Cluster Comput
DOI 10.1007/s10586-008-0054-y

Providing platform heterogeneity-awareness for data center power management

Ripal Nathuji · Canturk Isci · Eugene Gorbatov · Karsten Schwan

Received: 29 February 2008 / Accepted: 20 March 2008

Abstract Power management is becoming an increasingly critical component of modern enterprise computing environments. The traditional drive for higher performance has influenced trends towards consolidation and higher densities, artifacts enabled by virtualization and new small form factor server blades. The resulting effect has been increased power and cooling requirements in data centers which elevate ownership costs and put more pressure on rack and enclosure densities. To address these issues, we exploit a fundamental characteristic of data centers: "platform heterogeneity". This heterogeneity stems from the architectural and management-capability variations of the underlying platforms. We define an intelligent heterogeneity-aware load management (HALM) system that leverages heterogeneity characteristics to provide two data center level benefits: (i) power efficient allocations of workloads to the best fitting platforms and (ii) improved overall performance in a power constrained environment. Our infrastructure relies upon platform and workload descriptors as well as a novel analytical prediction layer that accurately predicts workload power/performance across different platform architectures and power management capabilities. Our allocation scheme achieves on average 20% improvements in power efficiency for representative heterogeneous data center configurations, and up to 18% improvements in performance degradation when power budgeting must be performed. These results highlight the significant potential of heterogeneity-aware management.

Keywords Power management · Distributed resource management · Heterogeneous systems

R. Nathuji (✉) · K. Schwan
Georgia Institute of Technology, Atlanta, GA 30032, USA
e-mail: rnathuji@ece.gatech.edu

K. Schwan
e-mail: schwan@cc.gatech.edu

C. Isci
VMware, Palo Alto, CA 93404, USA
e-mail: canturk@vmware.com

E. Gorbatov
Intel Corporation, Hillsboro, OR 97124, USA
e-mail: eugene.gorbatov@intel.com

1 Introduction

Power management has become a critical component of modern computing systems, pervading both mobile and enterprise environments. Power consumption is a particularly significant issue in data centers, stimulating a variety of research for server systems [2]. Increased performance requirements in data centers have resulted in elevated densities enabled via consolidation and reduced server form factors. This has in turn created challenges in provisioning the necessary power and cooling capacities. For example, current data centers allocate nearly 60 Amps per rack, a limit that is likely to become prohibitive for future high density rack configurations such as blade servers, even if the accompanying cooling issues can be solved [24]. In addition, a 30,000 square feet data center with a power consumption of 10 MW requires a cooling system which costs \$2–\$5 million [21]. In such a system, the cost of running the air conditioning equipment alone can reach \$4–\$8 million a year [24]. Coupled with the elevated electricity costs from high performance servers, these effects can substantially affect the operating costs of a data center. Overall, these trends in power/cooling delivery and cost highlight the need for power and thermal management support in data centers.

Previous work on server management has focused on managing heat during thermal events [21] or utilizing platform power management support, such as processor frequency scaling, for power budgeting [9, 18, 24]. In this paper, we approach the problem of managing data centers from a different perspective by considering how to intelligently allocate workloads amongst heterogeneous platforms in a manner that (i) improves data center power-efficiency while preserving/satisfying workload performance requirements, and (ii) meets data-center-level power budgets with minimal impact on workload performance. Typically, data centers statically allocate platform resources to applications based upon peak load characteristics in order to maintain isolation and provide performance guarantees. With the continuing growth in capabilities of virtualization solutions (e.g., Xen [1] and VMware [25]), the necessity of such offline provisioning is removed. Indeed, by allowing for flexible and dynamic migration of workloads across physical resources [6], the use of virtualization in future data centers enables a new avenue of management and optimization. Our approach begins to leverage some of these capabilities to enhance power efficiency by taking advantage of the ability to assign virtualized applications to varying sets of underlying hardware platforms based upon performance needs and power consumption characteristics.

Throughout their lifetimes, data centers continually upgrade servers due to failures, capacity increases, and migrations to new form factors [12]. Over time, this leads to data centers comprised of a range of heterogeneous platforms with differences in component technologies; power, performance and thermal characteristics; and power management capabilities. When provisioning resources to workloads in these heterogeneous environments, power efficiency can vary significantly based on the particular allocation. For example, by assigning a memory bound workload to a platform that can perform dynamic voltage and frequency scaling (DVFS), run-time power consumption can be reduced with minimal impact to performance [19]. We propose a novel heterogeneity-aware load management (HALM) architecture to achieve this power-friendly behavior in data centers.

Allocating power and cooling resources is another significant challenge in the modern data center. Though clearly beneficial for transient power delivery and cooling issues, power budgeting solutions can also be useful in the provisioning of these resources. Traditionally, power and cooling have been allocated based on the nameplate rating of the system power supply or its maximum output power. However, a fully utilized server with a typical configuration will see its electrical load between 60%–75% of the name plate rating with most enterprise workloads. Therefore, providing power and cooling capacity based on these worst case assumptions results in either over allocation of power and cooling capacity or underutilization of server rack space leading to increased capital costs and underutilized data centers. Allocating power and cooling capacity based on the average workload behavior within a server and across a data center allows significantly increased densities but requires dynamic protection mechanisms that can limit server power consumption when demand temporarily exceeds available capacity. These mechanisms have been recently proposed in the literature and explored in the industry [8]. While very effective in limiting power and protecting the infrastructure, they may result in nontrivial degradation of peak performance, especially when the power constraint is too prohibitive. In this paper we illustrate how HALM can lessen the performance impact of data center power budgeting strategies.

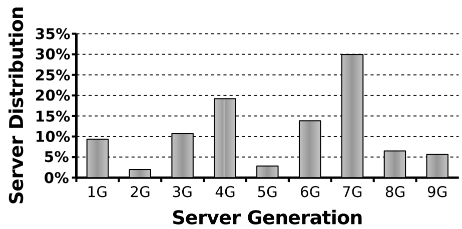
Intelligent mapping of applications to underlying platforms is dependent upon the availability of relevant information about workloads and hardware resources. As part of HALM, we extend the use of *workload* and *platform descriptors* for this purpose, which are then used by a *predictor* component that estimates the achievable performance and power savings across the different platforms in the data center. These predictions are finally used by an *allocation layer* that map workloads to a specific type of platform. This overall infrastructure is evaluated using data center configurations consisting of variations upon four distinct platforms. In summary, the main contributions of our HALM system are: (i) a platform heterogeneity-aware power management infrastructure that improves data center power efficiency under workload performance constraints and limited data center power budgets; (ii) an allocation infrastructure that uses workload and platform descriptors to perform mappings of hardware to virtualized workloads; and (iii) an intelligent load shedding policy to dynamically meet transient changes in power consumption limits. Evaluations of our system performed on state-of-the art platforms, including Intel® Core™ microarchitecture based hardware, demonstrate the benefits of exploiting platform heterogeneity for power management.

2 Motivation

2.1 Data center composition and exploiting heterogeneity

Data center deployments are inherently heterogeneous. Upgrade cycles and replacement of failed components and systems contribute to this heterogeneity. In addition, new processor and memory architectures appear every few years, and reliability requirements are becoming ever more stringent. The effect of these trends is reflected by a recent survey of data center managers that found that 90% of the facilities are expected to upgrade their compute and storage infrastructure in the next two years. Figure 1(a) shows a distribution of different systems in a representative enterprise data

« CLUS 10586 layout: Large v.1.3 reference style: mathphys file: clus54.tex (petras) aid: 54 doctopic: OriginalPaper class: spr-twocol-v1 v.2008/04/14 Prn:15/04/2008; 16:06 p. 3»



(a) Data Center Composition

	System A		System B	
	W1	W2	W1	W2
CPU Power	90W	40W	90W	20W
System Power	160W	120W	160W	120W
PSU Efficiency	86%	70%	87%	80%
Total Power	**291W**	**229W**	**287W**	**175W**

(b) Heterogeneity Management Example

Fig. 1 Data center heterogeneity and management benefits

center. As the figure shows, the data center contains nine different generations of systems that have either (1) different processor architectures, cores and frequencies; (2) varying memory capacity and interconnect speeds; or (3) different I/O capabilities. While all systems support the same software stack, they have very different and often asymmetric performance and power characteristics.

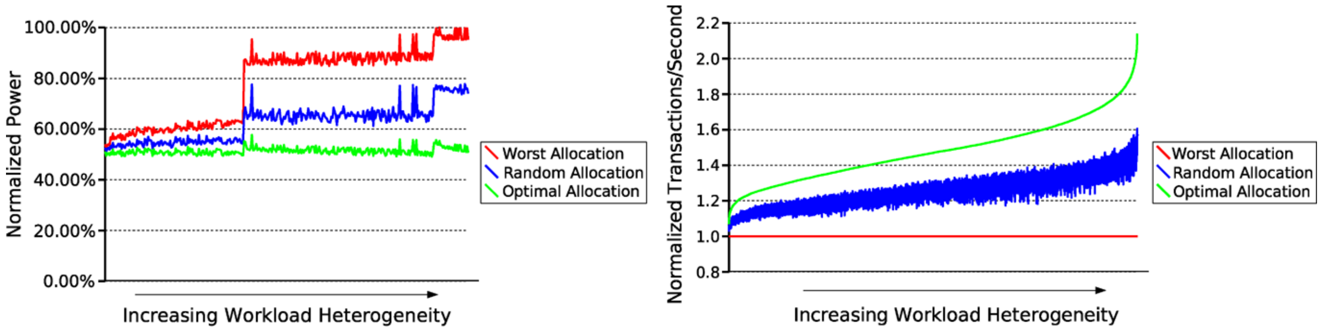
Traditionally, the non-uniformity of systems in a data center has been characterized by different levels of performance and power consumption. However, recently, another dimension has been added to this heterogeneity because server platforms are beginning to offer rich thermal and power management capabilities. Processors support DVFS and aggressive sleep states to conserve CPU power. New memory power management implementations allow different DRAM devices to go to lower power states when inactive, and enable bandwidth throttling for thermal protection. Server power supplies exhibit different conversion efficiencies under different loads, directly impacting the overall power efficiency of the system. Since power efficiency has become an important thrust in enterprise systems, we expect component and platform vendors to continue introducing new power and thermal management capabilities into their products, including I/O and system buses, chipsets, and network and disk interfaces, making future platforms even more heterogeneous.

Previous work has proposed different approaches for energy-efficient workload allocation in clusters, but none have accounted for system level power management and thermal characteristics. Therefore, the workload allocations proposed by previous approaches will yield less than ideal results since they are completely unaware of power and thermal management effects on system performance and power consumption. To illustrate this phenomenon, we experimentally compare two dual processor systems, A and B, running two different workloads, as shown in Fig. 1(b). The differences between the two systems are in the power supply unit (PSU) and processor power management capabilities. System A has a less efficient power supply at light load and has processors with limited power management support. System B, on the other hand, has a high efficiency power supply across all loads and processors that support a rich set of power management capabilities. We measure power consumption on these platforms using two different synthetic workloads: one with full utilization ($W1$) and one with a very low level of utilization ($W2$). $W1$ consumes about the same amount of power on both platforms. However, allocating the low-utilization $W2$ to system A leads to very power inefficient execution. Since A does not support power management and has low PSU efficiency at light load, its total system power is more than 50 W higher than that of system B. Thus, while both systems meet the performance demand of both workloads, heterogeneity-aware resource allocation can decrease total power by more than 10%, translating into millions of dollars in savings for large data centers. As this example shows, a full knowledge of system power and supported power management features is required to efficiently allocate workloads. Our HALM system is designed to provide such functionality.

2.2 Benefits of heterogeneity-aware management

To further motivate the need and benefits of heterogeneity-aware management in data centers, we perform two opportunity studies. The first study considers the possible benefits of allocating workloads by matching system capabilities and workload execution characteristics to reduce a data center's power profile while also meeting workload performance demands. We analyze an example of running a set of workloads in a data center configuration with four unique types of platforms described later in the paper, each with different power/performance characteristics. The set of workloads includes ten computational benchmarks (`swim`, `bzip2`, `mesa`, `gcc`, `mcf`, `art`, `applu`, `vortex`, `sixtrack`, and `lucas` from SPEC CPU2000) and one transaction-oriented workload (SPECjbb2005). We generate all subsets of four from these eleven benchmarks and compare three allocation policies for each of the subsets in Fig. 2(a). The 'worst case' allocation distributes the benchmarks across platforms to maximize power consumption, 'random' allocates workloads to platforms randomly, and 'optimal' distributes the workloads to minimize power consumption. For each workload, we allocate as many systems of a given type as necessary to meet workload throughput requirements. Subsets

« CLUS 10586 layout: Large v.1.3 reference style: mathphys file: clus54.tex (petras) aid: 54 doctopic: OriginalPaper class: spr-twocol-v1 v.2008/04/14 Prn:15/04/2008; 16:06 p. 4»



(a) Allocation Opportunity (b) Budgeting Opportunity

Fig. 2 Opportunity analysis of heterogeneity-aware management

that have benchmarks with more homogeneous behavior, i.e. similar processor and memory usage behavior, appear on the left side of the graph, while subsets with more heterogeneous benchmarks appear on the right. As can be seen from the figure, subsets of workloads with more heterogeneous behavior can substantially benefit from heterogeneity-aware resource allocation. Averaging across all subsets, the optimal policy can reduce total power by 18% when compared to random allocation and by 34% over worst-case allocation, without compromising workload performance.

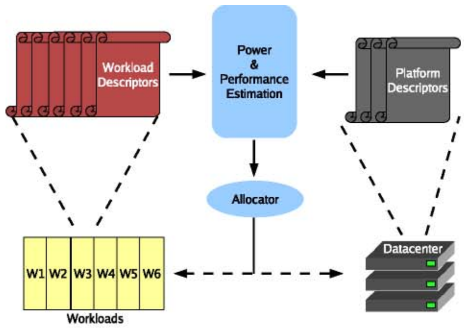
The second opportunity study considers how the aggregate throughput of a set of workloads varies within a given power budget based upon allocations. In particular, we assume that we have one of each of our four unique platforms and again generate subsets of four workloads from a set of SPEC CPU2000 benchmarks. For each subset, we calculate the minimum, average, and best case throughput across all permutations of possible allocations of the four workloads onto the four platforms. Figure 2(b) provides the results, where each scenario is normalized by the minimum throughput value to provide fair comparisons. We find that on average, the best case allocation provides a 23% improvement in performance over the random allocation, and a 48% improvement compared to the worst-case. These results highlight the relationship between allocation decisions and performance when a power budget must be imposed.

Summarizing, HALM addresses the power benefits of heterogeneity-aware allocation for two cases: (1) when there is no power budget and (2) when such a budget must be imposed temporarily due to power delivery or cooling constraints or as part of a power provisioning strategy [8].

3 Scalable enterprise and data center management

Our previous discussions have motivated the need to augment the behavior of data centers to improve manageability by leveraging the heterogeneity in platform capabilities. HALM extends this support with its heterogeneity-aware workload allocation infrastructure that utilizes the flexibility of rapidly developing virtualization technologies. Virtualization attempts to provide capabilities and abstractions that significantly impact the landscape of enterprise management. For example, there is active work to ensure performance isolation benefits, where it will be possible to run multiple virtual machines (VMs) within a given physical platform without interference among applications [15]. Currently, VMs can coexist on a platform with negligible performance interference as long as resources are not overcommitted. Approaches that allow for resource pools and reservations as well as dynamic resource sharing and reclamation can aid in providing isolation even when systems are over-provisioned. Secondly, by encapsulating application state within well defined virtual machines, migration of workloads among resources can be performed easily and efficiently. A more powerful contribution of virtualization, however, is the ability to combine multiple resources across physical boundaries to create virtual platforms for applications, providing a *scalable enterprise* environment. HALM assumes the existence of this flexible and powerful virtualization support.

The usage pattern of data centers is becoming increasingly service-oriented, where applications and workloads may be submitted dynamically by subscribers/clients. When managing these types of applications certain management actions, such as allocation decisions, happen at a coarse granularity with finer adjustments being made at runtime to address transient issues such as reduced power budgets. One can imagine how such a data center might be managed with the typically used assignment approaches. At some infrequent interval the pool of applications and service level agreements (SLAs) that specify their required performance, in metrics such as throughput or response time, are compiled. Applications are then assigned to platforms using a simple load balancing scheme based upon utilization or queue lengths, possibly even accounting for differences in the performance of the systems [26], so that SLAs are met.



(a) System Components

```
Component: Processor
  Architecture:     Netburst
  No_Cores:         4
  Frequency:        3.2 GHz
  p-state_support:  Yes

Component: Memory
  Cache_size:   2MB
  Memory_type:  DDR2
  Memory_size:  8GB
.
.
.
```

```
Attribute: MPI
  val1: <Memory.Cache_size=2MB>
  val2: <Memory.Cache_size=4MB>

Attribute: CPI_CORE
  val1: <Processor.Architecture=
        NetBurst>
  val1: <Processor.Architecture=
        Core Microarchitecture>
.
.
.
```

(b) Platform and Workload Descriptor Examples

Fig. 3 HALM architecture

When load must be reduced to address power budgeting requirements, load might be shed from workloads in a similarly random or round robin fashion. This approach clearly leaves room for improvement, since it does not consider power or platform differences in any way. HALM addresses this weakness by performing heterogeneity aware allocations as well as intelligent load shedding.

The HALM architecture can be organized into three major components: (1) platform/workload descriptors, (2) a power/performance predictor, and (3) an allocator, as shown in Fig. 3(a). We use platform and workload descriptors to provide our workload allocator with the differences amongst workloads and platforms. These descriptor inputs are utilized by the predictor to determine: (1) the relative performance of workloads on different types of platforms, and (2) the power savings achievable from platform power management mechanisms. Coupled with coarse platform power consumption information (obtained via online power monitoring) (3) the allocator, performs the assignments of workloads to the available resources.

The purpose of platform descriptors is to convey information regarding the hardware and power management capabilities of a machine. A platform descriptor is made up of multiple modules, representing different system components, as shown in Fig. 3(b). Each module specifies the type of component to which it refers, such as processor, memory subsystem, or power supply. Within each of these modules, various component parameters are defined. For example, a module describing the processor component may have attributes like its microarchitecture family, frequency, and available management support. Workload descriptors are also structured in modules, headed with attribute declarations. Within each module, a list of values for that attribute is provided. As workload attributes often vary with the platforms on which it executes, our descriptor design allows multiple attribute definitions, where each definition is predicated with component parameter values that correlate back to platform descriptors. Figure 3(b) illustrates the structure of the resulting workload descriptor. We further explain the meaning of the MPI (memory accesses per instruction) and $\mathrm{CPI}_{\mathrm{CORE}}$ (core cycles per instruction) attributes in subsequent sections.

Platform descriptor information can be provided in a variety of ways. It can be made readily available using platform support such as ACPI [13], and possibly also with some administrative input. To provide the required workload descriptors, we profile workloads on a minimal set of *orthogonal platforms*, with mutually exclusive component types. We then use an analytical prediction approach to project workload characteristics on all available platforms. As we discuss in Sect. 5, this approach provides accurate predictions that scale with increased amounts of heterogeneity.

4 Methodology

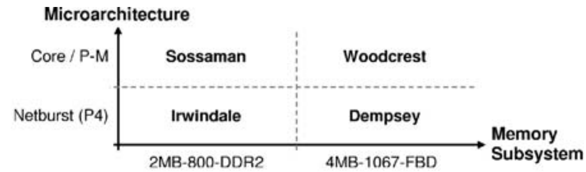
4.1 Platform hardware

Our hardware setup consists of four types of rack mounted server platforms summarized in Fig. 4(a), where LLC denotes last-level cache size. All four types of platforms contain standard components and typical configurations that entered production cycles. In our experiments Linux was installed on all systems for measurement of various attributes (e.g. CPI, MPI, etc.) as well as performance. We validated that the performance results matched those with Xen using a subset of workloads and platforms, but performed the majority of our experiments in a non-virtualized environment to have better access to performance counters used to measure other workload attributes.

The platform names are based on their processor code name in this paper. All four platforms are dual-processor systems. Woodcrest, Sossaman, and Dempsey are CMP dual-core processors, and Irwindale is a 2-way SMT processor supporting Hyper-Threading Technology. All platforms have 8 GB of memory. Woodcrest and Dempsey support Fully Buffered DIMM (FBD) memory with a 533 MHz DDR2 bus, while Sossaman and Irwindale support unregistered DDR2 400 MHz memory. Woodcrest and Dempsey

	Woodcrest	Sossaman	Dempsey	Irwindale
Processor	3GHz/4MB LLC Core architecture	2GHz/2MB LLC Pentium® M	3.7GHz/4MB LLC NetBurst/P4	3.8GHz/2MB LLC NetBurst/P4
FSB	1067 MHz Dual FSB	800 MHz	1067 MHz Dual FSB	800 MHz
Chipset	Blackford	Lindenhurst	Blackford	Lindenhurst
Memory	DDR2-533 FBD	DDR2-400	DDR2-533 FBD	DDR2-400

(a) Platform Characteristics



(b) Heterogeneity Quadrants

Fig. 4 Experimental platforms

Table 1 Levels of heterogeneity in our experimental platforms

	Across-Platforms				Within-Platform	DPM-Capability		Heterogeneous Configuration
	Microarchitecture		Memory		Frequency [GHz]	Enabled	Disabled	
	Core	Netburst	FBDIMM	DDR-2				
Woodcrest	X		X		3.0	x		1
							x	2
					2.6	x		3
							x	4
					2.3	x		5
							x	6
					2.0	x		7
							x	8
Sossaman	X			X	2.0	x		9
							x	10
					1.6	x		11
							x	12
					1.3	x		13
							x	14
					1.0	x		15
							x	16
Dempsey		X	X		3.7	x		17
							x	18
					3.2	x		19
							x	20
Irwindale		X		X	3.8	x		21
							x	22
					3.2	x		23
							x	24
					2.8	x		25
							x	26

have dual FSB architectures with two branches to memory and two channels per branch.

All four types of systems are heterogeneous in a sense that each has a unique combination of processor architecture and memory subsystem. If we assume that Intel Core microarchitecture/Pentium® M and NetBurst constitute two types of processors and LLC-4 MB/FSB-1066/FBD-533 and LLC-2 MB/FSB-800/DDR2-400 constitute two types of memory, all four platforms can be mapped as having unique processor/memory architecture combinations. Note that all four platforms also have vastly different power and performance characteristics. For example, the Intel Core microarchitecture is superior to NetBurst both in terms of performance and power efficiency. FBD based memory, on the other hand, provides higher throughput in our systems at the expense of elevated power consumption due to increased DDR2 bus speed and the power requirements of the Advanced Memory Buffer (AMB) on the buffered DIMMs. The four platforms occupy separate quadrants of a heterogeneity space with dimensions of microarchitecture heterogeneity and memory subsystem heterogeneity, as shown in Fig. 4(b). We refer to this initial level of heterogeneity as "*across-platform heterogeneity*". However, in addition to this, all these server platforms also support chip-level DVFS.

This leads to a second degree of heterogeneity, where one type of platform can have instances in a data center that are configured to operate at different frequencies. We refer to this as "*within-platform heterogeneity*". As process variations increasingly result in the *binning* of produced chips into different operating points, this within-platform heterogeneity becomes an inherent property of the general data center landscape. Finally, many of these platforms may incorporate some processor dynamic power management (DPM) techniques that adaptively alter platform behavior at runtime. This creates a third source of heterogeneity, "*DPM-capability heterogeneity*", where platforms with built-in DPM hooks exhibit different power/performance characteristics from the ones with no DPM capabilities. In Table 1, we show how these three levels of heterogeneity quickly escalate the number of distinct platform configurations in a data center scenario.

All experimental power measurements are performed using the Extech 380801 power analyzer. The power is measured at the wall and represents total AC power consumption of the entire system. The power numbers presented in this paper are obtained by averaging the instantaneous system power consumption over the entire run of each workload. Our assumption is that infrastructure support for monitoring power consumption will be utilized to obtain this type of workload specific power characteristics online, instead of parameterized models. For example, all power supplies, which adhere to the latest power supply monitoring interface (PSMI) specification, support out-of-band current/voltage sampling allowing for per platform A/C power monitoring reflected by our actual power measurements.

4.2 Application model

When power managing computing environments, improvements can be attained with a variety of approaches. In this work, we consider two scenarios. The first assumes a lack of budgeting constraints, concentrating on a workload allocation that reduces power consumption while maintaining baseline application performance. In other words, we maximize the performance per watt, while holding performance constant. The second addresses power budgeting by performing load shedding to reduce power consumption while minimizing performance impact to workloads. We consider application performance in terms of throughput, or the rate at which transaction operations are performed. Therefore, it is not the execution time of each transaction that defines performance, but the rate at which multiple transactions can be sustained. This type of model is representative of applications such as transaction based web services or payroll systems.

The goal of HALM is to evaluate the power-efficiency tradeoffs of assigning a workload to a variety of platforms. Since the performance capabilities of each platform are different, the execution time to perform a single operable unit, or atomic transaction, varies across them. As previously mentioned, virtualization technologies can help to extend the physical resources dedicated to applications when necessary to maintain performance by increasing the number of platforms used to execute transactions. In particular, transactions can be distributed amongst nodes until the desired throughput is reached.

For our analysis, we consider applications that mimic the high performance computational applications common to data center environments and also heavily exercise the power hungry components of server platforms, the processor and memory. Two aspects of these workloads are captured in our experimental analysis. First, these workloads are inherently transactional, such as the previous financial payroll example or the processing of risk analysis models across different inputs common to investment banking. Second, with the ability to incorporate large amounts of memory into platforms at relatively low costs, these applications often execute mostly from memory, with little or no I/O being performed. Though I/O such as network use can play a significant role in multi-tier enterprise applications, we leave consideration of such characteristics to future work. To realize our application model, while also providing deterministic and repeatable behavior for our experimentation, we utilize benchmarks from the SPEC CPU2000 suite as representative examples of transaction instances. SPEC benchmarks allow for the isolation of processor and memory components, while also generating different memory loads. Indeed, many SPEC benchmarks exhibit significant measured memory bandwidth of 5–8 GB/sec on our systems. In order to provide an unbiased workload set, we include all SPEC benchmarks in our experiments. For each application, we specify an SLA in terms of required transaction processing rate, equal to the throughput achievable on the Woodcrest platform.

5 Workload behavior estimation

The power/performance predictor component of our HALM framework can be implemented in multiple ways. For example, one can profile a set of microbenchmarks on all platform configurations and develop statistical mapping functions across these configurations. However, as the platform types and heterogeneity increase, the overhead of such approaches can be prohibitive. Instead, we develop a predictor that relies on the architectural platform properties and adjusts its predictions based on the heterogeneity specifications. We refer to this model as the "*Blocking Factor* (*BF*) *Model*". The BF model simply decomposes execution cycles into *CPU cycles* and *memory cycles*. CPU cycles represent

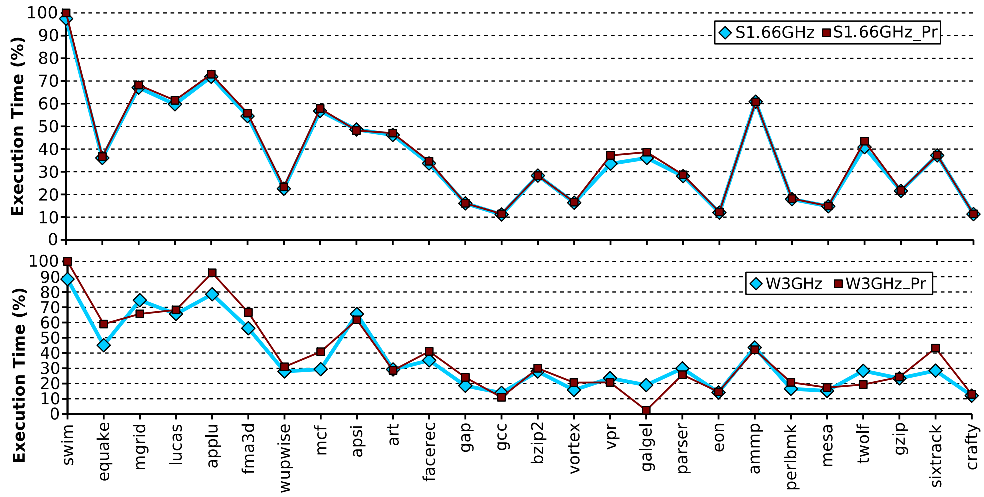


Fig. 5 Performance prediction results

the execution with a perfect last-level cache (LLC), while memory cycles capture the finite cache effects. This model is similar to the "*overlap model*" described by Chou et al. [5]. With the BF model, the CPI (cycles per instruction) of a workload can be represented as in (1). Here $\mathrm{CPI}_{\mathrm{CORE}}$ represents the CPI with a perfect LLC. This term is independent from the underlying memory subsystem. $\mathrm{CPI}_{\mathrm{MEM}}$ accounts for the additional cycles spent for memory accesses with a finite-sized cache:

$$\mathrm{CPI} = \mathrm{CPI}_{\mathrm{CORE}} + \mathrm{CPI}_{\mathrm{MEM}}. \tag{1}$$

The $\mathrm{CPI}_{\mathrm{MEM}}$ term can be expanded into architecture and workload specific characteristics. Based on this, the CPI of a platform at a specific frequency f_1 can be expressed as in (2). Here, *MPI* is the memory accesses per instruction, which is dependent on the workload and the LLC size, *L* is the average memory latency, which varies based upon the memory subsystem specifications, and *BF* is the *blocking factor* that accounts for the overlapping concurrent execution during memory accesses, which is a characteristic of the workload:

$$\mathrm{CPI}(f_1) = \mathrm{CPI}_{\mathrm{CORE}}(f_1) + \mathrm{MPI} \cdot L(f_1) \cdot \mathrm{BF}(f_1). \tag{2}$$

Using variants of (2), performance prediction can be performed relatively easily for within-platform heterogeneity, as well as across-platform heterogeneity. For within-platform heterogeneity, the frequency-dependent components of (2) are scaled with frequency to predict workload performance on a different frequency setting. The top chart in Fig. 5 provides results for an example of this type of prediction with an orthogonal platform (Sossaman). The figure contains the actual measured performance for our workloads together with the predicted performance.

In the latter case of across-platform heterogeneity, the natural decoupling of the microarchitectural and memory subsystem differences in the BF model enables us to estimate application performance on a platform lying on a different corner of the memory and microarchitecture heterogeneity space. Among our four experimental platforms, two "*orthogonal platforms*", which span two opposite corners of the platform heterogeneity quadrants in Fig. 4(b), can be used to predict performance on a third "*derived platform*". The lower chart in Fig. 5 shows the prediction results for the Woodcrest platform, whose performance is "*derived*" using the $\mathrm{CPI}_{\mathrm{CORE}}$ and $\mathrm{CPI}_{\mathrm{MEM}}$ characteristics of the orthogonal platforms (Sossaman and Dempsey respectively). Overall, for the orthogonal platforms, the BF model can very accurately predict performance with an average prediction error of 2%. For the derived platforms, our predictor can track actual execution times very well, though with an increased average prediction error of 20%. In the following sections, we show that this performance prediction methodology provides sufficient accuracy to represent workload behavior and allows HALM to achieve close to optimal allocations. Further details of this prediction methodology can be found in our previous work [22].

The final heterogeneity type supported by our predictor is the DPM-capability heterogeneity. For this, we consider a platform that enables DVFS during memory bound execution regions of an application. We implement this functionality as part of OS power management, based on prior work [14]. To incorporate DPM awareness, we extend the predictor component to estimate the potential power savings that can be attained when executing a workload on a DPM enabled platform. Experimental results show that there is a strong correlation between the MPI of a workload and its power saving potential. Therefore, we utilize the MPI attribute in the workload descriptors to predict the power saving potentials of workloads on DPM enabled platforms. Figure 6 shows that our MPI based prediction approach effectively captures the power saving potentials of different workloads and successfully differentiates applications that

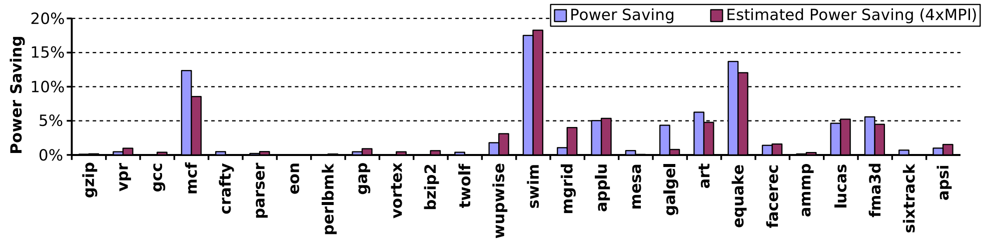


Fig. 6 Power saving predictions for DPM enabled platforms

can benefit significantly from being allocated to a DPM enabled machine. As we describe in Sect. 6.1, we use this predictor to choose workloads that should be assigned to the DPM enabled platforms.

6 Management policies

6.1 HALM allocation policy

After processing the workload and platform descriptors, and utilizing our BF model for performance prediction, the next step is to perform allocation of resources to a set of applications in a data center. Evaluations are based on a greedy policy for allocating workloads. In particular, with each application i, we associate a cost metric for executing on each platform type k. Workloads are then ordered into a scheduling queue based on their maximum cost metric across all platform types. The allocator then performs application assignment based on this queue, where applications with higher worst-case costs have priority. The platform type chosen for an application is a function of this cost metric across the available platforms as well as the estimated DPM benefits. As a cost metric for our policy, we define $N_{i,k}$, the number of platforms of type k required to execute a workload i, $N_{i,k}$. This value is clearly a function of both the performance capabilities of the platform and the SLA requirement of the workload. $N_{i,k}$ can be analytically defined, given the transaction based application model utilized in our work. For each application i, the service level agreement (SLA) specifies that X_i transactions should be performed every Y_i time units. If $t_{i,k}$ is the execution time of a transaction of application i on platform k, the resulting number of platforms required to achieve the SLA can be expressed with (3)

$$N_{i,k} = \left\lceil \frac{X_i}{\lfloor Y_i / t_{i,k} \rfloor} \right\rceil. \tag{3}$$

The $t_{i,k}$ values are provided by the performance predictor. It should be noted that there is a discretization in $N_{i,k}$, which is due to the fact that individual atomic transactions cannot be parallelized across multiple platforms. $N_{i,k}$ is therefore better able to handle errors due to the inherent discretization performed, making it a strong choice as a cost metric (other possible metrics are discussed and defined in our previous work [22]). Given the use of $N_{i,k}$ as our cost metric, our allocation approach first determines the platform types of which (1) there are enough available systems to allocate the workload and (2) the cost metric is minimized. We then use DPM savings to determine whether a more power efficient platform alternative should be used between those with the same cost value. In other words, if there are multiple platform types for which an application has the same $N_{i,k}$ value, we utilize a DPM specific threshold to decide whether or not it should be scheduled to a DPM enabled platform type. As we demonstrate in the following section, this threshold based approach can be effective in identifying workloads that can take advantage of DPM capabilities.

6.2 HALM power budgeting policy

In order to address transient power delivery or cooling issues, it is sometimes necessary to temporarily reduce power consumption in a data center. To provide this mechanism, we develop a load shedding policy based upon an existing workload allocation scheme. The goal of the policy is to reduce the amount of resources provided to applications in order to meet a power budget while still allowing all workloads to make some progress. In other words, application performance may be degraded compared to prescribed SLAs, but all applications achieve some fraction of their SLA.

Our power budgeting policy is, again, a greedy approach. For all applications with resources that can be shed, i.e. applications that utilize more than one platform, we define a power-efficiency metric as the throughput per Watt that is being provided by each resource. Afterwards, the resources with minimal power efficiency are shed until the power budget is met. As our experimental results demonstrate, this simple metric allows for improved performance when power budgeting must be performed, as well as better fairness across workloads in terms of performance degradation experienced.

7 Experimental evaluation

7.1 Increasing power efficiency

In order to evaluate our heterogeneity-aware allocation approach, we perform power and performance measurements of our SPEC based representative transactional workloads across each type of platform. To scale these results to the number of platforms present in data centers, this measured data is extrapolated analytically using a data center allocation simulator which combines real power and performance data, prediction output, and allocation policy definitions to calculate power efficiency in various data center configurations. In the simulator, we provide the output of the predictor as input to the allocation policy. We always assume that the platforms which are profiled are the 2 GHz Sossaman platform and the 3.7 GHz Dempsey system. Since we assume the workload attributes are profiled accurately on these systems, for fairness we also assume that for these two platforms performance data is obtained via profiling as well and is therefore known perfectly. We then consider three different scenarios: (1) all other platform performance information is known perfectly (oracle) (2) our BF model is used to predict performance for the remainder of platforms as described in Sect. 5 (BF model) (3) incorporating a simple statistical regression approach (Stat. Est.). For this regression method, we profile a subset of applications across all platforms to obtain linear performance prediction models parameterized by variables that can be obtained by profiling a workload on the 2 GHz Sossaman and 3.7 GHz Dempsey systems (CPI, MPI, etc.). The regression models can then be used to predict performance of any application. The baseline allocation scheme we compare against is a random one, since it closely estimates the common round-robin or utilization based approach.

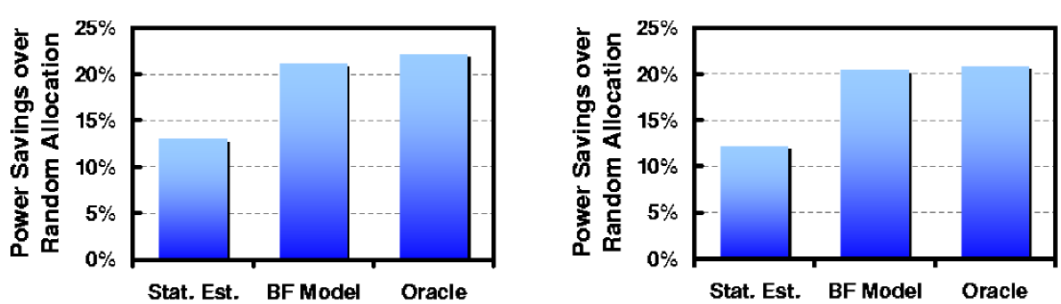
The efficiency improvements achievable in a data center are also dependent upon the system's current mix of applications. To obtain our results, we randomly pick applications and allocate them using the random approach until no more workloads can be scheduled. Using the resulting set of workloads, we then evaluate power consumption when using our prediction and allocation policies, and compare them against the random allocation result. This is repeated a hundred times for each of our data points.

We first look at the benefits achieved with HALM in data center configurations with across-platform and within-platform heterogeneity but no DPM support. In particular, we include the four base platforms, Woodcrest, Sossaman, Dempsey, and Irwindale, as well as the frequency variations of the platforms. We create data center configurations with equal numbers of each type of system. Trends are consistent across various data center sizes, so for brevity, we include here only results with 1000 platforms of each type. The resulting data center has 13 types of platforms, and power consumptions vary with allocation as shown in Fig. 7(a). The first interesting observation is that platform heterogeneity allows us to achieve improved benefits over a simple random approach. Indeed, we see improvements of 22% with perfect knowledge and 21% using our BF based prediction compared to a random allocation policy. We also observe a significant difference between the statistical and analytical prediction schemes. The regression approach is unable to scale in terms of accuracy with increased heterogeneity, whereas the BF approach achieves close to optimal power savings.

In order to evaluate how well our allocator can exploit DPM support, we extend the thirteen platform type configuration with an additional Woodcrest 3 GHz platform which provides DPM support. We again find that our BF prediction method can provide improved aggregate savings across all machines over the statistical approach as shown in Fig. 7(b). To more closely determine our ability to exploit DPM mechanisms, we also evaluate the power consumption of the thousand DPM-enabled platforms (all of which are active). We find that our BF model based allocation is able to improve the power efficiency of these platforms by 3.3%. This illustrates the potential of HALM to provide additional benefits when platforms vary in the power management they support.

7.2 Maximizing performance under power budgets

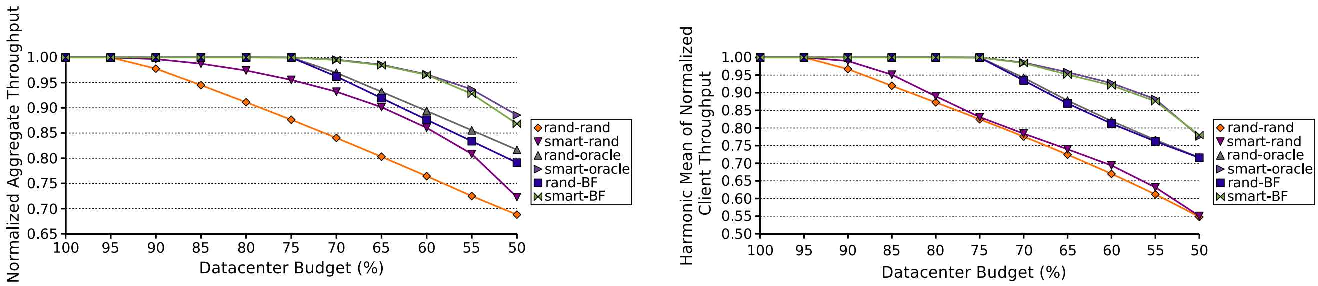
As a second benefit of HALM, we evaluate the ability to perform load shedding when power budgeting must be performed. In particular, our goal is to maximize performance obtained in the data center while observing power budgets. For the purposes of this paper, we consider transient power



(a) Thirteen Heterogeneous Platforms

(b) DPM-Aware Power Savings

Fig. 7 HALM power improvements



(a) Throughput Results (b) Client Fairness

Fig. 8 Power budgeting results

budgeting where workloads are not (re-)migrated, but instead, based upon an existing allocation, resources are temporarily withheld from applications and placed into low power states. For comparison, we consider three such initial allocations from the previous section, one based upon a random allocation, one based upon perfect oracle performance information, and finally an allocation based upon prediction with our BF approach. For each allocation, we consider two load shedding policies, one which randomly selects an application and sheds resources (i.e. a single compute node) if possible, and our greedy "smart" policy described in Sect. 6.2.

Figure 8(a) provides the aggregate performance across all workloads in the data center for different power budgets. The performance results are normalized with respect to the maximum achievable performance at the maximum unconstrained power budget (100%). The figure shows how the performance of different allocation policies decrease with decreasing power budgets. The range of applied power budgets is limited to 50% as there is no possible allocation that meets the budget below this point. The six scenarios considered are random shedding based upon a random allocation (rand-rand), our load shedding policy based upon a random allocation (smart-rand), similarly the two shedding approaches based upon oracle based allocation (rand-oracle and smart-oracle), and finally, shedding based upon a BF prediction based allocation (rand-BF and smart-BF). We see multiple interesting trends. First, the intrinsic benefits of the heterogeneity-aware allocations towards budgeting are apparent in the figure by the fact that performance does not begin to reduce until lower power budgets compared to a random allocation scheme. We also see that given a particular allocation, our shedding policy provides benefits in performance across the set of power budgets. Moreover, again, we find that our BF prediction model behaves very close to an oracle based allocation when our load shedding policy is used. Overall, we see benefits of up to 18% in performance degradation compared to a random load shedding policy based upon a random allocation. Figure 8(b) evaluates the performance degradations for the six approaches by also taking the fairness of load shedding into account. Here, we show the aggregate data center performance as the harmonic mean of the individual workload throughputs, which is a commonly used metric for evaluating fairness [20]. We see from the figure that a random allocation always exhibits poor performance regardless of the load shedding policy used. A heterogeneity-aware allocation, on the other hand, provides improved fairness, particularly when combined with our load shedding policy.

8 Related work

HALM builds upon existing work and extends the state of the art in power management research. A variety of mechanisms exist to provide power and thermal management support within a single platform. Brooks and Martonosi proposed mechanisms for the enforcement of thermal thresholds on the processor [3], focusing on single platform characteristics as opposed to the data center level management achieved with HALM. Processor frequency and voltage scaling based upon memory access behavior has been shown to successfully provide power savings with minimal impact to applications. Resulting solutions include hardware based approaches [19] and OS-level techniques, that set processor modes based on predicted application behavior [14]. HALM is designed to manage load while being aware of such underlying power management occurring in server platforms. Power budgeting of SMP systems with a performance loss minimization objective has also been implemented via CPU throttling [16]. Other budgeting solutions extend platform support for fine grain server power limiting [18]. The power budgeting achieved with HALM is based upon the use of resource allocation to reduce power consumption across multiple systems, as opposed to throttling performance of individual components.

At the data center level, incorporating temperature-awareness into workload placement has been proposed by Moore et al. [21], along with emulation environments for studies of thermal implications of power management [11]. HALM can use these thermal aware strategies to perform

« CLUS 10586 layout: Large v.1.3 reference style: mathphys file: clus54.tex (petras) aid: 54 doctopic: OriginalPaper class: spr-twocol-v1 v.2008/04/14 Prn:15/04/2008; 16:06 p. 12»

power budgeting based upon data center temperature characteristics. Chase et al. discuss how to reduce power consumption in data centers by turning servers on and off based on demand [4]. Utilizing this type of cluster reconfiguration in conjunction with DVFS [7] and the use of spare servers [23] has been investigated as well. As opposed to these approaches, HALM attempts to reduce power consumption by intelligently managing workloads across heterogeneous servers. Enforcing power budgets within data centers by allocating power in a non-uniform manner across nodes has been shown to be an effective management technique [9]. Techniques for enforcing power budgets at blade enclosure granularities have also been discussed [24]. HALM budgets aggregate power consumption via resource allocation without assigning per server power budgets as with these previous approaches.

Heterogeneity has been considered to some degree in prior work, including the evaluation of heterogeneous multi-core architectures with different core complexities [17]. In comparison, HALM considers platform level heterogeneity as opposed to processor asymmetry. In cluster environments, a scheduling approach for power control has been proposed for processors with varying fixed frequencies and voltages [10]. HALM supports heterogeneity across additional dimensions, such as power management capabilities and memory. A power efficient web server with intelligent request distribution in heterogeneous clusters is another example which considers leveraging heterogeneity in enterprise systems [12]. HALM goes beyond these methods, by considering not just the differences in platforms' performance capabilities, but also in their power management capabilities.

9 Conclusions and future work

Power management in data center environments has become an important area of research, in part because power delivery and cooling limitations are quickly becoming a bottleneck in the provisioning of performance required by increasingly demanding applications. This paper makes use of the management flexibility afforded by virtualization solutions to develop a heterogeneity-aware load management (HALM) system. HALM improves power management capabilities by exploiting the natural heterogeneity of platforms in data centers, including differences in dynamic power management support that may be available. We introduce a three phase approach to mapping workloads to underlying resources to improve power efficiency consisting of structured platform and workload descriptors, a prediction component to estimate the performance and power characteristics of various workload to platform mappings, and finally an allocator which utilizes policies and prediction results to perform decisions. We also evaluate a load shedding policy based upon resulting allocations to improve performance when power budgeting must be performed.

Our results underscore two major conclusions. First, we show that by intelligently considering the varying power management capabilities of platforms, the ability for these systems to obtain power savings using their management mechanisms can be vastly improved when compared to other assignment models. Using representative data center configurations consisting of older P4 based platforms up to Intel Core microarchitecture based systems, we find that our allocation architecture can improve power efficiency by 20% on average. In addition, our results show that by performing intelligent load shedding when power budgets must be observed, 18% improvements in performance degradation can be obtained when using HALM.

In this paper, we present the beginning of our investigation into exploiting platform heterogeneity and emerging virtualization support to improve the power characteristics of enterprise computing environments. As future work, we plan to integrate the management tradeoffs and lessons learned from this work into virtualization layer management applications. This includes the consideration of distributed virtualized workloads such as tiered web services where different components may be appropriate for each layer, including heterogeneous I/O devices. The results presented in this paper support the potential of this area of research for power managing heterogeneous computing systems.

References

1. Barham, P., Dragovic, B., Fraser, K., Hand, S., Harris, T., Ho, A., Neugebauer, R., Pratt, I., Warfield, A.: Xen and the art of virtualization. In: Proceedings of the ACM Symposium on Operating Systems Principles (SOSP), 2003
2. Bianchini, R., Rajamony, R.: Power and energy management for server systems. IEEE Comput. **37**(11), 68–76 (2004)
3. Brooks, D., Martonosi, M.: Dynamic thermal management for high-performance microprocessors. In: Proceedings of the 7th International Symposium on High-Performance Computer Architecture (HPCA), January 2001
4. Chase, J., Anderson, D., Thakar, P., Vahdat, A., Doyle, R.: Managing energy and server resources in hosting centers. In: Proceedings of the 18th Symposium on Operating Systems Principles (SOSP), October 2001
5. Chou, Y., Fahs, B., Abraham, S.: Microarchitecture optimizations for exploiting memory-level parallelism. In: Proceedings of the International Symposium on Computer Architecture (ISCA), June 2004
6. Clark, C., Fraser, K., Hand, S., Hansen, J.G., Jul, E., Limpach, C., Pratt, I., Warfield, A.: Live migration of virtual machines. In: Proceedings of the 2nd ACM/USENIX Symposium on Networked Systems Design and Implementation (NSDI), May 2005
7. Elnozahy, E.N., Kistler, M., Rajamony, R.: Energy-efficient server clusters. In: Proceedings of the Workshop on Power-Aware Computing Systems, February 2002
8. Fan, X., Weber, W.-D., Barroso, L.: Power provisioning for a warehouse-sized computer. In: Proceedings of the International Symposium on Computer Architecture (ISCA), June 2007

« CLUS 10586 layout: Large v.1.3 reference style: mathphys file: clus54.tex (petras) aid: 54 doctopic: OriginalPaper class: spr-twocol-v1 v.2008/04/14 Prn:15/04/2008; 16:06 p. 13»

9. Femal, M., Freeh, V.: Boosting data center performance through non-uniform power allocation. In: Proceedings of the Second International Conference on Autonomic Computing (ICAC), 2005
10. Ghiasi, S., Keller, T., Rawson, F.: Scheduling for heterogeneous processors in server systems. In: Proceedings of the International Conference on Computing Frontiers, 2005
11. Heath, T., Centeno, A.P., George, P., Ramos, L., Jaluria, Y., Bianchini, R.: Mercury and freon: Temperature emulation and management in server systems. In: Proceedings of the International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), October 2006
12. Heath, T., Diniz, B., Carrera, E.V., Meira, W. Jr., Bianchini, R.: Energy conservation in heterogeneous server clusters. In: Proceedings of the 10th Symposium on Principles and Practice of Parallel Programming (PPoPP), 2005
13. Hewlett-Packard, Intel, Microsoft, Phoenix, and Toshiba: Advanced configuration and power interface specification. http://www.acpi.info (2004)
14. Isci, C., Contreras, G., Martonosi, M.: Live, runtime phase monitoring and prediction on real systems with application to dynamic power management. In: Proceedings of the 39th International Symposium on Microarchitecture (MICRO-39), December 2006
15. Koh, Y., Knauerhase, R., Brett, P., Bowman, M., Wen, Z., Pu, C.: An analysis of performance interference effects in virtual environments. In: Proceedings of the IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), 2007
16. Kotla, R., Ghiasi, S., Keller, T., Rawson, F.: Scheduling processor voltage and frequency in server and cluster systems. In: Proceedings of the Workshop on High-Performance, Power-Aware Computing (HP-PAC), 2005
17. Kumar, R., Tullsen, D., Ranganathan, P., Jouppi, N., Farkas, K.: Single-Isa heterogeneous multi-core architectures for multithreaded workload performance. In: Proceedings of the International Symposium on Computer Architecture (ISCA), June 2004
18. Lefurgy, C., Wang, X., Ware, M.: Server-level power control. In: Proceedings of the IEEE International Conference on Autonomic Computing (ICAC), June 2007
19. Li, H., Cher, C., Vijaykumar, T., Roy, K.: Vsv: L2-miss-driven variable supply-voltage scaling for low power. In: Proceedings of the IEEE International Symposium on Microarchitecture (MICRO-36), 2003
20. Luo, K., Gummaraju, J., Franklin, M.: Balancing throughput and fairness in SMT processors. In: Proceedings of the IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), November 2001
21. Moore, J., Chase, J., Ranganathan, P., Sharma, R.: Making scheduling cool: Temperature-aware workload placement in data centers. In: Proceedings of USENIX '05, June 2005
22. Nathuji, R., Isci, C., Gorbatov, E.: Exploiting platform heterogeneity for power efficient data centers. In: Proceedings of the IEEE International Conference on Autonomic Computing (ICAC), June 2007
23. Rajamani, K., Lefurgy, C.: 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), March 2003
24. Ranganathan, P., Leech, P., Irwin, D., Chase, J.: Ensemble-level power management for dense blade servers. In: Proceedings of the International Symposium on Computer Architecture (ISCA), 2006
25. Sugerman, J., Venkitachalam, G., Lim, B.-H.: Virtualizing i/o devices on VMware workstation's hosted virtual machine monitor. In: Proceedings of the USENIX Annual Technical Conference, 2001
26. Zhang, W.: Linux virtual server for scalable network services. In: Ottawa Linux Symposium, 2000

Ripal Nathuji is a Ph.D. candidate in the School of Electrical and Computer Engineering at the Georgia Institute of Technology. He previously received his M.S. in Computer Engineering from Texas A&M University and his B.S. in Electrical Engineering and Computer Science from the Massachusetts Institute of Technology. His current research focuses on system-level power management of computing systems, with applications to virtualized enterprise server environments.

Canturk Isci is a senior member of technical staff at VMware. His research interests include power-aware computing systems and workload-adaptive dynamic management. He has a Ph.D. and an M.A. in electrical engineering from Princeton University, an M.S. in VLSI system design from University of Westminster, London, UK, and a B.S. in electrical engineering from Bilkent University, Ankara, Turkey.

Eugene Gorbatov is a researcher at Intel's Energy Efficient Systems lab. His current research focuses on platform power management, particularly the interaction of different platform components and system software.

Karsten Schwan is a professor in the College of Computing at the Georgia Institute of Technology, and is also the Director of the Center for Experimental Research in Computer Systems (CERCS). He obtained his M.S. and Ph.D. degrees from Carnegie-Mellon University in Pittsburgh, Pennsylvania, where he began his research in high performance computing, addressing operating and programming systems support for the Cm* multiprocessor. His current work ranges from topics in operating and communication systems, to middleware, to parallel and distributed applications, focusing on information-intensive distributed applications in the enterprise domain and in the high performance domain.