, increasing T_{s_proc} makes processors hotter and causes more leakage power. Therefore, it is desirable to find an optimal value for T_{s_proc} that minimizes the aggregate server fan power and processor leakage power.

One factor that complicates the optimization of aggregate server fan power and leakage power is that leakage power varies widely between processors that have identical architectures due to process variability. Any attempt to model processor leakage power and use the leakage model to optimize power would fail to reach the optimal thermal setpoint due to process variations. A statistical leakage model is useful to evaluate leakage distribution across a large batch of processor chips, but would not help at all for each individual chip. Alternatively, characterizing leakage power for each chip after fabrication at different temperatures is expensive. Therefore, a run-time optimization method using real-time measurement

is highly desirable.

By dynamically and moderately adjusting the processor thermal setpoint, this technique is able to reach an operational state during run-time that balances the thermally induced leakage power of all the processors and the power expended by the fans cooling those processors. Here, we only consider the case where processors hit their thermal setpoint first as most leakage power is from processors. Optimizing for other server components is left as future work. Furthermore, the machine runs in Turbo mode operation to demonstrate maximum system performance by running at the highest specified frequency.

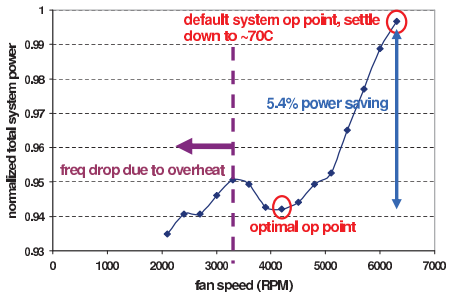

Figure 8: Measured total system total power as a function of manually swept fan speed, showing tradeoff between leakage power and fan power.

4.1 Manual Characterization

To get an idea of how much power we could potentially save, we manually swept the server fan speed from 2000RPM to 6300RPM. The fan power curve is shown in Figure 2. We ran *DAXPY*, a CPU intensive floating-point microbenchmark with L2-resident memory footprint (to minimize memory access and to fully heat up the processors), and observe from Figure 8 that there is clearly an optimal point around 4000-4500RPM where system power is minimized. In comparison, the default fixed thermal setpoint of 70°C results in a fan speed of >6000RPM, for this workload running at the maximum frequency allowed by the processors.

As we can see, by adjusting fan speed to its optimal value, we reduce total system power by 5.4%. In Figure 8, we can also see that if fan speed is further reduced from its optimal point, the increase in leakage power starts to dominate. The processors start to experience thermal emergency (e.g., >85°C) below 3200RPM. At this time, aggressive measures such as Dynamic Voltage and Frequency Scaling (DVFS) kick in to cool down the processors with a noticeable performance penalty. Because we keep processor at constant Turbo frequency (except for the points where DVFS is engaged due to thermal emergency), minimizing power is equivalent to maximizing power efficiency.

4.2 Measurement-Based Control Method for TAPO-server

Figure 9(a) illustrates the adaptive control mechanism that is used to find the optimal fan speed. Figure 9(b) is a simplified illustration of the relationship between fan RPM and system power (the center portion of Figure 8). By measuring the change in system power and the change in hottest processor temperature, we know whether the system is operating on the left- or right-hand side of the V-shaped curve, as well as in which direction it is moving. Based on this information, we make decisions about whether to increase or decrease the thermal setpoint.

For example, if we observe a decrease in power and an increase in temperature (caused by a decrease in fan speed) from Time 1 to Time 2, as marked in the right portion of Figure 9(b), we know that the system is operating on the right-hand side of the curve and is moving down towards the optimal point. Therefore, the corresponding decision is to further move down along the curve (indicated by the dashed arrow) by increasing the processor thermal setpoint. This causes fan speed to decrease because the system has a relaxed thermal requirement. Consequently, fan power also decreases and processor temperature starts to increase, which in turn leads to a moderate increase in leakage and possibly lower total system power. On the other hand, if we observe an increase in both power and temperature from Time 1 to Time 2, the system is operating on the left-hand side of the curve and is moving upward away from the optimal point. In this case, we step back by decreasing the thermal setpoint. With a harsher thermal setpoint fan speed goes up and leakage power is therefore reduced. Similar decision processes can be derived for the remaining two cases: system power and temperature decrease; system power increases and temperature decreases.

By repeating in a looped fashion, as shown in Figure 9(a), the system adjusts its thermal threshold toward an optimal fan speed minimizing the total system power. The search for optimal thermal threshold is continued till system power changes are within a pre-defined delta value. When the power changes exceed the delta again, singaling a possible in workload characteristics, the search is resumed again.

We take power measurements during an 8-second interval to get an average power value commensurate to the thermal time constants. This results in a smooth fan RPM vs. system power curve without noticeable local minima/maxima. Furthermore, we specify a minimum fan RPM change step that is large enough to get the control loop out of potential local minima. Our implementation shown by the flowchart in Figure 10 provides a graphical description, where T_{thr} is the thermal threshold, i.e. thermal setpoint. ΔT_{thr} is the step of thermal setpoint change, which is 1°C in this study. For each change in thermal setpoint and consequently each change in fan speed, a pre-defined time needs to be spent in waiting for processor temperatures to settle down before moving forward to get updated measurements.

Because this technique uses temperature measurements to make dynamic decisions, and system-level thermal response is usually on the order of tens of seconds to minutes, one limitation of this technique is that it best suits relatively constant workloads that run for a relatively long time (e.g., a few minutes or longer). However, it is worth noting that if the workload changes before convergence, the technique will adjust itself towards an new fan speed for the new workload. Although it has not converged for the previous workload, it still results in total system power reduction for the duration of the previous workload, albeit less than the optimal saving.

4.3 Results and Analysis of TAPO-server

We have implemented this technique on a prototype IBM POWER 750 server system. The system can be configured to operate at Turbo mode with the maximum allowed frequency for workloads that heavily exercise the processors. We choose *DAXPY* as a representative workload for this study. *DAXPY* adds a scalar multiple of a double precision vector to another double precision vector. To maximize the load on the processor cores, we limit the workload to be L2 cache resident such that the workload spends its entire time being executed on the microprocessor.

System-level dynamic thermal management (DTM) is also enabled for this prototype system. Based on the maximum temperature measurement of all the processors, system firmware issues a series of changes for cooling fan speed to the point where the hottest system component's temperature settles at the designated thermal setpoint, no matter whether it is fixed as in the default 70°C or is dynamically changed by our new technique.

We set the convergence criterion to be 5W change in system

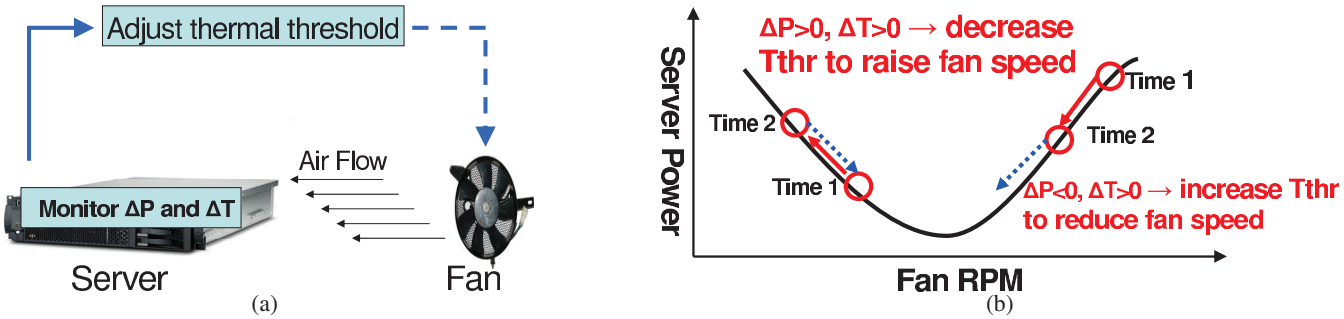

Figure 9: (a) Overview of the control loop for TAPO-server. (b) Tradeoff between server fan power and leakage power.

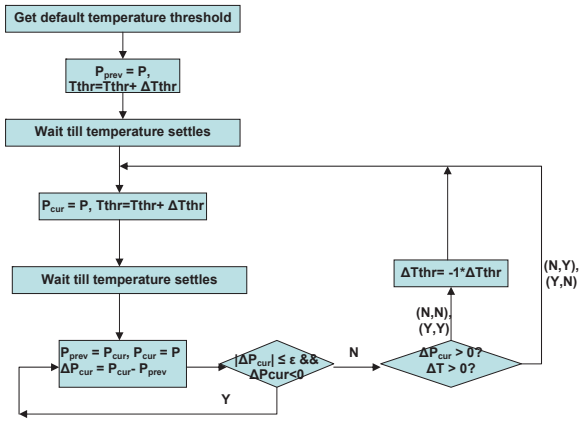

Figure 10: Measurement-based control flowchart of TAPO-server.

power, which is a reasonably small amount of power at peak performance, but also not too small to prevent the technique from reaching convergence.

Figure 11(a) shows the trace of normalized cooling fan power over time. After the system is already at the steady state running *DAXPY* at maximum frequency, we engage the technique. We can see the dynamic adjustment of the fan power during the convergence process until it finally settles down at around half the initial fan power.

As a means to engage the technique, we force an initial disturbance to the system by boosting the thermal setpoint by 1°C, which, for this workload, results in an immediate reduction in fan power and system power during the first few seconds, which can be seen in Figure 11(a) and (c) (curves seem not to start from 1.0 at time zero).

Figure 11(b) shows the corresponding change in the power of the four processors. We only show the processor power since we set *DAXPY* to be L2 resident and there are virtually no memory or disk accesses. Due to the decrease in fan speed and the increased leakage power, the final processor power is slightly higher by 2%.

Figure 11(c) shows the normalized total system power. We can see that with the engagement of the new technique, we are able to save 5% total system power for this CPU-intensive configuration. In the manual characterization in Figure 8, we showed that maximum potential power saving is about 5.4%, which is close to what we can actually achieve.

Figure 11(d) shows the dynamic adjustment of the processor thermal setpoint. Remember this thermal setpoint has been fixed and it is intended to control the leakage power. As we can see, there is some overshoot of the thermal setpoint up to 75°C, and finally settles to 73°C, which is only 3°C higher than the default 70°C. This indicates that for this CPU-intensive workload, the fan was operating at the steepest part of the fan power curve. The overshoot is caused by the delay of fan speed change in response to the thermal setpoint adjustment and the limited steps of fan speed change. This also explains the overshoot in Figure 11(a)-(c). The complexity of the system makes it difficult to completely eliminate the overshoot. We are working on an improved fan control algorithm to reduce the overshoot. Nevertheless, this adaptive technique is robust enough to quickly damp the overshoot and settle the fan speed to an operating point that is close to the optimal point.

We also apply TAPO-server to the SPECpower workload [22]. SPECpower is designed to provide a common standard to evaluate the power and performance characteristics of volume server class and multi-node class computers. It has 11 different load levels. TAPO-server makes a difference primarily at high load as at low loads fan speeds are low enough with insignificant fan power consumption. Especially when we run the 100% SPECpower workload level at a slightly elevated ambient temperature of 26°C, rather than the 23°C in the earlier experiments, we see a 5.2% total power saving with thermal setpoint converged to 78°C. This result implies that for data centers that attempt to save overall cooling energy by raising the ambient temperature, it is important to also take server fan power into account as discussed in Section 3. Additionally, in a warmer data center environment, TAPO-server also has the potential for server power saving for interactive workloads such as web services, where processor utilization is relatively low.

5. RELATED WORK

At the data center level, Mukherjee et al. [23] consider optimizing power of a data center infrastructure (such as HVAC) with thermal-aware scheduling. Bash et al. [4] include server inlet temperature into a distributed sensor network and use a PID controller to decide Computer Room Air Handler (CRAH) supply air temperature and flow rate, without explicitly considering tradeoff between HVAC and server fans. Chen et al. [5] propose the concept to integrate management of performance, power and cooling in data centers, based on virtualization. It has CRAH output temperature adjustments to keep servers operating at a constant ambient temperature. Pakbaznia et al. [6] also consider adjusting chiller thermal setpoint to save data center cooling power. It is based on models and does not explicitly optimize total power with consideration of server fan power. Das et al. [8] propose to use utility functions not only for self-management of power and performance objectives, but also temperature objectives as well, which results in total energy savings. ASHRAE's guideline also mentions the superlinear increase in server fan power when operating data centers at warmer ambient temperatures, but does not fully explore the cooling power tradeoff relationship. TAPO-dc is different from existing studies

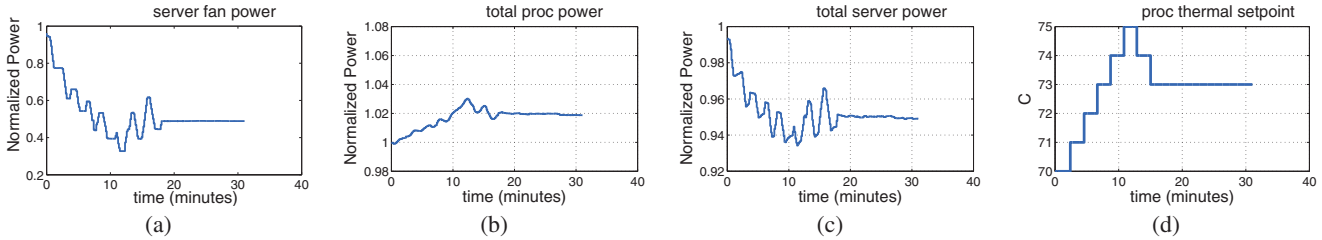

Figure 11: Normalized changes in (a) fan power, (b) total processor power, (c) total system power and (d) thermal setpoint.

in that it explicitly considers the tradeoff between HVAC power and *server fan power* by adjusting the chiller thermal setpoint. It is solely based on utilization measurements, and uses a simple binary control method to take into account the extremely complicated cooling mechanism and the large thermal time constant of a data center cooling zone.

At the server level, there are several existing studies on system power reduction by monitoring and controlling cooling. For example, in Wang et al. [3], a MIMO fan controller based on thermal models is used to achieve tighter server temperature control and reduced fan power. Economou et al. [24] propose a server power model for exploration of cooling and active power co-optimization. Shin et al. [25] build models for fan power and leakage power to minimize aggregate power by finding optimal fan speed based on the models. The thermal setpoint of the processor was not directly considered. All the above previous work relies on estimating or modeling (leakage) power, which becomes difficult with the presence of variations and workload dependency.

6. CONCLUSION

Cooling power has become a significant portion of both server and data center power consumption. In this paper, we explore ways to reduce energy consumption by carefully managing the cooling subsystems of servers and data centers.

At the data center level, we propose TAPO-dc, a mechanism that manages data center thermal setpoints based on IT equipment utilization to reduce cooling power by up to 12.4%-17%. TAPO-dc saves power by trading off chiller power and server fan power. In high-performance and energy-proportional servers, server fan power has a much wider power range and hence plays a more important role in total data center power.

At the server level, we propose TAPO-server, a mechanism that takes advantage of the relationship between server fan power and leakage power. We find that by adjusting a server's thermal setpoint and measuring the resulting impact on fan and leakage power, we can reduce the power of an IBM POWER 750 by roughly 5.4% for compute-intensive workloads.

Overall, by reclaiming and exploiting conservative cooling power margins, TAPO-dc and TAPO-server enable simple mechanisms to achieve significant power reduction.

7. REFERENCES

[1] EPA report to congress on server and data center energy efficiency, August 2007.

[2] L. Barroso and U. Holzle. *The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines*. Morgan and Claypool., 2009.

[3] Z. Wang, C. Bash, N. Tolia, M. Marwah, X. Zhu, and P. Ranganathan. Optimal fan speed control for thermal management of servers. In *Proc. of the ASME/Pacific Rim Technical Conf. and Exhibition on Packaging and Integration of Electronic and Photonic Systems, MEMS and NEMS (IPACK)*, July 2009.

[4] C. E. Bash, C. D. Patel, and R. K. Sharma. Dynamic thermal management of air cooled data centers. In *Proc. of the IEEE/ASME Tenth Intersociety Conf. on Thermal and Thermomechanical Phenomena in Electronic Systems (ITHERM)*, June 2006.

[5] Y. Chen, D. Gmach, C. Hyser, Z. Wang, C. Bash, C. Hoover, and S. Singhal. Integrated management of application performance, power and cooling in data centers. In *Proc. of Network Operations and Management Symposium (NOMS)*, April 2010.

[6] E. Pakbaznia, M. Ghasemazar, and M. Pedram. Temperature-aware dynamic resource provisioning in a power-optimized datacenter. In *Proc. of Design, Automation and Test in Europe Conf. and Exhibition (DATE)*, March 2010.

[7] H. F. Hamann, T. G. van Kessel, M. Iyengar, J.-Y. Chung, W. Hirt, M. A. Schappert, A. Claassen, J. M. Cook, W. Min, Y. Amemiya, V. Lopez, J. A. Lacey, and M. O'Boyle. Uncovering energy-efficiency opportunities in data centers. *IBM Journal of Research and Development*, 53(3), 2009.

[8] R. Das, J. O. Kephart, J. Lenchner, and H. Hamann. Utility-function-driven energy-efficient cooling in data centers. In *Proc. of Intl. Conf. on Autonomic Computing and Communications (ICAC)*, June 2010.

[9] Fed Data Centers 7 Percent Utilized, (http://www.datacenterknowledge.com/archives/2010/04/09/kundra-fed-data-centers-7-percent-utilized/),2009.

[10] Measuring Data Center Efficiency: Easier Said Than Done, (http://content.dell.com/us/en/enterprise/d/large-business/measure-data-center-efficiency.aspx),2009.

[11] 2008 ASHRAE Environmental Guidelines for Datacom Equipment — Expanding the Recommended Environmental Envelope., 2008.

[12] Microsoft's chiller-less datacenter, (http://www.datacenterknowledge.com/archives/2009/09/24/microsofts-chiller-less-data-center/),2009.

[13] Rackable Systems, (http://www.sgi.com/company_info/newsroom/press_releases/rs/2009/03182009.html),2009.

[14] Z. Wang, A. McReynolds, and C. Felix. Kratos: Automated management of cooling capacity in data centers with adaptive vent tiles. In *Proc. of ASME Intl. Mechanical Engineering Congress and Exposition (IMECE)*, November 2009.

[15] IBM POWER 750 Express server, (http://www-03.ibm.com/systems/power/hardware/750/index.html),2010.

[16] D. Wendel et al. The implementation of POWER7: A highly parallel and scalable multi-core high-end server processor. In *Proc. of Intl. Solid-State Circuit Conf. (ISSCC)*, February 2010.

[17] M. Ware, K. Rajamani, M. Floyd, B. Brock, J. C. Rubio, F. Rawson, and J. B. Carter. Architecting for power management: The IBM POWER7 approach. In *Proc. of Intl. Conf. on High-Performance Computer Architecture (HPCA)*, January 2010.

[18] H. F. Hamann, M. Schappert, M. Iyengar, T. van Kessel, and A. Claassen. Methods and techniques for measuring and improving data center best practices. In *Proc. of the IEEE/ASME Tenth Intersociety Conf. on Thermal and Thermomechanical Phenomena in Electronic Systems (ITHERM)*, June 2008.

[19] F. W. Yu and K. T. Chan. Low-energy design for air-cooled chiller plants in air-conditioned building. *Energy and Buildings*, 38(4), 2006.

[20] X. Fan, W.-D. Weber, and L. A. Barroso. Power provisioning for a warehouse-sized computer. In *Proc. of Intl. Symp. on Computer Architecture (ISCA)*, June 2007.

[21] Google's chiller-less datacenter, (http://www.datacenterknowledge.com/archives/2009/07/15/googles-chiller-less-data-center/),2009.

[22] SPECpower_ssj2008 Benchmark, (http://www.spec.org/power_ssj2008),2009.

[23] T. Mukherjee, Q. Tang, C. Ziesman, S. K. Gupta, and P. Cayton. Software architecture for dynamic thermal management in datacenters. In *Proc. of Intl. Conf. on Communications Systems Software and Middleware (COMSWARE)*, January 2007.

[24] D. Economou, S. Rivoire, C. Kozyrakis, and P. Ranganathan. Full-system power analysis and modeling for server environments. In *Proc. of Workshop on Modeling Benchmarking and Simulation (MOBS), held in conjuction with Intl. Symp. on Computer Architecture (ISCA)*, June 2006.

[25] D. Shin, J. Kim, N. Chang, J. Choi, S.-W. Chung, and E.-Y. Chung. Energy-optimal dynamic thermal management for green computing. In *Proc. of Intl. Conf. on Computer-Aided Design (ICCAD)*, November 2009.