Energy-Aware Accounting and Billing in Large-Scale Computing Facilities

Víctor Jiménez
Roberto Gioiosa
Barcelona Supercomputing Center

Francisco J. Cazorla
Spanish National Research Council

Mateo Valero
Polytechnic University of Catalonia

Eren Kursun
Canturk Isci
Alper Buyuktosunoglu
Pradip Bose
IBM Thomas J. Watson Research Center

Proposals have focused on reducing energy requirements for large-scale computing facilities (LSCFs), but little research has addressed the need for energy-usage-based accounting. Energy-aware accounting and billing benefits LSCF owners and users. This article makes a case for accurate cost accounting and billing, which accounts for user-specific energy usage, and identifies the hardware- and software-level changes necessary to support energy-aware accounting.

Energy and power trends in large-scale computing facilities (LSCFs) pose challenges that shape the design of next-generation facilities. The carbon footprint of societal energy consumption levels has seen intense scrutiny in recent years. According to recent statistics, the electricity demand from LSCFs shows the fastest growth among all sectors, and facilities consume several megawatts, enough to power small towns.[1] The US Environmental Protection Agency estimates that national energy consumption attributable to servers and data centers will soon reach more than 100 billion kW hours annually,[2] and recent studies estimate the corresponding electrical cost to be US$30 billion.[3]

The cost of energy is rising, further exacerbating the problem. Recent studies show that power accounts for 13 percent of the total cost of ownership (TCO) in an LSCF. This cost increases up to 31 percent if we add the cost for cooling and power infrastructure, becoming the second largest contributor to the TCO, behind server costs.[4] Additionally, while server cost has remained flat over successive generations, energy cost is expected to rise,[5] increasing the relative cost of energy.

Despite these energy consumption trends, user- or task-specific accounting for energy or power consumption is limited. The accounting method applied for user-level billing is usually based simply on the amount of time that a resource is used. However, this method typically doesn't consider the exact level of resource usage; power consumption attributable to a specific user job is either estimated on the basis of known peak (or nameplate) values for used resources or a derated estimation for the actual peak power consumption that the system can achieve under a realistic workload.[6] (In fact, several tools from server vendors let their

Published by the IEEE Computer Society

Related Work in Large-Scale Computing Facilities

Large-scale computing facilities are vast infrastructures with high operation costs. Any possible optimization that improves their efficiency can translate into a considerable cost reduction. Several proposals focus on improving data centers' energy efficiency.[1-3] Many of these proposals advocate for energy-proportional systems, in which the need for energy accounting is higher than in current systems.

Several of these works focus on either reducing power consumption when the system is idle or improving efficiency by consolidating more virtual machines in the same hardware. To this end, some proposals leverage workload heterogeneity to better schedule the workloads onto the computing resources, thus increasing resource usage.[1] Nathuji and Schwan proposed a mechanism to connect the low-power mechanisms available in the hardware with the power management requests and hints made by an operating system running within a virtual machine.[4]

Accounting users, tasks, and virtual machines for the energy they actually consume is orthogonal to the aforementioned proposals. On one hand, the potential to adapt to the workloads' heterogeneity increases with per-task energy accounting. On the other hand, energy accounting brings benefits itself, as we detail in the main article.

Kansal et al. presented initial steps for an accurate energy-accounting mechanism.[5] Their goal was to develop a better power-capping mechanism in the presence of multiple virtual machines on one node. However, more research is necessary to obtain a more accurate mechanism for use not only for power consumption estimation, but for billing users according to their energy consumption as well. For instance, their proposal uses simple ways to split static power consumption and power consumption caused by virtual machine interferences, among virtual machines. To obtain better accuracy, hardware and operating system support is necessary.

Bertran et al. presented an energy-accounting system for small-sized systems.[6] However, our work focuses on large-scale computing facilities, where other solutions are likely needed.

References

1. J. Karidis and J.E. Moreira, ''The Case for Full-Throttle Computing: An Alternative Datacenter Design Strategy,'' *IEEE Micro,* vol. 30, no. 4, 2010, pp. 25-28.
2. J. Karidis, J.E. Moreira, and J. Moreno, ''True Value: Assessing and Optimizing the Cost of Computing at the Data Center Level,'' *Proc. 6th ACM Conf. Computing Frontiers,* ACM Press, 2009, pp. 185-192.
3. D. Meisner, B.T. Gold, and T.F. Wenisch, ''PowerNap: Eliminating Server Idle Power,'' *ACM SIGPLAN Notices,* vol. 44, no. 3, 2009, pp. 205-216.
4. R. Nathuji and K. Schwan, ''VirtualPower: Coordinated Power Management in Virtualized Enterprise Systems,'' *SIGOPS Operating Systems Rev.,* vol. 41, no. 6, 2007, pp. 265-278.
5. A. Kansal et al., ''Virtual Machine Power Metering and Provisioning,'' *Proc. 1st ACM Symp. Cloud Computing,* ACM Press, 2010, pp. 39-50.
6. R. Bertran et al., ''Accurate Energy Accounting for Shared Virtualized Environments using PMC-based Power Modeling Techniques,'' *IEEE/ACM Int'l Conf. Grid Computing,* IEEE Press, 2010, doi:10.1016/j.future.2011.03.007.

customers estimate the expected maximum consumption in a much more accurate way than using nameplate values.) Nonetheless, this is a rough estimation typically based on average or worst-case behavior. Thus, using a more accurate method, such as energy-aware accounting, is beneficial for LSCF owners.

Although accounting based just on usage time and resource type and size is adequate in the present context, where static power dominates the total power consumption in current hardware, there's a clear movement toward energy-proportional systems.[7] In such systems, most of the energy an application consumes—and hence, its cost—is due to its activity. In this scenario, current accounting systems can be neither accurate nor fair. For instance, two customers can incur different utilizations across similarly allocated resources, and yet result in nearly identical usage time. Moreover, the facility owner's cost could vary significantly because of differences in power and energy consumption.

In this article, we highlight the importance and benefits of using energy-aware accounting and billing on current LSCFs such as data centers. We explore the opportunities, as well as the problems in implementing such technology. We leave detailed, specific solutions to these problems for future research work. Instead, we propose an accounting and billing method that would benefit the typical consumer in terms of (generally) reduced expenses, based on accurate measurements of actual resource usage levels. Additionally, we show that the facility owner's adoption of such accounting metrics would drive up energy efficiency in computing facilities. (For more information on other proposals, see the ''Related Work in Large-Scale Computing Facilities'' sidebar.)

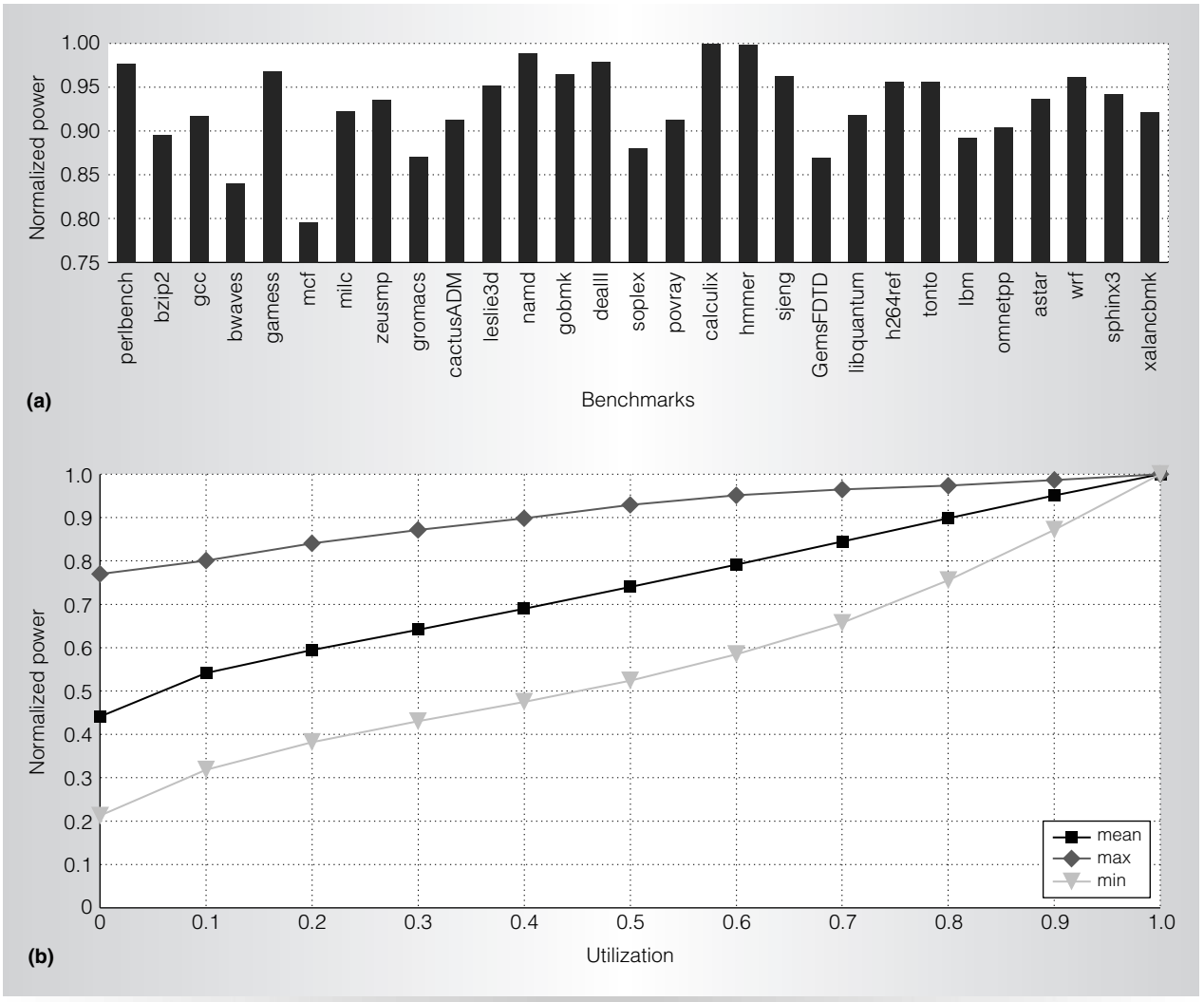

Figure 1. Power consumption for SPEC CPU2006 benchmarks measured on an Intel quad-core system (a) and for the available results for SPECpower at several CPU utilization levels (b). Max and min refer to the most- and least-consuming systems, respectively. Mean is the average for all submitted results.

Motivation examples

To elaborate on the need for accurate, energy-aware accounting principles, we consider several benchmarks as proxies for the behavior of applications executed by different users on a small system. Figure 1a shows the results for executing all the SPEC CPU2006 benchmarks on an Intel quad-core, single-socket server system. A 10-percent variation in power across workloads is typical, with the maximum variation being 20 percent (between `mcf` and `calculix`). So, `mcf`-like and `calculix`-like workloads executing for the same length of time on the same platform would incur energy usage levels that actually differ by a margin of 20 percent; yet, current accounting and billing practices would treat them equally. Figure 1b shows the power consumption at different usage levels for all the submitted SPECpower[8] results between 2007 and 2010 available at the SPEC website. This example illustrates variable-demand workloads, showing considerably different power consumption for CPU usage levels.

These variations are already significant and will probably increase in the future, when system vendors build more energy-efficient and energy-proportional systems. As idle consumption levels drive down, and

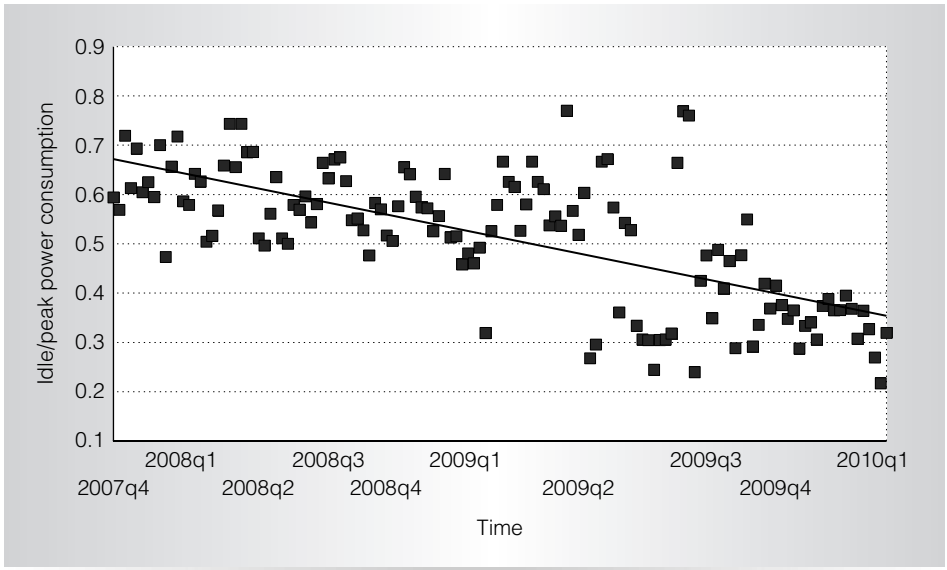

Figure 2. Comparison of idle and peak power consumption for SPECpower submitted results. The data shows a clear trend to reduce the significance of idle power consumption.

peak system power remains constant (or perhaps even higher), the variation in usage-driven power profiles across different workloads is bound to increase in the future. In fact, multiple ongoing initiatives are trying to reduce the significance of the static consumption fraction. Techniques implemented in current processors, such as dynamic voltage and frequency scaling (DVFS) and sleep modes with different depth levels, reduce static energy consumption. Yet for many hardware components, a high fraction of their power consumption is static regardless of their activity.

Although current systems aren't yet energy-proportional, the trend is moving toward this kind of system. Figure 2 shows the ratio of idle power consumption over peak consumption for all the SPECpower results submitted between 2007 and 2010. The data is sorted by submission date and shows a clear trend to reduce the idle power consumption's significance, and thus move toward energy-proportional systems. In the presence of truly energy-proportional systems, the static power cost would be almost entirely eliminated, and the dynamic cost would account for most of the energy consumption. Under this situation, all the energy that the systems consume will be a consequence of application activity; thus, considering energy consumption for accounting purposes becomes attractive.

Benefits of energy-aware accounting and billing

From both the user's and LSCF owner's perspectives, there are significant potential benefits from using energy-aware accounting and billing.

User's perspective

The first benefit for users would be more accurate and fair billing. Consider the consequences of current billing practices on the user community. Figure 3 shows the normalized power consumption as a function of usage for one system submitted to the SPECpower webpage. Under current accounting practices, if the user instance executes for T hours, the billing would effectively be based at the peak power rate, (P_{peak}), where usage is 1 (see Figure 3). Thus, the user's bill would be

$$bill_{conv} = K \times P_{peak} \times \mathrm{T} \quad (1)$$

where K is a constant value (measured in dollars per power unit per hour).

If the energy accounting were done accurately, the user would be charged depending on the average resource utilization. For example, in Figure 3, we observe that if the

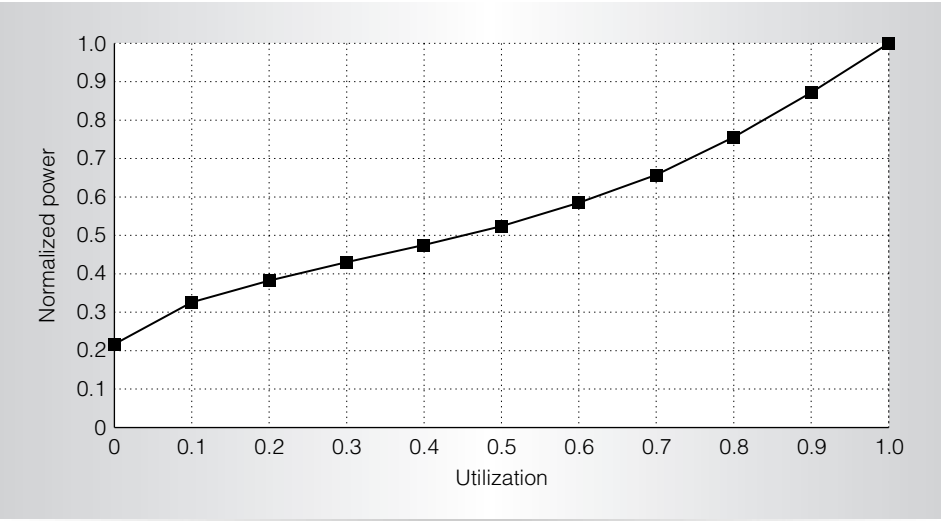

Figure 3. Power consumption as a function of usage for a system submitted to the SPECpower webpage. The values are normalized to the consumption when the system is fully utilized. The actual system is a Fujitsu Primergy TX150 S7 server, based on a quad-core Intel Xeon X3470 with 4 Gbytes of RAM. Its maximum power consumption is 112 W when utilization is 100 percent.

average CPU usage was recorded to be 40 percent (a study at Google revealed that most of the servers typically operate at 10- to 50-percent utilization[7]), the power consumption would decrease by slightly more than 50 percent and the fair bill would have been:

$$bill_{fair} = K \times P_{40\%} \times \mathrm{T} \approx 0.5 \times bill_{conv} \quad (2)$$

Energy consumption isn't the only cost for LSCFs; personnel, capital cost, and maintenance represent a significant part of the TCO. However, the cost for power plus cooling and power distribution accounts for up to 31 percent of the TCO.[4] Therefore, a 50-percent reduction in energy cost translates into a 16-percent reduction for the user's bill.

Energy-aware billing enables other end-user benefits. For instance, current facilities don't expose power consumption to their users. Exposing power consumption per task or virtual machine would let users understand their applications' power and energy profile and their power consumption versus execution time trade-off. Thus, users could optimize their applications and deployment configurations to reduce their bill. This green trend also benefits the data center owner and society in general.

Our approach shouldn't require users to have too-advanced computer science skills to exploit energy accounting. We envision a runtime system that will help users select proper setups for their applications and the underlying hardware to reduce energy. Energy accounting, on the other hand, could make users uncertain about the billing they'll receive, because it depends on the actual energy the applications use. The facility owner can provide bounds or estimates on the energy that users' applications will consume using profiling (for example, using Figure 3).

Owner's perspective

There are several reasons why a facility owner should invest in accurate, energy-aware accounting.

Finer-grained precision in allocating and managing cooling resources. Today's LSCFs design the cooling infrastructure so it can effectively dissipate the heat produced by the systems, under worst-case load scenarios, either based on the sum of nameplate powers across all the facility resources or on a derated estimation of the actual peak power consumption under realistic workloads. However, as we mentioned earlier, servers are typically underused. Therefore,

facilities might consider the possibility of reducing cooling costs by underprovisioning the cooling resources, based on typical or observed peak workloads in the facility. However, heuristically fixing thermal thresholds could lead to frequent need of engaging performance-throttling mechanisms or to tripping the fuses, producing unplanned server outages. Precise energy-accounting practices would result in better runtime task allocation and cooling resource allocation to prevent unplanned outages or performance shortfalls.

Safe workload consolidation. In prior non-virtualized systems, once a user instance received some physical resources, no other user would be able to share those resources. In such a situation, time is indeed money; so, even if the user instance isn't using the allocated resources, it would make sense to charge the user a flat, per-hour rental rate, because once a set of resources is tied up, the owner can't make rental income out of those resources from any other waiting customer.

With the advent of virtualized hardware, the owner can make money from multiple customers sharing a resource. The net resource usage could then approach 100 percent, a good business proposition. In this new scenario, the owner has no reason not to move to an energy-aware accounting system based on actual resource usage; because the total usage across all users approaches 100 percent, the net effect is that the total bill amount across multiple users sharing the same system basically follows Equation 1 again, with the total revenue approaching $K \times P_{peak} \times T$.

A built-in energy-accounting system could guide the workload management system to make scheduling decisions that result in safe, more efficient workload consolidation. For example, let's assume that a system can run N virtual machines simultaneously. When selecting a subset of virtual machines for execution, it's hard to determine whether the power or energy threshold would be exceeded, so the virtual machine manager must be conservative. With per-virtual machine energy accounting, at the time of composing a workload of N virtual machines, we know the power consumption of each virtual machine and thus the workload. Therefore, energy accounting improves efficiency, and we can consolidate more virtual machines simultaneously. By doing so, we can add more computing nodes and service more customers with the same power budget—a clear benefit for the data center owner.

Reduction in energy costs. Motivation for end users to reduce their energy consumption (and bill) will also drive down the total energy consumption incurred by the data center. In a context where energy costs are a significant part of the TCO, energy is becoming scarce—and, owing to difficulties associated with electric power transmission, the data center owner will welcome any reduction in power consumption.

Target facilities

Several types of facilities exist, each with a different business model. Energy accounting targets multiple facility types, though its potential benefits depend on their characteristics. We consider two major types of facilities:

- *Dedicated hosting services and colocation facilities.* In this case, the facility follows a dedicated provisioning in which some physical nodes (or the slots to place them) are leased to a given user. A user's application can span multiple nodes, and the overall provisioned capacity is dedicated to the deployed applications. In this model, the leased nodes' overall operation and power cost can be attributed to the running applications. Only a per-node energy accounting is needed. Although supercomputers aren't data centers, for accounting purposes we can accommodate them in this category.
- *Virtual private servers and cloud hosting.* Adoption of this type of facility is growing (for example, Amazon EC2), and we envision clear benefits from energy accounting in these types of facilities. In these facilities, the owner bills users according to the number of hours their instances (that is, virtual machines) are on, not considering the

detailed compute resource usage profile. Although some parameters—such as data transfer, I/O, and disk space—are used for billing purposes, two instances running for the same number of hours will be charged the same, in terms of wall-clock CPU time, regardless of what the actual CPU and memory usage is.

We can make several considerations when applying energy accounting in virtualized data centers. First, resource providers such as Amazon EC2 provision end users with virtual resources. Here, the direct mapping of the end user applications to actual physical resources isn't transparently known. Moreover, because the applications aren't directly mapped to physical hardware, direct hardware profiling isn't generally available at the application level. Instead, the virtual-machine manager has direct access to the hardware profiling and knows when applications are really mapped to hardware. This layer is an appropriate level for implementing the energy-accounting approach.

Second, virtualization vendors further provide additional resource management vehicles such as resource guarantees, limits, and shares. In this case, each application's and virtual machine's contribution to energy consumption depends on provisioned virtual resources, the imposed resource constraints, and the underlying resource-sharing mechanism. All these management vehicles are orthogonal to energy accounting. For instance, some applications handle asynchronous events and have hard latency requirements. To deal with this situation, the application or user must reserve resources in advance. From the energy-accounting point of view, this just implies that the user must pay the reserved resources' static power consumption. Once the user's application starts running, it follows our proposed energy-accounting policy.

Finally, many virtualization technologies also employ additional resource optimizations such as page sharing across compatible virtual machines, linked clones with shared based images, memory overcommitment, and dynamic memory ballooning. These techniques, while improving overall resource use efficiency, also blur the resource and energy usage association with individual applications and end users. Several changes are required at the software and hardware level to adopt energy accounting.

Energy-accounting design and trade-offs

There are challenges and opportunities associated with energy and power accounting at various granularities in a large-scale computing environment. Some changes are also required, both at the hardware and software levels, to provide accurate energy accounting. The infrastructure required to accurately track peak power and energy dissipation can vary significantly over the computing spectrum. However, several common considerations apply to all systems.

Granularity versus overhead

A critical point in an energy-accounting system is to decide the level at which energy is tracked. The hardware and software overhead increases for per-user rack and node-level accounting. Within a node, the accounting becomes even more challenging. At the hardware level, we must decide the area, power, and cost overhead of the additional hardware blocks to provide accurate accounting. At the software level, we must decide how much overhead time we'll allow for tracking energy consumption.

Fairness

From the user's perspective, an important principle to follow is that different runs of the same application with the same input exhibit a similar energy profile. This is called the principle of accounting, and it's currently applied to CPU time accounting.[9,10] In an ideal scenario, the application reaches the same energy-accounting result for the same input, regardless of the applications it's coscheduled with. However, in reality, several factors complicate the ideal case, potentially causing significant variation for repeated runs. Accurate, fair energy-aware accounting and billing should account for this.

Power versus time trade-off

More computing resources imply less execution time and higher power but reduce static consumption's significance.

Energy-aware accounting and billing lets the user (likely in collaboration with the data center owner) find the best design point for the client applications. Fine-grained energy accounting lets us find the best power and time trade-off. Service-level agreements complicate this process, making it more difficult to perform optimizations by adjusting performance-related parameters. However, for cases that are targeted to fill the unused data center capacity (as in Amazon EC2 Spot Instances), the agreement's flexibility could provide potential for such optimization.

Static and dynamic power consumption

To accurately track energy consumption, we must first break down power-related costs between static and dynamic costs. The former accounts for the power that doesn't depend on the system activity (for example, the power consumption of an idle machine that is not running any user process). The latter is related to the extra power consumed when there's user activity on the system. The fraction between static and dynamic power depends on both the system under consideration and the workload itself.

For the dedicated data center case, where users don't share nodes, that distinction isn't really necessary, because the total power consumption can be typically measured at the node level. (If some external resources are shared, some of the following discussion might apply to the accounting for these resources.) However, for virtualized data centers, we must estimate the fraction of these components that must be attributed to each virtual machine running on the system.

Static power. Splitting the cost of static power consumption among virtual machines depends on the level at which resources are shared, leading to several possibilities with different associated accuracies and overheads. The easiest solution is to split the static consumption among all the virtual machines mapped to that node either evenly among them all or proportional to each virtual machine's dynamic power consumption.[11] If we want higher accuracy, we can individually look at the system's subcomponents. However, we need either hardware support to derive the static power consumption or the hardware vendor to provide these values. Current performance-monitoring counters aren't enough to derive, for example, the static power consumption of a system's individual subcomponents.

We differentiate two subcomponent types on the basis of their nature.

- *Spatial sharing.* In spatially shared subcomponents (such as cache or memory), there's a linear relation between the amount of space a virtual machine demands and the cost of static power. If in a given instant a resource with an associated space of M_{total} bits has a static power consumption of S_{total} watts, it can be broken down among N virtual machines as follows: $S_i = (M_i/M_{total}) \times S_{total}$, in which $\sum_{i=1}^{N} M_i = M_{total}$ and $\sum_{i=1}^{N} S_i = S_{total}$, where M_i and S_i are the amount of space used and the static consumption incurred by virtual machine i, respectively.
- *Temporal sharing.* Temporally shared components (such as the CPU or hard drive) consume static power proportionally to the duration they're enabled. In this case, we can use an *interval-based accounting* approach. Let's assume we divide the time into intervals of fixed length I. If during a given interval a certain amount of virtual machines access the device, all of its static power consumption is charged to these virtual machines. The other running virtual machines shouldn't be charged, because we assume that the subcomponents can go into a low-power mode if they're not accessed for an interval I. Thus, the static energy consumption for virtual machine i during time interval k, when Nk virtual machines are accessing the device, is $S_{i,k} = S_k/N_k$, where S_k is the static energy consumed by a device during interval k. It follows that the static power charged to virtual machine i after N intervals is $\sum_{k=1}^{N} S_{i,k}$.

We can find subcomponents that, depending on their power-saving capabilities,

present both spatial and temporal sharing characteristics. In that case, we can apply a hybrid combination of the methodology we discussed in the previous paragraph.

Dynamic power. Splitting the dynamic power consumption among virtual machines is a complex task that in some cases might require hardware or software support. We can use several approaches for attributing energy consumption to multiple virtual machines sharing a node. CPU usage is a high-level metric that typically correlates well with power consumption, and thus, energy consumption.[7,11] Its main advantage is that it's easy to collect, thereby reducing the complexity and the overhead for energy-accounting implementation.

Additionally, if we want a higher accuracy level, we can estimate energy consumption on the basis of lower-level metrics, such as events in the system. We can use different sources to collect events: performance counters such as instructions per cycle (IPC) and cache misses, and operating system statistics such as I/O operations. Multiple studies demonstrate the high correlation between system events and power consumption,[12] with generally higher associated overheads compared to high-level metrics.

The type of metrics required to estimate power consumption also depends on the workloads being executed within the virtual machines. For instance, in CPU-intensive workloads, high-level generic metrics generally are less useful. CPU usage for these kinds of workloads is mostly close to 100 percent, rendering CPU usage-based power estimations inapplicable. Despite this fact, as Figure 1a shows, significant power consumption variation exists among workloads running at 100-percent CPU usage. We can use workload-specific, high-level metrics, but this solution isn't portable among different workloads, and it might not be easy to make that metric visible from outside the virtual machine. Therefore, in the case of CPU-intensive workloads, event-based metrics are a much better fit to accurately estimate energy consumption.

Application interference and system activity

In shared environments, there's generally interference among virtual machines accessing the same hardware resources. Nowadays, most facilities use processors that can concurrently execute more than one thread (based on chip-level multiprocessing [CMP], simultaneous multithreading [SMT], or a hybrid approach). In these systems, two different virtual machines share certain resources when they're executed at the same time. Although program output won't change, the actions that the system takes to obtain this output could differ compared to when a virtual machine is executed in isolation. For instance, the aggregated memory footprint of both virtual machines can exceed the amount of cache or memory installed in the system, leading to memory or disk accesses that wouldn't occur if the virtual machines ran in isolation. Luque et al. show that the interaction between multiple applications running on a CMP can lead to errors in CPU time accounting up to 19 percent.[9,10] Including hardware support tracking intra- and inter-task interferences can reduce the error to 1 percent. Including similar mechanisms based on tracking per-thread subcomponent usage would make energy-aware accounting more precise.

Another source of interference is system activity caused by housekeeping (for example, freeing virtual memory and cleaning system logs). Finally, optimizations across virtual machines create interactions among them as well. The challenge here is to determine how to account for the energy that the system consumes considering such interference. Current solutions, such as Kansal et al.,[11] don't focus on these issues because hardware and operating system support would be necessary to increase the accuracy of their energy-accounting proposal.

Hardware and software support for energy accounting

As we've shown, several shortcomings exist in obtaining accurate energy accounting with low overhead. However, new hardware support could overcome some of these problems. First, some current systems already let us obtain power measurements at the processor level. A standard, accurate

way to obtain similar measurements for a system's most consuming subcomponents can greatly enhance the accuracy of energy accounting.

Second, an easier way to derive power consumption from performance counters is desirable. Kadayif et al. presented a framework based on performance counters to obtain energy measurements.[13] However, a native hardware implementation will probably prove more accurate. For instance, the IBM Power7 processor internally uses a power proxy based on more than 50 architectural events to estimate the power consumption for each core.[14,15]

Third, although we can use performance counters as a power proxy, other possibilities are required, because we can't use current performance counters to derive the static power consumption of certain devices (such as the memory). For instance, including hardware support to obtain the instruction mix per thread can already significantly increase the accuracy on power consumption estimation.

Fourth, as we mentioned earlier, hardware support to overcome application interference can also help improve the accuracy of energy accounting.

Software support can improve energy accounting's accuracy as well. For example, the operating system or the virtual machine manager can help by tracking the time that resources are being used by the operating system itself, without contributing to a direct profit for the user. Also, interaction between the accounting system and the virtual machine monitor can help track energy usage in the presence of virtual machine optimizations such as those we described earlier.

The complexity of implementing an energy-aware accounting mechanism depends on the facility and application characteristics. Although the case of dedicated systems is considerably simple, shared environments face multiple research challenges, which will constitute future research. Interaction among the different system layers (hardware, hypervisor, and software) is necessary to obtain accurate accounting systems.

We argue for the importance of continuing the trend toward energy-proportional systems. In fact, energy-aware accounting will benefit from this trend and, at the same time, can accelerate it, as demand for greener computing grows. MICRO

Acknowledgments

This work was supported by a collaboration agreement between IBM and BSC with funds from IBM Research and IBM Deep Computing. It was also supported by the Ministry of Science and Technology of Spain under contracts TIN-2007-60625 and JCI-2008-3688, as well as the HiPEAC Network of Excellence (ICT-217068). The authors thank the anonymous reviewers for their constructive comments and suggestions.

References

1. C. Belady and C. Malone, ''Data Center Power Projections to 2014,'' *IEEE Intersociety Conf. Thermal and Thermomechanical Phenomena in Electronics Systems,* IEEE Press, 2006, pp. 439-444.
2. *EPA Report to Congress on Server and Data Center Energy Efficiency,* tech. report, US Environmental Protection Agency, 2007.
3. R. Raghavendra et al., ''No 'Power' Struggles: Coordinated Multilevel Power Management for the Data Center,'' *ACM SIGOPS Operating Systems Review,* vol. 42, no. 2, 2008, pp. 48-59.
4. J. Hamilton, ''Overall Data Center Costs,'' blog, 18 Sept. 2010; http://perspectives.mvdirona.com/2010/09/18/OverallDataCenterCosts.aspx.
5. L.A. Barroso, ''The Price of Performance,'' *Queue: Multiprocessors,* vol. 3, no. 7, 2005, pp. 48-53.
6. ''Proper Sizing of IT Power and Cooling Loads,'' white paper, Green Grid, 2009.
7. L.A. Barroso and U. Hölzle, ''The Case for Energy-Proportional Computing,'' *Computer,* vol. 40, no. 12, 2007, pp. 33-37.
8. K.-D. Lange, ''Identifying Shades of Green: The SPECpower Benchmarks,'' *Computer,* vol. 42, no. 3, 2009, pp. 95-97.
9. C. Luque et al., ''ITCA: Intertask Conflict-Aware CPU Accounting for CMPs,'' *Proc. 2009 18th Int'l Conf. Parallel Architectures*

and Compilation Techniques, IEEE CS Press, 2009, pp. 203-213.
10. C. Luque et al., ''CPU Accounting in CMP Processors,'' *IEEE Computer Architecture Letters,* vol. 8, no. 1, 2009, pp. 17-20.
11. A. Kansal et al., ''Virtual Machine Power Metering and Provisioning,'' *Proc. 1st ACM Symp. Cloud Computing,* ACM Press, 2010, pp. 39-50.
12. W. Bircher and L. John, ''Complete System Power Estimation: A Trickle-Down Approach Based on Performance Events,'' *IEEE Int'l Symp. Performance Analysis of Systems & Software,* IEEE Press, 2007, pp. 158-168.
13. I. Kadayif et al., ''vEC: Virtual Energy Counters,'' *Proc. 2001 ACM SIGPLAN-SIGSOFT Workshop Program Analysis for Software Tools and Eng.,* ACM Press, 2001, pp. 28-31.
14. M. Floyd et al., ''Adaptive Energy-Management Features of the IBM POWER7 Chip,'' *IBM J. Research and Development,* vol. 55, no. 3, 2011, pp. 8:1-8:18.
15. M. Floyd et al., ''Introducing the Adaptive Energy Management Features of the POWER7 chip,'' *IEEE Micro,* vol. 31, no. 2, 2011, pp. 60-75.

Víctor Jiménez is a doctoral candidate in the Computer Architecture Department at the Polytechnic University of Catalonia, Spain, and a resident student at the Barcelona Supercomputing Center. His research interests include performance- and power-efficient systems, operating systems, and parallel architectures. Jiménez has an MSc in computer science from the Polytechnic University of Catalonia.

Roberto Gioiosa is a senior researcher in the group on operating system/computer architecture interaction at the Barcelona Supercomputing Center. His research interests include operating systems, high-performance computing, real-time systems, large parallel clusters, embedded systems, network protocols, processors, and network cards. Gioiosa has a PhD in computer science from the University of Rome Tor Vergata.

Francisco J. Cazorla is a researcher in the Spanish National Research Council. He leads the group on operating system/computer architecture interaction at the Barcelona Supercomputing Center. His research interests include multithreaded architectures for both high-performance and real-time systems. Cazorla has a PhD from the Polytechnic University of Catalonia.

Mateo Valero is a professor at the Polytechnic University of Catalonia and director of the Barcelona Supercomputer Center. His research interests include high-performance architectures. Valero has a PhD in telecommunications from the Polytechnic University of Catalonia. He's an IEEE Fellow, an Intel Distinguished Research Fellow, and an ACM Fellow.

Eren Kursun is a research staff member in the Computer Architecture Department at the IBM Thomas J. Watson Research Center. Her research interests include power and temperature management of microprocessor architectures and low-power design. Kursun has a PhD in computer science from the University of California, Los Angeles. She's a member of IEEE and the ACM.

Canturk Isci is a research staff member in the Distributed Systems Department at the IBM Thomas J. Watson Research Center. His research interests include virtualization, data center energy management, and microarchitectural and system-level techniques for workload-adaptive and energy-efficient computing. Isci has a PhD in computer engineering from Princeton University.

Alper Buyuktosunoglu is a research staff member in the Reliability and Power-Aware Microarchitecture Department at the IBM Thomas J. Watson Research Center. His work has included support for IBM p-series and z-series microprocessors in the areas of power-aware computer architectures, dynamic power management, and high-level power modeling. Buyuktosunoglu has a PhD in electrical and computer engineering from the University of Rochester. He's a senior member of IEEE and an editorial board member for *IEEE Micro*.

Pradip Bose is a research staff member and manager of the Reliability and Power-Aware

Microarchitecture Department at the IBM Thomas J. Watson Research Center. His work has included presilicon modeling and definition of IBM Power-series microarchitectures, beginning with the research precursor of the first RS/6000 product. Bose has a PhD in electrical and computer engineering from the University of Illinois at Urbana-Champaign. He's a fellow of IEEE and an advisory board member of *IEEE Micro*.

Direct questions or comments about this article to Víctor Jiménez, Barcelona Supercomputing Center, Nexus II Building c/ Jordi Girona, 29, 08034 Barcelona, Spain; victor.javier@bsc.es.