Delivering Energy Proportionality with Non Energy-Proportional Systems – Optimizing the Ensemble

Niraj Tolia, Zhikui Wang, Manish Marwah, Cullen Bash
Parthasarathy Ranganathan, Xiaoyun Zhu
HP Labs, Palo Alto

Abstract

With power having become a critical issue in the operation of data centers today, there has been an increased push towards the vision of "energy-proportional computing", in which no power is used by idle systems, very low power is used by lightly loaded systems, and proportionately higher power at higher loads. Unfortunately, given the state of the art of today's hardware, designing individual servers that exhibit this property remains an open challenge. However, even in the absence of redesigned hardware, we demonstrate how optimization-based techniques can be used to build systems with off-the-shelf hardware that, when viewed at the aggregate level, approximate the behavior of energy-proportional systems. This paper explores the viability and tradeoffs of optimization-based approaches using two different case studies. First, we show how different power-saving mechanisms can be combined to deliver an aggregate system that is proportional in its use of server power. Second, we show early results on delivering a proportional cooling system for these servers. When compared to the power consumed at 100% utilization, results from our testbed show that optimization-based systems can reduce the power consumed at 0% utilization to 15% for server power and 32% for cooling power.

1 Introduction

With power having become a critical issue in the operation of data centers, the concept of energy-proportional computing [3] or energy scaledown [13] is drawing increasing interest. A system built according to this principle would, in theory, use no power when not being utilized, with power consumption growing in proportion to utilization. Over the last few years, a number of techniques have been developed to make server processors more efficient, including better manufacturing techniques and the use of Dynamic Voltage and Frequency Scaling (DVFS) [5, 7, 12, 17, 18] for runtime power optimization. While the savings are significant, this has lead to CPUs no longer being responsible for the majority of power consumed in servers today. Instead, subsystems that have not been optimized for power-efficiency, such as network cards, hard drives, graphics processors, fans, and power supplies, have started dominating the power consumed by systems, especially during periods of low utilization.

Instead of waiting for all these different technologies to deliver better energy-efficiency, this paper advocates that energy-proportional computing can be approximated by using software to control power usage at the ensemble level. An ensemble is defined as a logical collection of servers and could range from an enclosure of blades, a single rack, groups of racks, to even an entire data center.

It was previously difficult to dynamically balance workloads to conserve power at the ensemble level for a number of reasons. Unnecessarily shutting down applications to simply restart them elsewhere is looked upon as a high-risk, high-cost change because of the performance impact, the risk to system stability, and the cost of designing custom control software. However, the re-emergence of virtualization and the ability to "live" migrate [6] entire Virtual Machines (VMs), consisting of OSs and applications, in a transparent and low-overhead manner, will enable a new category of systems that can react better to changes in workloads at the aggregate level. By moving workloads off under-utilized machines and then turning idle machines off, it should now be possible to approximate, at an ensemble level, the behavior found in theoretical energy-proportional systems.

However, virtualization is not a magic bullet and a naïve approach to consolidation can hurt application performance if server resources are overbooked or lead to reduced power savings when compared to the maximum possible. This paper therefore advocates a more rigorous approach in the optimization of ensemble systems including the use of performance modeling, optimization, and control theory. Finally, instead of only looking at more traditional server components such as storage, CPUs, and the network, optimization-based systems should also consider control of other components such as server fans and Computer Room Air Conditioners (CRACs) as cooling costs are rapidly becoming a limiting factor in the design and operation of data centers today [10, 16, 20].

We use two case studies to demonstrate that it is possible, by applying optimizations at the ensemble layer, to de-

liver energy proportionality with non energy-proportional systems. First, we show how the use of a VM migration controller, that can also turn machines on or off in addition to DVFS in response to demand changes, can reduce the power consumed by servers and exhibit power-usage behavior close to that of an energy-proportional system. Second, we demonstrate how a power and workload-aware cooling controller can exhibit the same behavior for cooling equipment such as server fans.

2 Case Studies

It has been advocated that optimization-based algorithms should be preferred over ad hoc heuristics in making system runtime and management decisions [11]. We will therefore not stress this point further but, through the use of two case studies, show how optimization can be used to deliver energy proportionality at the ensemble layer. It should be stressed that the focus of this section is not on the use of any *particular* algorithm but instead on how non energy-proportional systems can be combined to approximate the behavior of an energy-proportional system.

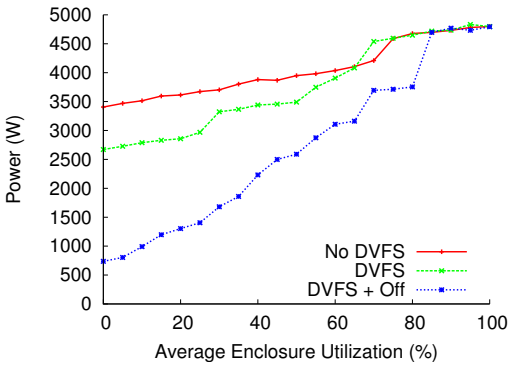
2.1 Experimental Setup

To evaluate energy proportionality in the case studies, we used an HP c7000 BladeSystem enclosure with 16 ProLiant BL465c server blades and 10 fans. Each blade was equipped with 16 GB of RAM and two AMD 2216 HE dual-core processors. Each processor has 5 P-states (a voltage and frequency setting) corresponding to frequencies of 2.4, 2.2, 2.0, 1.8, and 1.0 GHz. The blades and the fans are equally divided among two rows on the front and back ends of the enclosure respectively. Each blade is cooled by multiple fans, and each fan draws cool air in through the front of the blades. The enclosure allows us to measure blade temperatures and the power consumed by different components.

We used Xen 3.2.1 [2] with both the administrative domain and the VMs using the 2.6.18.8 para-virtualized Linux kernel. The Xen administrative domain is stored on local disks while the VMs use a storage area network (SAN). We used 64 VMs configured with 128 MB of RAM, a 4.4 GB virtual hard drive, and one Virtual CPU. We set each VM's memory to a low value to allow us to evaluate a large configuration space for workload placement. In production environments, the same effect could be achieved at runtime through the use of page sharing or ballooning [21]. We used *gamut* [8] to experiment with different VM (and physical server) utilizations.

2.2 Server Energy Proportionality

We examined the energy proportionality in the enclosure layer using three different policies. The first policy, No DVFS, uses no power-saving features. The second policy, DVFS, uses hardware-based voltage and frequency scaling and is very similar to Linux's OnDemand governor [15]. The third policy, DVFS+Off, uses DVFS plus a VM migration controller that consolidates virtual machines and turns idle machines off. Its algorithm uses the blade's power models for the different P-states and sensor readings from resource monitoring agents. The optimization problem is similar to that in [18], with constraints on the blade CPU and memory utilization to prevent overbooking. Examples of other constraints that could be modeled include network and storage utilization. Our experience has shown that algorithms such as bin-packing or simulated annealing work well in this scenario.



Each result presented above is an average of approximately 90 readings over a 15 minute interval.

Figure 1: Enclosure Power Usage (Blades, Network)

For this case study, we varied each VM's utilization in the range of $[0,5,\ldots,100]\%$ (of a single core) and measured the power consumed by the enclosure. The results are presented in Figure 1, where the x-axis shows the enclosure utilization by all 64 VMs as a percentage of total capacity of all the 16 blades. As each server has 4 cores, the utilization also corresponds to each VM's utilization (as a percentage of a single core). The measured power, shown on the y-axis, includes the power consumed by the blades as well as the networking, SAN, and management modules. With No DVFS, we notice a very small power range (1,393 W) between 100% and 0% utilization and the minimum power used is 3,406 W, or 71% of the power consumed at 100% utilization, significantly away from the theoretical minimum of 0 W. Once DVFS is enabled, the power range increases but, as seen in the figure, the system is still not energy-proportional and consumes 2,670 W at 0% utilization. It is only when we look at the DVFS+Off policy that the system starts approximating energy proportionality at the ensemble level. The range of power used between 100% and 0% utilization is 4,164 W and at 0% utilization, there is only a 737 W difference, or 15% of the power consumed at 100% utilization, from the theoretical minimum of 0 W.

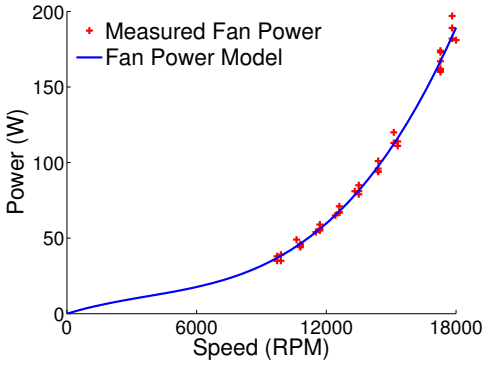
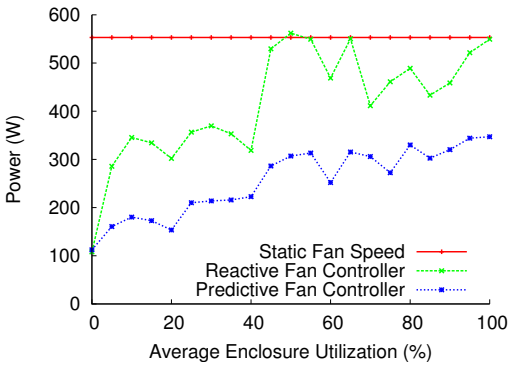


Figure 2: Fan Power Consumption and Model



Each result presented above is an average of approximately 90 readings over a 15 minute interval.

Figure 3: Fan Power

Note from the figure that, at high utilizations (above 80%), all three policies have the same power usage as they are all running near peak performance. While the DVFS policy only shows a noticeable difference when it can switch to a lower P-state, the DVFS+Off policy starts showing benefit around 80% utilization as it can start consolidating VMs and turning machines off.[1]

Even at 0% utilization, zero power usage was not achieved with DVFS+Off for a number of reasons. First, at least one active server is needed to host all the VMs. Second, our enclosure contained two network switches, two SAN switches, two management modules, and six power supplies. Most of these components are not energy-proportional. Finally, like all industry standard servers, even off servers have an active management processor that is used for network management tasks such as remote KVM and power cycling.

2.3 Cooling Energy Proportionality

While we showed how energy proportionality could be achieved for server power in Section 2.2, cooling equipment also consumes a significant part of the total power used by a data center. In particular, server fans can consume between 10–25% of total server power and, at the data center level, cooling can account for as much as 50% of the total power consumed [10, 16, 20]. In this case study, we therefore examine how intelligent fan control can be used to achieve better energy proportionality for server cooling resources. We continue to use the experimental setup described in Section 2.1 and use the DVFS+Off policy, presented in Section 2.2, for managing server power. The objective of fan control is to provide enough cool air for the blades so that the server temperatures can be maintained below thresholds for thermal safety reasons. In this paper, we specifically evaluate two different fan controllers – a Reactive Fan Controller (RFC) and a Predictive Fan Controller (PFC).

The Reactive Fan Controller (RFC) is a simple feedback controller that changes fan speeds based on the data gathered from the hardware temperature sensors present on all the blades. Because of the complexity of sharing 10 fans between 16 blades, the RFC synchronously changes the speed of all fans in each row (there are two rows of blades and fans in the enclosure) in response to the maximum observed temperature for that row. When the maximum temperature is above the threshold, the fan speeds are increased to provide a larger volume of air flow to cool the servers, and vice versa. The RFC is similar to commercial fan controllers used in industry today.

The Predictive Fan Controller (PFC) aims to minimize the total fan power consumption without violating temperature constraints. It uses temperature sensors in the blades as well as software sensors to monitor server utilization. Figure 2 shows the power model for a fan in our enclosure, where the fan power is a cubic function of the fan speed, which closely matches the measured fan power. Note that the volume air flow rate is approximately proportional to the fan speed. Also note that each fan provides different levels of cooling resource (i.e. cool air) to each individual blade according to the location of the fan with respect to the blade. In this regard, each fan has a unique cooling efficiency with respect to each blade. This provides an opportunity to minimize the fan power consumption by exploring the variation in cooling efficiency of different fans for different blades along with the time-varying demands of the workloads. We built a thermal model [19] empirically that explicitly captures the cooling efficiency between each pair of blade and fan. This thermal model, together with the blade and the fan power models, is used by the PFC to predict future server temperatures for any given fan speed and measured server utilization. By representing this as a convex, constrained optimization problem [19], the PFC is able to use an off-the-shelf convex optimization solver to set the fan speeds to values that potentially minimize the

[1]One anomalous reading noticeable in Figure 1 is for the No DVFS and DVFS settings at 70% and 65% utilization levels. While the processor reports being at the highest P-state, we recorded a sharp drop in power usage at these points. We believe that the processor is using an internal power-saving scheme that it disables at high utilizations.

aggregate power consumption by the fans while keeping the blade temperatures below their thresholds.

The results for the two controllers are presented in Figure 3. The figure also includes the power consumed by setting the fans to a static speed that, independent of the workloads, is guaranteed to keep temperatures below the threshold under normal ambient operating conditions.

When examining the RFC's performance, it is helpful to note the relationship between VM utilization and fan speed. For a given row of fans, the fan speed is directly controlled by the maximum CPU temperature in the row. Furthermore, CPU temperature is a function of blade utilization and blade ambient, or inlet, temperature. Blade utilization, however, does not always directly correlate to VM utilization. For example, with each VM utilization set to 85% (one VCPU), the system cannot fit more than 4 VMs per machine with an overall blade utilization of 340% (four CPUs). However, with a reduced VM utilization of 80%, each blade can accommodate 5 VMs with an overall blade utilization of 400%. As a blade's CPU temperature will be much higher with a utilization of 400% vs. 340%, it will require a greater amount of cooling and therefore use more power due to increased fan speeds. Similarly, if a blade is located in an area of increased ambient temperature, that blade could also drive fan speed higher, if and when it becomes utilized, even if the utilization levels are relatively low. These factors are responsible for the RFC operating in two power bands; approximately between 410–560 W when the utilization ranges between 45–100% and between 290–370 W when the utilization ranges between 5–40%. Even at 5% utilization, the RFC still uses 52% of the peak power used at 100% utilization.

In contrast, due to its knowledge of the cooling efficiencies of different fans, the demand levels of the individual blades, and their ambient temperatures, the PFC is able to set fan speeds individually and avoid the correlated behavior exhibited by controllers like the RFC. Overall, the PFC induces an approximately linear relationship between the fan power and the aggregate enclosure utilization and, at 0% utilization, the PFC only consumes 32% of the power it uses at 100% utilization. The use of model-based optimization also allows the PFC to perform significantly better than the RFC. When we compare the two controllers, the PFC can reduce fan power usage by $\sim 40\%$ at both 100% and 5% utilization. At 5% utilization, the PFC only consumes 29% of the power used by the RFC at 100%.

However, the PFC (and the RFC) with zero load is unable to reduce its power usage to 0 W. This was not a deficiency of our optimization-based approach but was due to the fact that the fans were also responsible for cooling the blade enclosure's networking and management modules. We therefore had to lower-bound the fan speeds to ensure that these non energy-proportional components did not accidentally overheat.

3 Assumptions

Even though we used a homogeneous set of machines in our cases studies, our experience has shown that these algorithms can be extended with different power and thermal models to control an ensemble composed of heterogeneous hardware. Further, current generation of processors from both Intel and AMD support CPUID masking [1] that allows VMs to migrate between processors from different families. This work also assumes that it is possible to migrate VM identities such as IP and network MAC addresses with the VMs. While this is generally not a problem in a single data center, migrating Storage Area Network (SAN) identities can sometimes be problematic. However, vendor products such as HP's Virtual Connect and Emulex's Virtual HBA have introduced a layer of virtualization in the storage stack to solve this problem.

Note that it might be difficult to adopt the optimization-based solutions similar to those proposed in this paper to applications that depend on locally-attached storage for their input data. However, the fact that such locally-attached storage systems usually replicate data for reliability and availability reasons [9] might provide a possible solution. In such a scheme, the optimization algorithms could be made aware of the dependencies between the VMs and the locations of their datasets. Given this information, the algorithms could find a suitable consolidated mapping of VMs to physical machines. In order to make this scheme effective, a higher degree of replication might be needed to give the algorithms more flexibility in making placement decisions. This change essentially boils down to a tradeoff between the cost of increased storage capacity versus energy savings.

Finally, our approach approximates energy proportionality by turning machines off. Concerns have been previously raised about reliability of both servers and disk drives due to an increased number of on-off cycles. However, our conversations with blade-system designers have shown that it should not affect server-class machines or, given the large number of on-off cycles supported by disk drives, their internal storage systems during their normal lifetime. Aggressive consolidation might also hurt application availability if the underlying hardware is unreliable. As most applications tend to be distributed for fault-tolerance, introducing application awareness into the consolidation algorithm to prevent it from consolidating separate instances of the same application on the same physical machine will address this issue.

4 Concluding Remarks

This paper has shown that it is possible to use optimization-based techniques to approximate energy-proportional behavior at the ensemble level. Even though our techniques result in some added complexity in the system, in the form

of models, optimization routines, and controllers, we believe the power savings are significant enough to justify it. However, better instrumentation can help us get even closer to theoretical energy proportionality. For example, if temperature sensors, which are relatively cheap to deploy, were installed in the networking and SAN switches, it would have allowed our fan controller to have more complete knowledge of the thermal condition inside the enclosure so that more efficient fan speed optimization could be achieved. In addition, we have used CPU utilization in this case study as a proxy for application-level performance. To directly evaluate and manage application performance, we will need sensors that measure application-level metrics such as throughput and response time. These sensors can be provided by parsing application logs or by monitoring user requests and responses.

While the controllers shown in this paper assumed a single management domain, further study needs to be done to show how they would work in a federated environment where information would need to be shared between controllers at different management layers and possibly from different vendors. We believe that some of the Distributed Management Task Force (DMTF) standards would help address at least some of these issues.

Prior work exists that looked at data center level cooling efficiency by manipulation of CRAC unit settings [4] or by temperature-aware workload placement [14]. However, given that these studies only looked at total power consumption, a more careful investigation is needed in the context of this paper. For example, the model-based optimization approach we used in the Predictive Fan Controller may be applied to the CRAC unit control problem studied in [4] to achieve energy proportionality for the cooling equipment at the data center level.

Finally, even though this paper demonstrated energy proportionality at the ensemble layer, this does not preclude the need for better energy efficiency for individual components such as disks, memory, and power supplies. We believe that new hardware designs with finer levels of power control will help in designing energy efficient systems at both the single server and ensemble layer.

References

[1] AMD. Live migration with AMD-V extended migration technology. Whitepaper, 2007.

[2] P. Barham, B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho, R. Neugebauer, I. Pratt, and A. Warfield. Xen and the art of virtualization. In *Proc. of the 19th ACM Symposium on Operating Systems Principles (SOSP)*, pages 164–177, 2003.

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

[4] C. E. Bash, C. D. Patel, and R. K. Sharma. Dynamic thermal management of air cooled data centers. In *Proc. of the 10th International Conference on Thermal and Thermomechanical Phenomena in Electronics Systems (ITHERM)*, pages 445–452, San Diego, CA, May 2006.

[5] Y. Chen, A. Das, W. Qin, A. Sivasubramaniam, Q. Wang, and N. Gautam. Managing server energy and operational costs in hosting centers. In *Proc. of the International Conference on Measurements and Modeling of Computer Systems (SIGMETRICS)*, pages 303–314, Banff, Canada, June 2005.

[6] C. Clark, K. Fraser, S. Hand, J. G. Hansen, E. Jul, C. Limpach, I. Pratt, and A. Warfield. Live migration of virtual machines. In *Proc. of the 2nd Symposium on Networked Systems Design and Implementation (NSDI)*, pages 273–286, Boston, MA, May 2005.

[7] E. N. Elnozahy, M. Kistler, and R. Rajamony. Energy-efficient server clusters. In *Proc. of the 2nd International Workshop on Power-Aware Computer Systems (PACS)*, pages 179–196, Cambridge, MA, Feb. 2002.

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

[9] S. Ghemawat, H. Gobioff, and S.-T. Leung. The google file system. In *Proc. of the 19th ACM Symposium on Operating Systems Principles (SOSP)*, pages 29–43, Bolton Landing, NY, Oct. 2003.

[10] S. Greenberg, E. Mills, B. Tschudi, P. Rumsey, and B. Myatt. Best practices for data centers: Results from benchmarking 22 data centers. In *Proc. of the 2006 ACEEE Summer Study on Energy Efficiency in Buildings.*, pages 76–87, Pacific Grove, CA, Aug. 2006.

[11] K. Keeton, T. Kelly, A. Merchant, C. Santos, J. Wiener, X. Zhu, and D. Beyer. Don't settle for less than the best: use optimization to make decisions. In *Proc. of the 11th Workshop on Hot Topics in Operating Systems (HOTOS '07)*, San Diego, CA, May 2007.

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

[13] R. N. Mayo and P. Ranganathan. Energy consumption in mobile devices: Why future systems need requirements-aware energy scale-down. In *Proc. of 3rd International Workshop on Power-Aware Computer Systems (PACS)*, pages 26–40, San Diego, CA, Dec. 2003.

[14] J. Moore, J. Chase, P. Ranganathan, and R. Sharma. Making scheduling "cool": Temperature-aware workload placement in data centers. In *Proc. of the USENIX Annual Technical Conference*, pages 61–75, Anaheim, CA, Apr. 2005.

[15] V. Pallipadi and A. Starikovsky. The ondemand governor - past, present, and future. In *Proc. of the 2006 Linux Symposium*, volume 2, pages 215–229, Ottawa, Canada, July 2006.

[16] C. D. Patel, C. E. Bash, R. Sharma, M. Beitelmam, and R. J. Friedrich. Smart cooling of data centers. In *Proc. of the Pacific Rim/ASME International Electronic Packaging Technical Conference and Exhibition (IPACK'03)*, pages 129–137, Kauai, Hawaii, July 2003.

[17] T. Pering, T. Burd, and R. Brodersen. The simulation and evaluation of dynamic voltage scaling algorithms. In *Proc. of the International Symposium on Low Power Electronics and Design (ISLPED)*, pages 76–81, Monterey, CA, Aug. 1998.

[18] R. Raghavendra, P. Ranganathan, V. Talwar, Z. Wang, and X. Zhu. No "power" struggles: Coordinated multi-level power management for the data center. In *Proc. of the 13th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, pages 48–59, Seattle, WA, Mar. 2008.

[19] N. Tolia, Z. Wang, M. Marwah, C. Bash, P. Ranganathan, and X. Zhu. Zephyr: a unified predictive approach to improve server cooling and power efficiency. Technical Report HPL-2008-107, HP Laboratories, Palo Alto, CA, 2008.

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

[21] C. A. Waldspurger. Memory resource management in vmware esx server. In *Proc. of the 5th Symposium on Operating System Design and Implementation (OSDI 2002)*, pages 181–194, Boston, MA, 2002.