Proceedings of the ASME/Pacific Rim Technical Conference and Exhibition on Packaging and Integration of Electronic and Photonic Systems, MEMS, and NEMS
InterPACK '09
July 19-23, 2009, San Francisco, USA

IPACK2009-89075

UNIFIED THERMAL AND POWER MANAGEMENT IN SERVER ENCLOSURES

Niraj Tolia, Zhikui Wang, Parthasarathy Ranganathan, Cullen Bash, Manish Marwah

Hewlett-Packard Laboratories
1501 Page Mill Road, MS 1183
Palo Alto, California 94304
Email: firstname.lastname@hp.com

Xiaoyun Zhu

VMware
3401 Hillview Avenue
Palo Alto, California 94304
Email: xzhu@vmware.com

ABSTRACT

Improving the cooling efficiency of servers has become an essential requirement in data centers today as the power used to cool the servers has become an increasingly large component of the total power usage. Unfortunately, most previous approaches have individually focused on reducing either the server power or the power used by the fans to cool the servers. This paper presents Zephyr, a systems approach to managing fan power that combines conventional server power optimizations with fan power management to optimize *overall* energy efficiency. By combining distributed system design with concepts from heat transfer theory, Zephyr can reduce cooling power by up to 21% when compared to a feedback-based controller and up to 30% when combined with cooling-aware workload migration policies. Overall, the combined Zephyr system can reduce total system power by up to 29% without impacting application performance.

1 Introduction

Power consumption is a critical issue in the design and operation of enterprise servers and data centers today. For 2006, the Environmental Protection Agency (EPA) reported that 60 billion kWh, or 1.5% of the total U.S.A. electricity consumption, was used to power data centers [1]. This is expected to rise to 100 billion kWh by 2012. In response to this problem, there have been several studies on server and cluster power management [2, 3, 4, 5, 6, 7, 8, 9, 10]. However, server power is only one component of the total power consumed by a data center. The other significant component is power consumed by cooling equipment (e.g., fans, computer room air conditioners). Several studies [11, 12] have shown that every 1W of power used to operate a server often requires an *additional* 0.5-1W of power, needed by the cooling equipment, to extract the heat at the data center level. For a large data center (e.g., 30,000 square feet, 10MW), the yearly electricity costs for cooling can reach millions of dollars [12]. The same trends are applicable at the individual server level. In particular, with increasingly dense compute infrastructures, such as blade servers, and more powerful processors, the server fans can often consume a significant amount of power. Peak power usage by fans can be as high as 2000W, comprising 23% of the typical system power. While a few studies have examined cooling power, they have mainly examined data center level issues [13, 14, 15, 16, 12] or have looked at server and cooling power as two separate aspects of the problem.

In contrast, this paper presents Zephyr, a model-based systems approach to managing fan power in servers that combines conventional server power optimization with fan power control to optimize *overall* energy efficiency. We believe this is the first work to study such a model-based approach to unified power and cooling management of servers. Zephyr uses a unique model-based fan controller to manage the distribution of cooling resources according to the needs of individual components in an enclosure [17]. Models are built that can determine the individ-

ual and collective impact of multiple actuators, including adjusting fan speeds, using Dynamic Voltage and Frequency Scaling (DVFS) or P-states [18, 3], and turning on/off servers [19], on a server's temperature. Zephyr also uses live Virtual Machine (VM) migration [20] to reduce the number of active servers and to place workloads in cooling-efficient regions of a blade enclosure. As such, Zephyr uses a multiplicity of cooling and workload management actuators to minimize overall enclosure power consumption while meeting application performance requirements.

The rest of this paper is organized as follows: Section 2 provides a background into the bladed servers used to evaluate Zephyr and introduces various means for workload and power management in servers that Zephyr exploits. Section 3 describes the overall design of Zephyr while section 4 discuss the three controllers used to manage resources. Finally, compared to previous studies on power and cooling, Section 5 quantifies Zephyr's benefits through a real prototype that works with commercial, off-the-shelf hardware. We measure both power savings and impact on application performance with workload traces gathered from real data centers. The results show that, without impacting performance, Zephyr's unified model-based approach can reduce cooling power by up to 21% when compared to a feedback-based controller and up to 30% when combined with cooling-aware workload migration policies. Overall, the combined Zephyr system can reduce total system power by up to 29% without impacting application performance.

2 Background

We begin this section with background on blade servers, a system that can benefit greatly from Zephyr, and also cover related work in the areas of virtualization and power and cooling.

2.1 Blade Servers

There has been a rapid growth in the use of blade servers, or simply blades, in data centers in recent years [21]. Commercially available blade systems include HP C-Class blades, IBM BladeCenter, and Dell PowerEdge blade servers. A survey of 166 data center operators [22] showed that 76% of operators were using blade servers within their data centers, with a further 14% having plans to deploy them in the near future. The faster growth of blade servers in the overall server market is being driven by their compact design, their high density in terms of compute and storage capacity, and their lower management costs.

Figure 1 illustrates an example of a typical HP blade enclosure. It has a total of sixteen blades in the front, eight on the top and eight on the bottom. The blades are cooled by a total of ten fans in the back, five on the top and five on the bottom. The airflow generated by the fans is pulled through the blades towards the back of the enclosure with each fan contributing to the blade-level airflow rate. While Zephyr is not limited to use in blade enclosures, the blade environment provides a challenge for model-based techniques due to the complex interaction of multiple variables and is therefore a suitable environment for Zephyr development and validation.

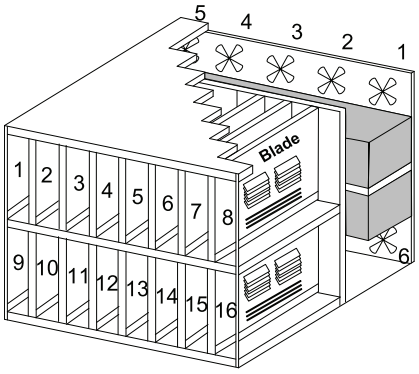


Figure 1. Enclosure Design

2.2 Virtualization in the Data Center

While virtualization is an old technology [23], its use has experienced a resurgence within data centers in the last few years, especially due to the availability of Virtual Machine Monitors (VMMs) from companies such as VMware, Citrix, and Microsoft, and hardware support found in Intel and AMD processors.

The biggest driver behind this transformation is consolidation. It is estimated that the average resource utilization in data centers ranges between 5–20% [24]. This allows operators to use virtualization to reduce the number of physical machines in their data centers. Apart from raising the per-server utilization, this consolidation also reduces power, floor space, cooling, and management costs. The use of VMs also provides other features such as security, performance, and fault isolation. Due to these reasons, IDC estimated that almost 75% of large organizations ($>$ 10,000 employees) were using server virtualization in their IT infrastructure in 2006 and predicted that 44% of servers being deployed by 2010 will be encapsulated within a VM [25].

Virtualization is also a key technology behind the growth of utility computing services such as virtual clusters [26, 27], Amazon's EC2 [24], and grid-based clusters [28]. VMMs have been shown to be extremely useful in these environments as live VM migration [20, 29] allows for autonomic management of resources [30, 31] in order to meet application performance goals [32] and even allows for high-availability [33] that was traditionally only available through customized hardware or software. In Zephyr, the same VM migration mechanism used by the above systems is leveraged to help optimize both power and cooling usage simultaneously in a blade enclosure.

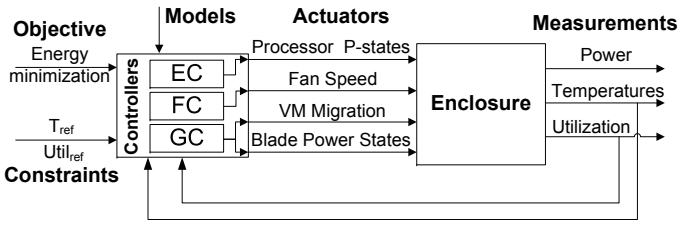


Figure 2. Zephyr Control System

2.3 Power and Cooling

There have been several studies on server and cluster power management. These studies have proposed solutions that enable the power consumed to track the resource demands of the application to reduce server electricity costs. The techniques used include low-power states (e.g., sleep and hibernate modes, DVFS [3, 5, 6, 34, 8]) and on/off states [2] at the local server level, and resource redirection [4, 7, 9] and task scheduling [26] at the cluster level. Of these studies, Chen et al. [34] also formally address the dynamic optimization problem of server provisioning and frequency control to reduce power while minimizing SLA violations. More recent work has examined approaches for power management for virtual machines. The VirtualPower [35] system explored exposing virtual P-states to VMs to guide actual changes in the underlying hardware while Stoess et al. [36] have proposed an energy management framework for VMs. Heo et al. [19] studied the potential conflict between a DVFS adaptation policy and a server on/off policy in a server farm when they are not coordinated. Raghavendra et al. [8] studied the interaction between multiple power management controllers at different levels of a data center.

Similarly, several studies have examined the cooling power, mainly at the data center level [13, 14, 15, 16, 12]. These include techniques to change workload placement to reduce air conditioning costs [16], as well as techniques to dynamically vary air flows to specific locations to improve cooling efficiency [12].

Zephyr is unique from these studies in its model-based techniques for fan power control, as well as its unified approach to power and cooling management for bladed environments.

3 Design and Implementation

Zephyr's goal is to minimize the total energy consumption by *both* the servers and the fans of a blade enclosure. Figure 2 illustrates the key components of the Zephyr control system, including the objective function and the constraints, the measurements and the actuators used, and the models needed.

Objective function and constraints: For an enclosure with **I** fans and **J** blades, Zephyr aims to minimize the total power consumed, i.e.,

$$\min\left(\sum_i P_{F_i} + \sum_j P_{B_j}\right), \tag{1}$$

where P_{F_i} is the power consumed by fan i ($i = 1,\ldots,I$), and P_{B_j} is the power consumed by blade j ($j = 1,\ldots,J$).

In addition, Zephyr ensures that the following two requirements are satisfied:

Thermal safety requirement: The temperature of each blade, T_j, should be maintained below T_{ref}, a reference threshold that depends on the tolerance level of the electronic components, i.e.,

$$T_j \leq T_{ref}, \quad \text{for any blade j.} \tag{2}$$

We currently use CPU temperature (T_{CPU_j}) as a proxy for T_j but could also include other sensors such as those that measure memory or motherboard temperatures.

Application performance requirement: Reducing power by aggressively tuning P-states or consolidating workloads may adversely affect application performance by creating resource bottlenecks on a server. To prevent this, the resource utilization level of each blade, $Util_j$, needs to be kept below a threshold, i.e.,

$$Util_j \leq Util_{ref}, \quad \text{for any blade j.} \tag{3}$$

In this paper, we use CPU utilization, measured as a percentage of total CPU capacity, as a proxy for server resource utilization and never over-commit memory. It should be straightforward to extend our system later to include other resources such as the network [32].

The above objective function (Equation 1) and constraints (Equations 2-3) together define an optimization problem. The time-varying and sometimes unpredictable nature of application demands requires that this optimization problem be solved at runtime. Zephyr therefore uses real-time measurements for temperature and resource utilization and adjusts its actuators dynamically to optimize the total power consumption.

The design of Zephyr includes the following components:

Measurements and Actuators: The current utilization and temperature information is available to the controllers from various software and hardware sensors located on a blade. Zephyr uses the following actuators to control power usage: processor P-states, fan speeds, workload placement through live VM migration, and blade on/off states. Manipulating these actuators will affect the power consumption of the blades and/or the fans as well as the temperature and utilization of the blades.

Controllers: Designing a single controller that can simultaneously utilize all the above actuators is challenging because dif-

ferent actuators work at different time scales and can be implemented across hardware and software. For example, while P-states can change every second, VM migrations may only happen every few minutes or even hours. In addition, fan control may be embedded in hardware whereas P-state control can be done in either hardware or software. Therefore, Zephyr uses a federation of three different controllers instead of a single controller. A fan controller (FC) adjusts the fan speeds periodically to minimize fan power while satisfying the thermal safety requirement in Equation 2. A group controller (GC) dynamically reassigns individual workloads to blades using live VM migration. Each blade also includes an efficiency controller (EC) that adjusts its processor P-state (and thus capacity available and power consumed) to match the resource demands of its workloads. The controllers and the challenges in their design and implementation are described in greater detail in Section 4.

Models: To solve the optimization problem defined in Equations 1-3, the controllers need models to determine the impact of actuator changes on the objective function and the constraints. For example, the FC requires models correlating fan speed and resource utilization with both fan power and blade temperature. The same set of models are needed by the GC, in addition to the models for predicting blade power for given workload placement. These models, described in greater detail elsewhere [17], summarize complex relationships between multiple variables and are often nonlinear.

4 Controllers

As mentioned earlier, Zephyr employs a federation of multiple controllers instead of a single unified controller to minimize the total power consumption of the blades and the fans in an enclosure. This is primarily because the actuators work at very different timescales ranging from milliseconds to minutes or even hours. Furthermore, system designers typically embed controllers like the FC in hardware. In this section, we first briefly describe some of the challenges in designing the controllers, and then present the design and implementation of Zephyr's three controllers.

4.1 Challenges

We faced the following three key technical challenges in the controller design. First, the control and optimization problems became more complex due to the presence of nonlinear relationships. For example, the cubic function that relates fan power to fan speed precluded the use of efficient optimization techniques such as linear programming. Furthermore, the nonlinearity also prevented us from using the large body of well established control techniques for linear systems.

Second, while some actuators have discrete values such as workload placement, others, like fan speeds, are continuous-valued. This created a heterogeneous search space for the optimization problem such that it cannot be solved using conventional optimization techniques such as Lagrange multipliers.

Finally, zonal variations and complex interdependencies exist between the fans and the blades. Building a controller for such a multiple-input multiple-output (MIMO) system implies that one cannot optimally control the temperature of an individual blade without considering the impact on the other blades surrounding it.

4.2 Fan Controller

The fan controller (FC) periodically adjusts the fan speeds in order to minimize the total fan power while satisfying the cooling requirements of all the blades. Since the fans are shared among the blades, the controller needs to consider all the fans and the blades simultaneously. By formulating this as a convex optimization problem, the fan controller computes optimal fan speeds for every control interval. A complete description of the controller can be found in our earlier paper [17].

4.3 Group Controller

The group controller (GC) operates at the highest level in Zephyr's control architecture and aims to minimize the power consumption of the entire enclosure. For a given set of workload demands, it addresses the problem of workload placement onto the blades such that the *total* blade and fan power consumed is minimized without violating any constraints. The GC plays a critical role since the workload placement it generates is used as a starting point for finer grain control and optimization by the fan and the efficiency controllers.

The GC takes four sets of inputs: (1) workload demand, (2) blade utilization, (3) current VM-to-blade assignment, and (4) blade ambient temperatures. As each workload is hosted within a VM, the demand is available as the utilization of individual VMs. The main actuator available to the GC is VM migration, that is, dynamic assignment of VMs to blades. Depending on its configuration, the GC can also, based on the demand, turn blades on or off. All idle blades that have no resident VMs can be powered off and they will be powered back on as the demand increases. While turning machines on and off has been previously cited as a reliability concern [34], this should not affect newer server-class machines or their internal storage systems during their normal lifetime.

Using workload placement, the GC solves the optimization problem defined in Equations 1-3 in the beginning of Section 3. However, this optimization problem is much harder than that for the fan controller. Since a VM must migrate atomically, the granularity of changes in utilization depends on the utilization levels of the VMs. The constraints are more complex as well due to the atomic migration operation. These differences make the problem intractable using conventional optimization techniques.

Simulated annealing was therefore chosen to solve the optimization problem in the GC. It is a randomized search algorithm for optimization in large search spaces. In particular, it is useful for searching discrete spaces and is less likely to get stuck at a local minimum [37]. While it is possible that other similar search techniques may be as effective, simulated annealing has worked well for us in practice. The simulated annealing algorithm requires generation and comparison of feasible candidate solutions to move towards a better solution. In order to narrow the search space and increase the probability of generating a better solution candidate, a heuristic function is used to choose assignments onto blades that have lower inlet temperatures. In order to compare two candidate assignments, both blade power and fan power need to be computed. While blade power is easily computed from the utilization values, computation of fan power requires determining the optimal fan speed settings for that assignment. This becomes an optimization problem in itself, similar to the one solved in the fan controller. However, since the GC operates over a coarser time scale, a simpler, steady-state thermal model is used. Again, this is a convex optimization problem that is solved using `cvxopt` [38].

Although the main objective is to minimize power, if multiple assignments exist with equal power cost, the one with the minimum number of VM migrations is chosen to reduce overhead. In addition, the GC algorithm places an upper bound on the number of VM migrations allowed in an assignment when generating candidate solutions.

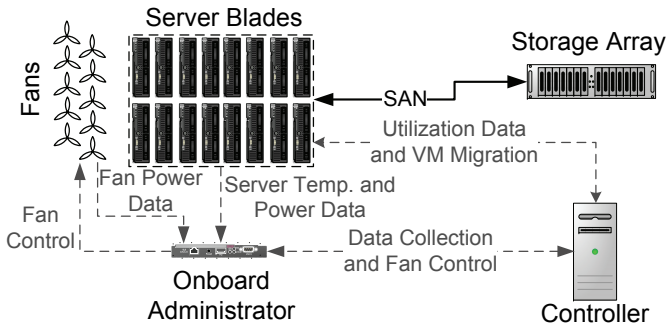
Ideally, the control interval of the GC should be shorter than the rate at which the workload demand changes. However, due to the overheads associated with VM migration and turning blades on and off, it will be inefficient to use a small time scale, say seconds. While our prototype uses a control interval of ten minutes for the GC, the techniques would work equally well with longer intervals. It should be noted that extending the GC to include scenarios where network or memory pressure is a concern [32] would be straightforward.

4.4 Efficiency Controller

The efficiency controller (EC) is very similar to the OnDemand governor [39] found in Linux. The EC is a purely local controller and adjusts the processor's voltage and frequency such that the utilization for the given P-state is never higher than 80%. In overloaded conditions, the EC fixes P-state to the highest frequency. While the EC can be run at sub-second granularity, its control interval was set to one second in our prototype.

4.5 Integral Fan Controller

To compare the performance of the Fan Controller, we designed the Integral Fan Controller (IFC). Unlike the predictive FC, the IFC is a reactive controller that increases or decreases the fan speeds based on the error between the temperature reference and current measurement. As the IFC has also been described earlier [17], we omit a detailed description here but would like to note that this controller is similar to commercial controllers used in industry today.



The dashed lines above represent sensor data and control flow.

Figure 3. Experimental Setup

5 Evaluation

We have evaluated Zephyr in a real data center. However, while this setup allowed us to gather realistic results, the fact that this facility also hosted services belonging to other groups prevented us from exploring different scenarios such as operation at elevated inlet temperature - a practice gaining attention due to its impact on reducing energy consumption at the data center level. Similarly, without access to larger number of servers, it was difficult to quantify Zephyr's effectiveness on different hardware. In this paper, we therefore present results from both our real testbed and our simulator and describe their configurations below.

5.1 Hardware Setup

For our data center experiments, we used an HP c7000 BladeSystem enclosure, shown in Figure 3, with 16 ProLiant BL465c server blades and 10 fans. As shown in Figure 1, the blades and fans within this enclosure are equally divided into two rows. Each blade is equipped with 16 GB of RAM and two AMD 2216 HE processors with two cores each. Each processor has 5 P-states corresponding to frequencies of 2.4 GHz, 2.2 GHz, 2.0 GHz, 1.8 GHz, and 1.0 GHz. Each blade also has two 72 GB 10,000 RPM SAS disks in a RAID-1 configuration. However, the local drives are only used for the administrative domain. To provide shared storage for the VMs, each blade uses QLogic QMH2462 Fibre Channel adapters to connect to a HP EVA 8000 storage array over a 2 Gbit/s FC connection.

The blades used come with seven pre-installed temperature sensors each. Three sensors are located in the CPU region, two for the memory regions, one near the front to measure the inlet air temperature, and one that measures the motherboard temperature. The enclosure also contains an Onboard Administrator

(OA), an embedded module running Linux, that provides integrated enclosure management. The OA allows us to record all the temperature readings as well as the power used by the entire enclosure and each individual fan. It also allows us to control the speed of individual fans between 3,000 and 18,000 RPM.

While we used Xen 3.2.1 [40] as the Virtual Machine Monitor (VMM), the techniques described in this paper would work equally well with any other VMM implementation, including VMware, that includes support for live VM migration [20, 29]. Xen's administrative domain used Ubuntu 7.10 as the operating system and the 2.6.18.8-xen para-virtualized Linux kernel. All VMs were configured with 1 GB RAM, a 4.4 GB virtual hard drive, two Virtual CPUs, and ran the Fedora Core 8 OS with the 2.6.18.8-xen para-virtualized Linux kernel. 64 VMs were used with the traces described below in Section 5.3. We set $Util_{ref} = 75\%$ for all experiments. While the enclosure can handle higher temperatures, we set $T_{ref} = 65$C and the minimum fan speed to 4,000 RPM to ensure equipment safety while conducting our experiments. For our experiments, we determine that the equipment has reached thermal overload if the temperature reaches 70C.

5.2 Simulation Setup

For accurate simulation, we started with the previously validated power and temperature models for the blades and fans that captured the computing and thermal properties of a blade system. The controllers for the simulated fans and blades, in fact, used most of the real system's code. Using the simulator, we were able to evaluate the sensitivity of the power and performance results with respect to the parameters we are interested in. For instance, an increase of 15C for each ambient temperature sensor can model a significantly warmer data center, or a change in a blade's physical configuration that exposes components to higher temperatures.

The accuracy of this simulator was verified by running the IT workload traces, described below, on the real hardware running Zephyr and recording both the system input (inlet air temperature, blade utilization, etc.) and output (fan power, enclosure power, CPU temperature, etc.). The input was then replayed for the simulator and the simulator output was compared to the system output. In most cases, the resulting difference between the two systems was less than 3% with a slightly higher difference when the Zephyr On/Off simulation picked a different consolidation solution than the real experiment.

We then used the simulator to model a number of different experimental configurations and we present the scenario that models a warmer data center in this paper. We used the inlet temperature traces gathered from our real experiments [17] and increased the readings by 15C ($T_{amb} += 15C$) for the simulator. For this scenario, we also model 13 different servers, shown in Table 1, as measured by their idle and peak power. Server 1 represents the real hardware used in our data center experiments. Servers 2–7 represent theoretical energy-proportional systems that are characterized by a negligible power draw when idle ($P_{idle} = 0$) [41]. The remaining servers represent a more practical system where new manufacturing techniques have reduced the power consumed when idle but have possibly higher power usage due to an increase in the number of sockets or cores found in the system. This range of servers are meant to broadly represent future platforms including the use of low-power or embedded processors, a shift to more energy-efficient servers, and scale-up servers being used for workload consolidation.

Server #	P_{idle}	P_{max}	Server #	P_{idle}	P_{max}
1	171	239	8	86	120
2	0	34	9	86	154
3	0	68	10	86	190
4	0	102	11	86	222
5	0	136	12	86	256
6	0	170	13	86	290
7	0	204			

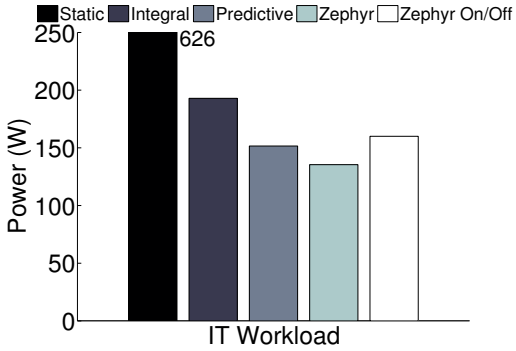
Table 1. Simulated Server Properties

5.3 Benchmarks

To obtain a realistic estimate of the possible savings of our system, we used traces gathered from 64 servers in real data centers. These traces have also been previously used to evaluate the power-efficiency of data center servers [42]. The traces, called the IT Workload in the rest of the evaluation, were gathered from servers running e-commerce and database workloads and are representative of a traditional IT environment found in large corporations. An analysis showed that 80% of the traces had an average utilization lower than 24%. While this low utilization is similar to other data center environments [24], it does not mean that the resource usage is uniform over the entire time period. Our analysis of the traces showed that not only were they bursty, but they also exhibited periodicity. While the traces were gathered over a period of a number of days, in the interests of time, our experiments used a representative four hour-long segment from the busy periods.

5.4 Trace Replay

We used *gamut* [43], a tool used by other researchers to measure performance in a VM [26], to replay the traces. Gamut measures performance in units of work performed per second. Note that if the CPU is overloaded during trace replay, gamut does not allow work between different time intervals to overlap but instead allows time to dilate beyond the specified amount. For example, if a trace indicates that 75% of CPU resources should be consumed during the first five seconds, gamut will translate



Note that the SFC bar is truncated at 250W. The actual value can be found on the label next to it.

Figure 4. Fan Power

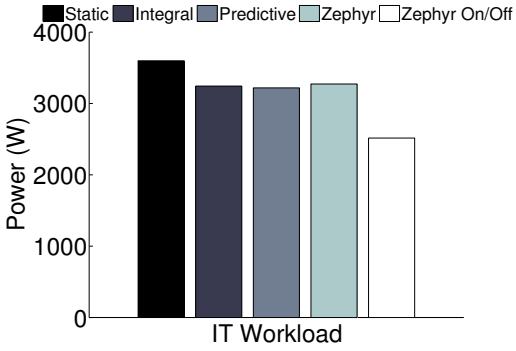


Figure 5. Total Server Enclosure Power

this requirement into units of work and, in overload conditions, it will wait for the given amount of work to be completed instead of simply starting the next portion of the trace after five seconds. Only after completion of the specified units of work will it move on to the next time interval. This behavior captures the overheads of our system. As the total amount of work to be done is fixed by the trace, the increased time taken in overload conditions or due to the overhead of VM migration will result in a reduced number of units of work that can be performed per second.

We were unable to replay disk or network activity because our traces did not include enough information (such as individual network connections or disk requests) to faithfully recreate the original workload. However, we believe that this absence did not significantly change our results. First, all VMs were stored on a SAN and virtual disk activity would not impact power usage at the enclosure level. In this scenario, other complementary techniques [44, 45] could be applied to reduce SAN power usage. Second, the power usage of most network ports tends to be independent of the network utilization [46]. Also, any increased CPU utilization due to I/O overhead in virtualized systems would be automatically included in the power model.

5.5 Experimental Methodology

The evaluation of Zephyr aims to answer two main questions. First, what power savings are possible from using Zephyr? Second, what is the impact of the controllers on performance?

To answer the first question and to individually quantify the power benefits from each part of our system, we used the above traces in five different configurations. The EC was used in all of these configurations. All 64 VMs were distributed over the blades in the same round-robin manner at the beginning of each experiment. All power results are the average over a four hour duration with measurements recorded every 10 seconds.

Static: Apart from the efficiency controller, a Static Fan Controller (SFC) was used to set the fans to a speed where, even under maximum load, the blade temperatures could not exceed T_{ref}. We experimentally determined this to be 12,500 RPM for our hardware setup. The SFC is based on fan controllers found in prior generations of hardware.

Integral: Apart from the efficiency controller, the integral fan controller (IFC) was used to control the fan speeds. No VM consolidation was performed.

Predictive: Apart from the efficiency controller, Zephyr's predictive fan controller (FC) was used to control the fan speeds.

Zephyr: In addition to Zephyr's fan controller and efficiency controller, the group controller (GC) was used to place workloads in cooling-efficient portions of the enclosure but did not turn idle blades off.

Zephyr On/Off: While Zephyr's fan, efficiency, and group controllers are all used, the GC was able to turn blades on or off in response to changes in demand.

To answer the second question about performance and to provide a performance baseline to compare Zephyr against, we also ran the workloads in two different "vanilla" configurations without any controllers. In the *Isolated* configuration, each trace was individually run in its VM on a blade with no other co-located workloads. In the *Co-located* configuration, all traces were simultaneously run in their VMs with an identical VM placement to that of the above five configurations. As we saw no performance difference between the two, because of the low average utilization of the traces, we use the co-located performance results as the baseline.

5.6 Data Center Results

Figure 4 shows the fan power savings obtained from using the different controllers. Without workload consolidation, we observe that Zephyr's FC, when compared to the static baseline (SFC), can reduce fan power usage by 74%. However, the SFC is a naïve baseline and is only shown in this figure to give an estimate of the maximum power that could be used. A more representative comparison is that of Zephyr's FC with the Integral

controller (IFC). As seen in the figure, Zephyr's FC can reduce the IFC's power usage by 21%.Overall, more power was consumed by the IFC as the fans under its control were driven to higher speeds than those of the FC. Unlike the IFC, where the fans in the same row run at the same speed, the FC varies the fan speeds with a much finer granularity and is able to provide "on-demand" cooling to the blades.

Once Zephyr's GC was allowed to move workloads to more cooling-efficient regions of the enclosure, we see that the FC was able to further reduce fan power usage by 30% when compared to the IFC. In the Zephyr On/Off case, where the GC is allowed to turn idle machines off, the fan power consumed actually rises when compared to the above scenarios. This occurs because consolidation causes the average utilization to increase and the consequent temperatures rise in the active blades requires an increase in fan speeds. This behavior is still desirable because, as shown later in Figure 5, the savings from turning blades off outweighs the increase in cooling costs for the current system. Note that, if evenly distributing load had been more power-efficient than consolidation, the GC would have consolidated less aggressively. However, even with consolidation, Zephyr's FC can reduce the fan power consumed by the IFC by 17%. In all of our experiments, the thermal overload temperature (70C) was never reached and the average CPU temperature was below T_{ref} (65C).

The power usage of the enclosure in all its different configurations is shown in Figure 5. Apart from the power consumed by the blades and the fans, the figure also includes the power used by the enclosure's networking and SAN switches, and the OA. Without consolidation, the power between the SFC and the IFC and Predictive FC configurations differs mainly due to the power usage of the different cooling controllers. Even though the Predictive FC and Zephyr configurations showed wins for fan power, the total enclosure power when compared with the IFC is within 1% of each other for both workloads as the fan power was a very small fraction of total power when all the servers were on. However, we believe this result to be more of an artifact of our current experimental system and data center environment rather than a measure of Zephyr's effectiveness at the enclosure level. As shown later in Section 5.7, Zephyr's predictive fan controller and cooling-aware GC can show higher savings in terms of total enclosure power when used in warmer data centers or on hardware with different power characteristics. Finally, once the Zephyr On/Off configuration is introduced, it can rapidly consolidate workloads and turn idle machines off. For example, when compared to the FC configuration, Zephyr On/Off can reduce power usage by 23% for the IT workload.

5.7 Simulation Results

The results for our 13 simulated servers, described in Table 1, running in a warmer data center where the ambient temperature has been raised by 15C are presented in Figure 6. Figure 6 (a) and (b) presents the power used and temperature violations results for the Predictive, Zephyr, and Zephyr On/Off systems. The power and temperature violations for these three systems is compared to the IFC over the simulated four hour run. Note that each bar in Figure 6 (a) represents the power usage of the *entire* enclosure and includes both servers and fans. Each bar is broken up into two colors that show the fraction of savings derived from each component (servers or fans). Each bar in Figure 6 (b) represents the reduction in violations measured as a fraction of the total number of readings $\left(\frac{Viol_{new}-Viol_{old}}{TotalReadings}\right)$.

If we only compare the power used by Zephyr's Predictive FC to the IFC, we see from Figure 6 (a) that the power used by the FC is lower in most cases with the exceptions of the server 1, 12, and 13 cases. The reason for the increased power usage can be seen in Figure 6 (b) as the FC has less temperature violations than the IFC. Therefore, the increased power usage is due to improved correctness.

When comparing Zephyr's predictive FC and its GC that can place workloads in cooling efficient regions of the enclosure, labeled Zephyr in the figure, to the IFC, we see a reduction in total system power up to 17% and this is entirely due to the fans[1]. Further, by looking at Figure 6 (b), we see that this is accompanied by reductions in violations of up to 7%. The two cases where we see an increase of violations (servers 9 and 10) can be treated as noise as it is less than 0.1%.

Finally, when we allow Zephyr to turn machines on or off in response to total load, labelled Zephyr On/Off in the figure, we see even higher savings of $\sim$24%. By examining the different servers, we can also see the relation between a server's power model and the amount of savings possible through consolidation vs. cooling. Once again, these savings are obtaining with no significant increase in temperature violations and a significant reduction in some cases.

5.8 Performance Results

To examine Zephyr's impact on performance, measured in terms of gamut units of work done per second, we compared it to the Co-located performance baseline described in Section 5.5. Overall, for both the real hardware and simulated cases, Zephyr has a small, and almost negligible impact on performance. As performance is not impacted by fan speeds, the performance for both workloads with the SFC, IFC, and FC controllers was identical to the baseline. Even with the GC configuration turning machines on and off, Zephyr's impact on performance, compared to the baseline, was between 0–2% as live migration is very efficient [20]. A detailed description of these results is therefore omitted here in the interests of brevity.

[1]There is a small server power saving (<1%) in a few cases where the slight performance loss during migration manifests itself as power savings.

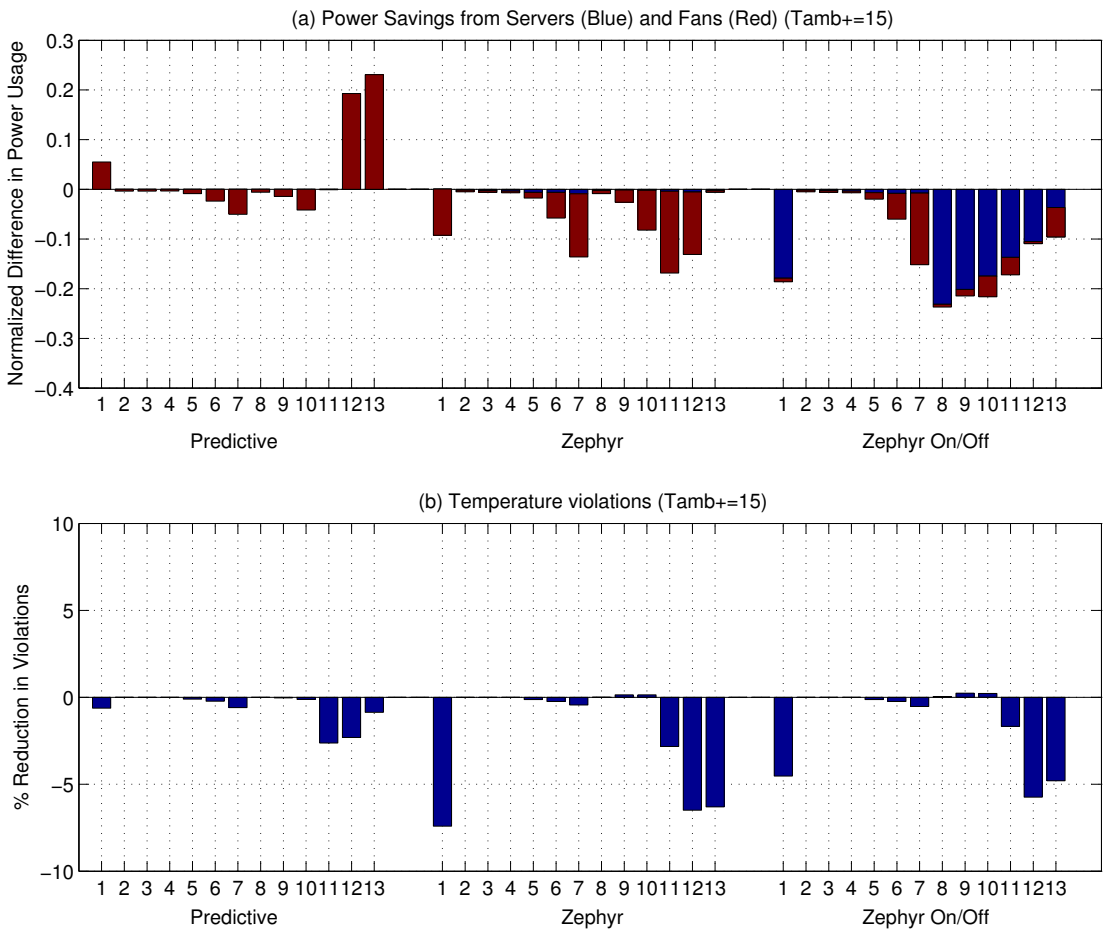


Figure 6. Simulation Results with $T_{amb}+=15C$

6 Conclusion

This paper has presented Zephyr, a unified power and cooling management system for blade servers that addresses *overall* power efficiency including the power consumed for cooling. Our work combines system design and virtualization with fundamental concepts from heat transfer theory to develop powerful models and controllers, and our results, based on a full prototype implementation exercised by real-world enterprise traces, demonstrate significant benefits (up to 30% improvement in cooling power usage and 29% enclosure power reductions). Overall, as power management continues to increase in importance for enterprise environments, we believe our approach has significant promise and will likely be a critical part of future solutions.

REFERENCES

[1] U.S. Environmental Protection Agency (EPA), 2007. Report to congress on server and data center energy efficiency, public law 109-431, Aug.

[2] Chase, J. S., Anderson, D. C., Thakar, P. N., Vahdat, A. M., and Doyle, R. P., 2001. "Managing energy and server resources in hosting centers". In Proceedings of the 18th ACM Symposium on Operating Systems Principles, pp. 103–116.

[3] Elnozahy, E. N., Kistler, M., and Rajamony, R., 2002. "Energy-efficient server clusters". In Proceedings of the 2nd International Workshop on Power-Aware Computer Systems (PACS), pp. 179–196.

[4] Heath, T., Diniz, B., Carrera, E. V., Jr., W. M., and Bianchini, R., 2005. "Energy conservation in heterogeneous server clusters". In Proceedings of the ACM Symposium on Principles and Practice of Parallel Programming (PPOPP), pp. 186–195.

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

[6] Pering, T., Burd, T., and Brodersen, R., 1998. "The simulation and evaluation of dynamic voltage scaling algorithms". In Proceedings of the 1998 International Symposium on Low Power Electronics and Design (ISLPED), pp. 76–81.

[7] Pinheiro, E., Bianchini, R., Carrera, E. V., and Heath, T., 2003. "Dynamic cluster reconfiguration for power and performance". *Compilers and Operating Systems for Low Power*, pp. 75–93.

[8] Raghavendra, R., Ranganathan, P., Talwar, V., Wang, Z., and Zhu, X., 2008. "No "power" struggles: Coordinated multi-level power management for the data center". In Proceedings of the 13th International Conference on Architectural Support for Programming Languages and Operating Systems

(ASPLOS), pp. 48–59.

[9] Rajamani, K., and Lefurgy, C., 2003. "On evaluating request-distribution schemes for saving energy in server clusters". In Proceedings of the IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), pp. 111–122.

[10] Ranganathan, P., Leech, P., Irwin, D. E., and Chase, J. S., 2006. "Ensemble-level power management for dense blade servers". In Proceedings of the 33rd International Symposium on Computer Architecture (ISCA 2006), pp. 66–77.

[11] Greenberg, S., Mills, E., Tschudi, B., Rumsey, P., and Myatt, B., 2006. "Best practices for data centers: Results from benchmarking 22 data centers". In Proceedings of the 2006 ACEEE Summer Study on Energy Efficiency in Buildings.

[12] Patel, C. D., Bash, C. E., Sharma, R., Beitelmam, M., and Friedrich, R. J., 2003. "Smart cooling of data centers". In Proceedings of IPACK'03, The Pacific Rim/ASME International Electronic Packaging Technical Conference and Exhibition.

[13] Bash, C., and Forman, G., 2007. "Cool job allocation: Measuring the power savings of placing jobs at cooling-efficient locations in the data center". In Proceedings of the USENIX Annual Technical Conference, pp. 363–368.

[14] Bash, C. E., Patel, C. D., and Sharma, R. K., 2006. "Dynamic thermal management of air cooled data centers". In Proceedings of the 10th International Conference on Thermal and Thermomechanical Phenomena in Electronics Systems (ITHERM), pp. 445–452.

[15] Heath, T., Centeno, A. P., George, P., Ramos, L., Jaluria, Y., and Bianchini, R., 2006. "Mercury and freon: Temperature emulation and management for server systems". In Proceedings of the 12th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), pp. 106–116.

[16] Moore, J., Chase, J., Ranganathan, P., and Sharma, R., 2005. "Making scheduling "cool": Temperature-aware workload placement in data centers". In Proceedings of the USENIX Annual Technical Conference, pp. 61–75.

[17] Wang, Z., Bash, C., Tolia, N., Marwah, M., Zhu, X., and Ranganathan, P., 2009. "Optimal fan control for thermal management of servers". In Proceedings of the ASME/Pacific Rim Electronic Packaging Technical Conference and Exhibition (InterPACK '09).

[18] Bianchini, R., and Rajamony, R., 2004. "Power and energy management for server systems". *IEEE Computer,* **37**(11), pp. 68–74.

[19] Heo, J., Henriksson, D., Liu, X., and Abdelzaher, T., 2007. "Integrating adaptive components: An emerging challenges in performance-adaptive systems and a server farm case-study". In Proceedings of the 28th IEEE International Real-Time Systems Symposium (RTSS).

[20] Clark, C., Fraser, K., Hand, S., Hansen, J. G., Jul, E., Limpach, C., Pratt, I., and Warfield, A., 2005. "Live migration of virtual machines". In Proceedings of the 2nd Symposium on Networked Systems Design and Implementation (NSDI), pp. 273–286.

[21] IDC, 2008. Worldwide quarterly server tracker, Q4 2007, Feb.

[22] Data Center Users' Group, 2008. Spring 2008 Data center users' group survey results. `http://datacenterug.org/`.

[23] Goldberg, R.P., 1974. "Survey of Virtual Machine Research". *IEEE Computer,* **7**(6), June.

[24] Vogels, W., 2008. "Beyond server consolidation". *ACM Queue,* **6**(1), pp. 20–26.

[25] IDC, 2007. Worldwide virtual machine software 2006-2010 forecast, Jan.

[26] Grit, L., Irwin, D., Yumerefendi, A., and Chase, J., 2006. "Virtual machine hosting for networked clusters: Building the foundations for "autonomic" orchestration". In Proceedings of the 2nd International Workshop on Virtualization Technology in Distributed Computing (VTDC), p. 7.

[27] McNett, M., Gupta, D., Vahdat, A., and Voelker, G. M., 2007. "Usher: An extensible framework for managing clusters of virtual machines". In Proceedings of the 21st Conference on 21st Large Installation System Administration (LISA), pp. 1–15.

[28] Figueiredo, R. J. O., Dinda, P. A., and Fortes, J. A. B., 2003. "A case for grid computing on virtual machines". In Proceedings of the 23rd International Conference on Distributed Computing Systems (ICDCS), pp. 550–559.

[29] Nelson, M., Lim, B.-H., and Hutchins, G., 2005. "Fast transparent migration for virtual machines". In Proceedings of the 2005 Annual Conference on USENIX Annual Technical Conference, pp. 391–394.

[30] Ruth, P., Rhee, J., Xu, D., Kennell, R., and Goasguen, S., 2006. "Autonomic live adaptation of virtual computational environments in a multi-domain infrastructure". In Proceedings of The 3rd IEEE International Conference on Autonomic Computing (ICAC), pp. 5–14.

[31] Steinder, M., Whalley, I., Carrera, D., Gaweda, I., and Chess, D. M., 2007. "Server virtualization in autonomic management of heterogeneous workloads". In Proceedings of the 10th IFIP/IEEE International Symposium on Integrated Network Management (IM), pp. 139–148.

[32] Wood, T., Shenoy, P. J., Venkataramani, A., and Yousif, M. S., 2007. "Black-box and gray-box strategies for virtual machine migration". In Proceedings of the 4th Symposium on Networked Systems Design and Implementation (NSDI), pp. 229–242.

[33] Cully, B., Lefebvre, G., Meyer, D., Feeley, M., Hutchinson, N., and Warfield, A., 2008. "Remus: High availability via asynchronous virtual machine replication". In Proceedings of the 5th Symposium on Networked Systems Design and Implementation (NSDI), pp. 161–174.

[34] Chen, Y., Das, A., Qin, W., Sivasubramaniam, A., Wang, Q., and Gautam, N., 2005. "Managing server energy and operational costs in hosting centers". In Proceedings of the International Conference on Measurements and Modeling of Computer Systems (SIGMETRICS), pp. 303–314.

[35] Nathuji, R., and Schwan, K., 2007. "Virtualpower: coordinated power management in virtualized enterprise systems". In Proceedings of the 21st ACM Symposium on Operating Systems Principles (SOSP), pp. 265–278.

[36] Stoess, J., Lang, C., and Bellosa, F., 2007. "Energy management for hypervisor-based virtual machines". In Proceedings of the 2007 USENIX Annual Technical Conference, pp. 1–14.

[37] Russell, S., and Norvig, P., 2003. *Artificial Intelligence: A Modern Approach*, 2nd edition ed. Prentice-Hall.

[38] cvxopt. `http://abel.ee.ucla.edu/cvxopt/`.

[39] Pallipadi, V., and Starikovsky, A., 2006. "The ondemand governor - past, present, and future". In Proceedings of the 2006 Linux Symposium, Vol. 2, pp. 215–229.

[40] Barham, P., Dragovic, B., Fraser, K., Hand, S., Harris, T., Ho, A., Neugebauer, R., Pratt, I., and Warfield, A., 2003. "Xen and the art of virtualization". In Proceedings of the 19th ACM Symposium on Operating Systems Principles, pp. 164–177.

[41] Barroso, L. A., and Hölzle, U., 2007. "The case for energy-proportional computing". *IEEE Computer,* **40**(12), pp. 33–37.

[42] Meisner, D., Gold, B. T., and Wenisch, T. F., 2009. "Powernap: Eliminating server idle power". In Proceedings of the 14th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS '09).

[43] gamut. `http://issg.cs.duke.edu/cod/`.

[44] Joukov, N., and Sipek, J., 2008. "Greenfs: Making enterprise computers greener by protecting them better". In Proceedings of the 2008 EuroSys Conference, pp. 69–80.

[45] Narayanan, D., Donnelly, A., and Rowstron, A., 2008. "Write off-loading: Practical power management for enterprise storage". In Proceedings of the 6th USENIX Conference on File and Storage Technologies, pp. 253–267.

[46] Fan, X., Weber, W.-D., and Barroso, L. A., 2007. "Power provisioning for a warehouse-sized computer". In Proceedings of the 34th Annual International Symposium on Computer Architecture (ISCA '07), pp. 13–23.