

A Comparative Study of the Effectiveness of CPU Consolidation Versus Dynamic Voltage and Frequency Scaling in a Virtualized Multicore Server

Inkwon Hwang, *Student Member, IEEE*, and Massoud Pedram, *Fellow, IEEE*

Abstract—Companies operating large datacenters are focusing on how to reduce the electrical energy costs of operating datacenters. A common way of cost reduction is to perform a dynamic voltage and frequency scaling (DVFS), thereby matching the CPU's performance and power level to incoming workloads. Another power saving technique is CPU consolidation, which uses the minimum number of CPUs necessary to meet the service request demands and turns OFF the remaining unused CPUs. DVFS has been already extensively studied and verified its effectiveness. On the other hand, it is necessary to study more about the effectiveness of CPU consolidation. Key questions that must be answered are how effectively the CPU consolidation improves the energy efficiency and how to maximize the improvement. These questions are addressed in this paper. After understanding modern power management techniques and developing an appropriate power model, this paper provides an extensive set of hardware-based experimental results and makes suggestions about how to maximize energy efficiency improvement through CPU consolidation. In addition, this paper also presents new online CPU consolidation algorithms, which reduce the energy-delay product up to 13% compared with the Linux default DVFS algorithm.

Index Terms—Algorithm, consolidation, energy efficiency, virtualization.

I. INTRODUCTION

DATACENTERS consist of a very large number of server machines, and the number of servers has been rapidly increasing. Although the energy efficiency of server machines has been improving, this efficiency advances have not kept pace with the increase in the number and size of data centers. As a result, an ever increasing amount of electrical energy is being consumed in today's data centers, giving rise to concerns about the carbon emission footprint of data centers and the costs of operating them. The latter is especially important concern from the viewpoint of datacenter owners and operators (as well as their customers who must eventually pay the bill).

Manuscript received February 2, 2015; revised June 4, 2015 and August 31, 2015; accepted October 14, 2015. Date of publication December 8, 2015; date of current version May 20, 2016. This work was supported by Semiconductor Research Corporation under Grant 2012-HJ-2292.

The authors are with the Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089 USA (e-mail: inkwonhw@usc.edu; pedram@usc.edu).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TVLSI.2015.2499601

Two widely employed techniques for reducing the costs of operating data centers are server consolidation and DVFS. The former aims at minimizing the number of ON servers in a datacenter by consolidating all the incoming jobs into as few servers as possible whereas the latter attempts to match the performance of each ON server to the current workloads so that energy can be saved at the workload level of each server. Server consolidation is needed and complements DVFS technique because of the energy non-proportional behavior of modern servers [1]. This is expected because an electronic circuit consumes static power (leakage in CMOS digital circuits) regardless of whether it provides any computational services. The issue is, however, worse than simple leakage and has to do with the fact that many components within a modern server system cannot be scaled/modulated to exhibit a linear relationship between their power consumption and delivered performance levels.

A data center is typically under-utilized; it has been designed to provide the required performance and satisfy its service level agreements (SLAs) with clients even during peak workload hours, and hence, at other times its resources are vastly under-utilized. For example, the minimum and the maximum utilization of the statically provisioned capacity of Facebook's data center are 40% and 90%, respectively [2]. Hence, in light of the energy non-proportionality of today's server base, a greater amount of energy costs can be reduced by the server consolidation. The server consolidation has been studied very well, and many studies have suggested the use of virtual machine (VM) migration as a means of doing server consolidation [3]–[7].

Although server consolidations can greatly lower a data center's total energy consumption, there is still room for further energy savings due to the limitations and overheads associated with the server consolidation. For one, it is difficult to conduct server consolidation very frequently because the migration of tasks or VMs causes high overheads; e.g., heavy network traffic, high network latency, and additional energy consumption. Because of these overheads, there is a relatively long period between server consolidation decision times. To avoid the SLA violations during each timing period when VM to server assignments are fixed, VMs (or tasks) are not too tightly consolidated into the active server set. For the

longer the period between migrations more server machines should be activated, each at a lower average utilization rate. Hence, server machines are still under-utilized after server consolidation, which implies the potential of further energy savings through additional resource management techniques.

There are a number of resources in a computer system such as computing, storage, and I/O bandwidth. This study focuses on the computing resource, i.e., the CPU, which is a major energy consumer. A well-known and popular energy-aware CPU management technique is a Dynamic Voltage and Frequency Scaling (DVFS) [3], [8], [9]. The DVFS was introduced decades ago, and it has been one of the most effective power saving techniques. The amount of the energy savings by DVFS, however, is decreasing due to the following reasons. First, the supply voltages have already become quite low; hence, the remaining headroom for further supply voltage reductions is small and shrinking. Second, many modern servers have two or more processor chips, each chip containing multiple CPUs (cores). Because one on-chip power distribution network is shared by all the CPUs, the CPUs on the same chip must operate at the same supply voltage level clock frequencies. Unless we do “perfect” load balancing among CPUs in a chip, the voltage level is set for the most highly loaded CPU. This results in energy waste because all the other CPUs run at higher frequency than what is actually needed. Third, in a virtualized server system, it is difficult to gather sufficient information about the running applications, which is necessary to choose the optimal clock frequency and voltage level for the CPUs. This is because the virtual machine manager (hypervisor), which conducts DVFS, resides in a privileged domain whereas the applications are running in a VM domain [4].

Another well-known CPU energy management technique is Dynamic Power State Switching. Many modern processors support multiple power states (known as C, Core C, and Package C-states). Each C-state specifies the circuits that are turned ON or OFF. Based on the history of recent workloads of CPUs, the operating system (OS) suggests C-state of each CPU. Based on this suggestion, appropriate CC and PC-states are chosen by a power control unit (PCU). This PCU is programmed using fine-tuned algorithms. These algorithms are well designed but still not perfect, so we believe that further energy savings can be achieved if there is software-level assistance; we suggest a CPU consolidation technique as the assistance. This technique explicitly defines sets of active and inactive (sleep) CPUs while meeting performance target and reducing energy consumption.

There are research studies that investigate the effectiveness of the CPU consolidation. In [10], the authors show that consolidation across CPUs in a single processor and two processor systems offers a very small amount of energy savings. They used their own benchmark which is not the standard and may not create realistic workloads. In [11], Jacob *et al.* compare core-level power gating (CPG) with DVFS and show that CPG saves more energy by 30% than DVFS. This result implies the energy savings by the CPU consolidation may be larger if the processor supports the CPG. However, the reported results are calculated from a combination of real

measurements and their leakage estimator (the adopted leakage power model is somewhat simple). In [2], the authors present a technique called core count management (CCM), which is a variant of the CPU consolidation technique, and report 35% energy savings. However, all results are obtained from a simulator, and the power and performance models are again fairly simplistic.

This paper is differentiated from the prior work because of the following reasons. First, well-designed benchmarks (PARSEC and SPECWeb2009) are used for the experiment, and we believe the workloads from these benchmarks are realistic. Second, it is intensively analyzed and discussed how CPU consolidation improves energy efficiency rather than just showing final results. Third, this paper compares energy efficiency improvement by CPU consolidation, DVFS, and combined techniques (i.e., both consolidation and DVFS are combined). Forth, all results are collected from experiments on a server system comprising two Intel® Xeon® Westmere E5620 processors. It is true that our conclusions are therefore applicable to the server under consideration. However, this server is an archetypical Intel server that can be used in a wide range of scenarios from small cluster of servers to larger data center computing. Hence, we believe there is value and benefit in understanding and quantifying the energy efficiency gains achieved by using the right combination of DVFS and CPU consolidation in such a server. Clearly, the basic power management strategy must be enhanced when we go from the level of an individual server to a very large number of servers in a data center [12]–[14]. However, further study on the server level is also important in deriving higher-level power management strategies. Moreover, to the best of our knowledge, this is the only published result that provides a detailed discussion of the power states, develops and validates power and latency models as a function of the deployed DVFS and/or consolidation policy, and suggests effective yet simple, solutions for applying a combination of DVFS and CPU consolidation to these servers (esp. for I/O bound applications).

A preliminary version of this work has been published in [15]. This archival paper is a substantially extended version, which includes a completely new power model, using a new benchmark (PARSEC), vastly more detailed experimental results and discussions, and a more efficient online CPU consolidation algorithm.

The remainder of the paper is organized as follows. Existing CPU power management technologies are reviewed in Section II. In Section III we present the power and delay models. Based on these models we discuss how CPU consolidation can improve energy efficiency. The metrics for quantifying energy efficiency are presented in Section IV. The experimental system setup is explained in Section V. Section VI presents our extensive experimental results and discussions. Finally, we summarize the results and provide final conclusions and insights in Section VII.

II. BACKGROUND—POWER MANAGEMENT TECHNOLOGIES

The purpose of this paper is to understand how effectively CPU consolidation improves the energy efficiency of

server systems so as to maximize the improvement. The consolidation interacts with the existing power management technologies, so it is helpful to understand these technologies. In this section, the processor power and performance states are briefly reviewed. In addition, we review Intel Quickpath Interface technology, which may affect the energy savings by consolidation. Before starting a discussion, a few confusing terms are clearly defined as follows.

- 1) *CPU*: All circuits used to perform arithmetic/logic operations and L1 and L2 cache memories. A term core is used interchangeably with a word CPU in this paper.
- 2) *Uncore*: All components in a processor except cores.
- 3) *Package*: A physical unit which has core and uncore. A word package is used interchangeably with a word processor.
- 4) *CPU Consolidation*: It is simply called consolidation unless it is confusing.
- 5) *Total Utilization*: Sum of percentage of times when CPUs are running codes. For example, when two CPUs are fully utilized, the total utilization is 200%.
- 6) *Average Utilization*: Average utilization per core. It is calculated by dividing total utilization by the number of active CPUs, so it should be equal to or less than 100%. This term is simply called utilization or util.
- 7) *Throughput*: The number of tasks (jobs) processed in a second.
- 8) *Delay*: Total amount of time spent for executing a task. This includes the time when a task is suspended and that waiting in a queue of the CPU scheduler.

A. Processor Power States (C, CC, PC-States)

The advanced configuration and power interface specification was developed as an open standard for OS-directed power management. Many modern OSs meet this specification. First, the specification defines C-states as processor power states; when a processor is in a higher numbered C-state, which is also called a deeper sleep state, a larger area of internal circuitry is turned OFF or inactive, which reduces power dissipation. On the other hand, it also takes longer time to go back to the operating state (i.e., C0 state). The number of supported C-states is processor-dependent. For example, the Intel Core i7 processor (code-named Nehalem) supports the following core states: 1) C0 (normal operating mode—cores in this state are either executing code or are in standby); 2) C1 (autoHALT—a low power state entered when all threads within the core execute an HLT or MWAIT instruction); 3) C1E (autoHALT with lowest frequency and voltage operating point); 4) C3 (deep sleep—cores in this state flush their L1 instruction cache, L1 data cache, and L2 cache to the L3 shared cache; clocks are shut OFF to each core); and 5) C6 (deep power down—cores in this state save their architectural state before removing core voltage). See Intel Core™ i7 datasheets for more detailed information.

The C-state is also known as a logical C-state. An OS requests a change in C-state of logical cores,¹ but the request

may be denied (e.g., autodemotion). The decision about demotion is made based on each core's immediate residency history (think of this as the breakeven time for the proposed power state transition); if an amount of future residency in idle, which is estimated from the residency history, is insufficient, then a request to transition a core into a deeper power state will be ignored. In general, the entry/exit costs (latency and energy overheads) increase when the processor/core escapes from a deeper sleep state; hence, the autodemotion prevents unnecessary excursions into deeper power states, and thereby, it reduces both latency and energy overheads of power state switches.

In addition to the logical C-states, there are two more types of hardware C-states: 1) core (CC_n) and 2) package C-state (PC_n). Based on logical C-state switch request from an OS, a PCU decides about the CC and PC-states. Each core and/or package can switch its power state independently; that is, CC-state of a core may be different from that of other cores. Likewise, PC-state of one package could be different from that of another package. The CC₀ state is a special state that deserves more discussion; a core is in this state when it is executing tasks or when it is standing by for the next task to arrive (but is still in the normal operating mode). Note that a core does not immediately switch from CC₀ state to a deeper sleep state when it becomes idle.

The PC-state of a package depends on the CC-states of cores in the package. In particular, a package state can be PC_k only when all of its cores are in CC_k state or deeper CC-states. It is because some resources shared by cores cannot be turned OFF unless all cores are inactive. For example, the Intel i7 processor's L3 cache is shared by multiple cores, so a package should stay in active state if any core is still active. Otherwise, the active cores cannot function properly.

B. Processor Performance State (P-States)

While a processor executes codes, it can be in one of several performance states (P-state), which specify the clock frequency and corresponding voltage level. At the higher frequency, the performance is higher, but its power dissipation is also higher. Similar to the C-states, the number of supported P-states is processor-dependent, and a frequency is higher at lower numbered P-states, that is, P0 is the highest performance state.

An OS chooses the P-state based on the historical workload information. The OS may not choose the same state for all cores, but all cores in Intel processors will run at the same clock frequency. Therefore, even if the OS sets different P-states for the cores, only one state is selected and applied. In general, the highest performance state is chosen, but another decision policy may be used. Because of this hardware constraint of the current Intel processors, it is recommended to distribute the workload evenly among all active cores. Otherwise, the selected P-state will be appropriate only for some cores, but not for the others.

C. Core-Level Power Gating

Recent state-of-the-art Intel processors are capable of CPG; processors can completely shut down some of the cores (the

¹A logical core is identical to a physical core unless Intel hyperthreading is enabled. In this paper, hyperthreading was disabled.

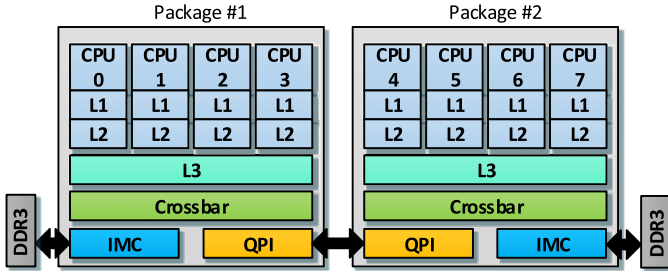


Fig. 1. Intel QPI block diagram.

OFF cores consume nearly zero power). Processors with the power-gating feature have additional C-state at which power dissipation is nearly zero, but with the largest entry/exit costs. Note that the processor used in this paper supports CPG.

D. Intel Quickpath Interconnect

In the past, there was only one memory controller (MC) in a system because one MC was enough for single core/processor systems. However, modern server systems, which have multiple packages (processors), need multiple MCs to achieve a target performance level. With a single MC, performance of all packages reduces as soon as a CPU in one package uses MC heavily. In order to address such a memory bottleneck issue, Intel invented quickpath interconnect (QPI); each package owns an integrated MC and communicates with other packages through the QPI. An example of a dual processor system is shown in Fig. 1. Half of DDR3 slots are directly connected to a first package, while the other half is connected to a second package. Therefore, if a core in package #1 needs data located in a remote DDR3, which is connected to package #2, the data are requested and received through QPI. Hence, the PC-state of some package may not be the deepest sleep state even when all CPUs in that package are inactive. This is because any CPU in other packages may be requesting data from the DDR3 slots that are directly connected to the package in question. Due to this phenomenon, it is expected that power savings achievable by package-level consolidation are negligible. This will be discussed in detail in Section VI-B.

III. POWER, DELAY, AND CONSOLIDATION

In this section, we present power and delay models for chip multiprocessor (CMP) server systems. Using these models, we will discuss how the consolidation affects the energy efficiency and delay. This discussion is abstraction level, so it may be too simplified to cover everything of real situations. However, we believe it is sufficient for deriving insights. The discussion about the power/latency tradeoffs will be verified by empirical results in Section VI. Note that thermal issues (e.g., leakage power variation as a function of chip temperature) are not considered. This is because we can do consolidation only when the system is underutilized, which also implies that the temperature of processor chips is not very high.

A. Power Model

This section presents a full platform-level power dissipation model, accounting for the power consumed by all components within a server system. This model estimates the system power dissipation by using statistical data reported by the system, i.e., the percentage of time spent in specific CC and PC-states.

The processor power dissipation consists of core and uncore power dissipations. Some notations used for the model and their definitions are shown as follows.

- 1) P_{active}^{core} : Power dissipation by a core when the core is active and executing code. This active power is a function of P-state of active cores. This is also dependent on the type of workload, but we do not consider this factor because consolidation does not change the characteristics of the current workload.
- 2) $P_{CC_n}^{core}$: Power dissipation by a core at CC_n state. Note that $P_{CC_0}^{core}$ is different from P_{active}^{core} ; a core may be simply standing by while it is in CC_0 (i.e., it is not executing any tasks although it is fully ON).
- 3) $P_{PC_n}^{uncore}$: Power dissipation by the uncore when a package is in the PC_n state.
- 4) $T_{active}^{core_i}$: Percentage of time when a core is active and executing tasks, which is also called core utilization ($util_i$).
- 5) $T_{CC_n}^{core_i}$: Percentage of time when a core is in CC_n state.
- 6) $T_{PC_n}^{uncore}$: Percentage of time spent when a package is in PC_n state.

Total (server platform) power dissipation is the sum of the processor power dissipation (P^{proc}) and the power consumed by other system components (P^{other}), e.g., I/O, memory, and hard disk drive (HDD). P^{other} is fixed and independent of the DVFS or CPU consolidation, so it acts as a fixed offset on top of the P^{proc}

$$P^{total} = P^{proc} + P^{other} = \left(\sum_i P^{core_i} + \sum_j P^{uncore_j} \right) + P^{other}. \quad (1)$$

The core power dissipation can be estimated using P_{active}^{core} , $P_{CC_n}^{core}$, $T_{active}^{core_i}$, and $T_{CC_n}^{core_i}$ as shown below. CC_0 is a special state; a core is in the CC_0 state when the core is executing codes. However, the core may be in the CC_0 state even when it is in idle. For example, a CPU stays in the CC_0 state for a certain amount of time (i.e., a timeout period) before switching to deeper power state

$$\begin{aligned} P^{core_i} &= P_{active}^{core} \cdot T_{active}^{core_i} + P_{CC_0}^{core} \cdot (T_{CC_0}^{core_i} - T_{active}^{core_i}) \\ &\quad + \sum_{n \geq 1} P_{CC_n}^{core} \cdot T_{CC_n}^{core_i} \\ &= (P_{active}^{core} - P_{CC_0}^{core}) \cdot T_{active}^{core_i} + \sum_{n \geq 0} P_{CC_n}^{core} \cdot T_{CC_n}^{core_i}. \end{aligned} \quad (2)$$

Similar to the core power dissipation, the uncore power dissipation is

$$P^{uncore_j} = \sum_n P_{PC_n}^{uncore} \cdot T_{PC_n}^{uncore_j}. \quad (3)$$

Note that $\forall i : \sum_{n \geq 0} T_{CC_n}^{core_i} = 1$ and $\forall j : \sum_{n \geq 0} T_{PC_n}^{uncore_j} = 1$.

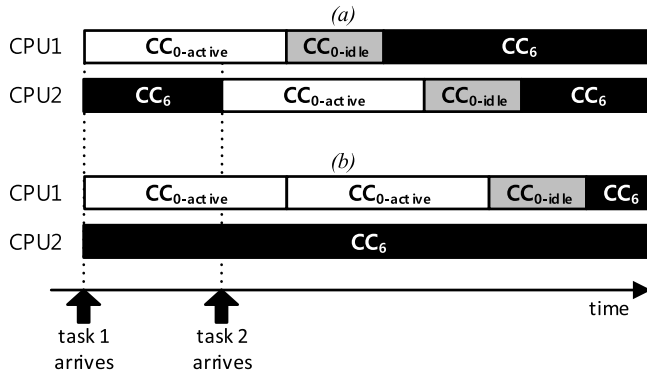


Fig. 2. Example of CC-state switch by consolidation. (a) 2 active CPUs. (b) 1 active CPU (consolidation).

B. CPU Consolidation and Power Dissipation

In this section, we discuss whether consolidation reduces power dissipation or not. We do not cover its impact on delay, which is discussed in Section III-C. The discussion here is based on an assumption that consolidation is performed in a correct way so that throughput remains the same; in other words, a sufficiently large number of CPUs are always active. Lower power dissipation at the same throughput means less energy is consumed for the same workloads, which also means more energy-efficient operation. The discussion in this section focuses on how power dissipation changes with consolidation. Therefore, we can see if energy efficiency is improved or not.

Consolidation reduces the number of active CPUs, that is, the type and level of workloads do not change. Therefore, it is expected that P^{other} is not affected by consolidation. Hence, we focus on changes in the core and uncore power dissipations

$$\Delta P^{\text{total}} = \Delta \left(\sum_i P^{\text{core}_i} \right) + \Delta \left(\sum_j P^{\text{uncore}_j} \right). \quad (4)$$

We start with the power impact on cores

$$\Delta \left(\sum_i P^{\text{core}_i} \right) = (P^{\text{core}}_{\text{active}} - P^{\text{core}}_{\text{CC}_0}) \Delta \left(\sum_i T^{\text{core}_i}_{\text{active}} \right) + \sum_{n \geq 0} P^{\text{core}}_{\text{CC}_n} \Delta \left(\sum_i T^{\text{core}_i}_{\text{CC}_n} \right). \quad (5)$$

In the above equation, term $\sum_i T^{\text{core}_i}_{\text{active}}$ is not affected by consolidation because the workload level does not change (i.e., $\Delta \sum_i T^{\text{core}_i}_{\text{active}} = 0$). Therefore

$$\Delta \left(\sum_i P^{\text{core}_i} \right) = \sum_{n \geq 0} P^{\text{core}}_{\text{CC}_n} \Delta \left(\sum_i T^{\text{core}_i}_{\text{CC}_n} \right). \quad (6)$$

As shown in the above equation, the power savings of consolidation is a function of changes in the sum of $T^{\text{core}_i}_{\text{CC}_n}$.

Let us assume that power state transition is ideal: 1) CC-state immediately switches to the deepest sleep state (CC₆) without any delay when a core becomes idle (i.e., $T^{\text{core}_i}_{\text{active}} = T^{\text{core}_i}_{\text{CC}_0}$) and 2) there is no power state switch cost: additional delay and power consumption when power

state switches. Based on these assumptions, there is negligible change in core power dissipation by consolidation because all cores are in the CC₆ state when they are idle

$$\begin{aligned} \Delta \left(\sum_i P^{\text{core}_i} \right) &= P^{\text{core}}_{\text{CC}_6} \Delta \left(\sum_i T^{\text{core}_i}_{\text{CC}_6} \right) \\ &= P^{\text{core}}_{\text{CC}_6} \Delta \left(\sum_i (1 - T^{\text{core}_i}_{\text{active}}) \right) \\ &= P^{\text{core}}_{\text{CC}_6} \left(-\Delta \sum_i T^{\text{core}_i}_{\text{active}} \right) = 0. \end{aligned} \quad (7)$$

However, this assumption is not realistic. Because of non-negligible switch costs, a core may not switch promptly its power state when it becomes idle; if the core is in low power state for the very short time, then switch costs could be greater than power savings by the switch. Consolidation can decrease power consumption by reducing the number of switch (hence, reducing switch costs). Fig. 2 shows an example which shows how consolidation reduces the costs; there are two CPUs and two CC-states available: 1) CC₀ and 2) CC₆. When a task is given to a CPU, the CPU executes the task (CC₀-active). When the execution is done, the CPU stays in the CC₀ state (CC₀-idle) for certain amount of time before switching to CC₆. For the rest of period, the CPU is in the CC₆ state. From the upper case in the figure [Fig. 2(a)], we can see one CC₆-to-CC₀ switch and two CC₀-to-CC₆ switches. On the other hand, from a consolidation case [Fig. 2(b)], there is only one switch: CC₀-to-CC₆. In addition, CPUs reside in the CC₀-idle state for shorter amount of time. Therefore, in this example, consolidation reduces power dissipation, which also means it improves energy efficiency. However, consolidation may increase execution time of a task. In this example, the second task cannot be executed promptly because a CPU is running the previous task (task 1) and this CPU is the only one active CPU. Therefore, we have to consider performance degradation and decide whether performs consolidation or not.

Second, we discuss about power dissipation impact on uncore power by consolidation. As discussed in Section II-A, a package can switch its power state to deeper one only when all cores in the package are idle. A uncore can stay longer at deeper power state when both CPUs are active [Fig. 2(a)]; there is overlap where both CPUs are in active state, so both CPUs are in the CC₆ state for longer period than the second case, where only one CPU is active. In other words, consolidation may increase uncore power. However, consolidation can reduce the percentage of time spent in the CC₀-idle state, so if the reduction is greater than the overlap, uncore power dissipation may be reduced by consolidation.

We have discussed about impacts on power dissipation using the power model, but real system is too complicated for the model to consider all factors, which affect power dissipation. Therefore, we run experiment on a real system and quantify the power savings by consolidation.

C. Delay Model

As discussed in Section III-B, consolidation may increase delay of tasks. In this section, we discuss the impact of

consolidation on delay. The proposed delay model is a function of core utilization ($T_{\text{active}}^{\text{core}_i}$). In general, the delay increases rapidly when a CPU approaches full utilization [5]

$$D_i = \frac{e}{1 - T_{\text{active}}^{\text{core}_i}} + f \quad (8)$$

where D_i is the delay of the i th CPU (core_i). Coefficient e represents how sensitive the delay is to the core utilization; with larger e , delay increases more rapidly as the core utilization approaches 1. Another coefficient f represents a lower bound on the delay, that is, the delay may not reduce below certain value even when the core utilization is very low ($D_i \geq e + f$). These coefficients are task-dependent, that is, coefficients for one task might be different from those for another task. They are also hardware-dependent. This delay is affected by consolidation because $T_{\text{active}}^{\text{core}_i}$ is a function of the active CPU count. When K tasks are assigned to the system every second, the tasks are evenly distributed to the m active CPUs by a scheduler; therefore, each CPU is assigned K/m tasks every second. The core utilization $T_{\text{active}}^{\text{core}_i}$ is linearly proportional to the workload (K/m)

$$T_{\text{active}}^{\text{core}_i} = d(K/m). \quad (9)$$

Coefficient d represents the amount of CPU resource (i.e., the number of CPU cycles) needed for executing a task. A task with higher d needs more CPU cycles compared with another task with smaller d . Now, we can model the delay as a function of the active CPU count (m) and the total number of tasks (K)

$$D = \frac{e}{1 - d(K/m)} + f. \quad (10)$$

Delay increases as the core utilization $T_{\text{active}}^{\text{core}_i}$ increases. The increase rate at high utilization is greater than that at low utilization. Consolidation increases the core utilization, but if we keep the core utilization lower than certain level (threshold), then delay increase by consolidation will be insignificant. Hence, it is important to find this threshold and keep core utilization lower than the threshold.

Coefficients in the delay model may be application-dependent, so a threshold for one application may be different from that for another application. Therefore, we will find thresholds for various kinds of benchmark tests. From experiments, we recommend that the average core utilization is no more than 70% for CPU-bound applications. Note that for memory-bound applications, where the execution time limit is tight, contention can occur on other shared resources (including bus, second level cache, and main memory), and hence, a limit on average CPU utilization will not be sufficient. However, as we will show later for such applications CPU consolidation is not an effective technique anyways. Details will be presented in Section VI.

IV. ENERGY EFFICIENCY METRICS

In the previous discussion, the term energy efficiency has been used without defining it. In order to determine if consolidation improves the energy efficiency or not, we have to precisely define what energy efficiency is. Depending on how it is defined, consolidation may or may not enhance the

energy efficiency. In this paper, we use two metrics for energy efficiency: 1) energy per task (E/task) and 2) energy-delay product per task (ED/task).

A. Energy per Task

This metric is often used for comparing energy efficiency among different platforms. A task denotes an instance of executing a specified benchmark. This metric is simply calculated using average power consumption (P_{avg}) and throughput (i.e., the number of tasks processed in a second)

$$E/\text{task} = \frac{E_{\text{gross}}}{\# \text{ of tasks}} = \frac{P_{\text{avg}} \times \text{Time}}{\# \text{ of tasks}} = \frac{P_{\text{avg}}}{\text{throughput}}. \quad (11)$$

The consolidation may decrease this metric, but it can also reduce performance. If throughput is selected as a performance indicator, then this metric also includes performance information in it. If consolidation reduces E/tasks , we can say energy savings dominate performance degradation (i.e., throughput reduction). However, for another performance definition, this metric may be insufficient; if we have to care of execution time as well as throughput, this metric does not include performance information. For example, if the execution time increases due to consolidation but throughput does not change, this metric shows that the energy efficiency is improved without any performance degradation. This may mislead into a wrong decision. Hence, we introduce another metric in Section IV-B.

B. Energy-Delay Product per Task

A delay in this metric is the average execution time of tasks. This also includes period when a task is suspended by the CPU scheduler, and CPUs execute other tasks. This metric is calculated using the average power dissipation, throughput, and execution time

$$ED/\text{task} = (E/\text{task}) \times \text{delay} = \frac{P_{\text{avg}} \times \text{delay}}{\text{throughput}}. \quad (12)$$

Depending on a metric, a different power management technique can be determined as the best one. Hence, we will report energy efficiency improvement of both metrics.

V. EXPERIMENTAL SETUP

A goal of this paper is to quantify energy efficiency improvement of consolidation and to find a way to maximize the improvement. In addition, we will compare consolidation with DVFS, which is the most popular technique. Because a real system is too complicated to be well simulated, all data shown in Section VI are measured from experiments (not simulations).

A. Hardware Test-Bed and XEN

The server system under test has two Intel Xeon Westmere E5620 processor packages, and each package in turn includes four CPUs in it (Fig. 1). As mentioned in Section II-B, all CPUs in the same package run at the same clock frequency and voltage. However, the power state of a CPU can be different from that of the other CPUs in the same package. Each 64-bit CPU has its own dedicated

256-kB L1 and 1-MB L2 caches but shares a 12-MB L3 cache with the other CPUs. The total size of the system memory is 6 GB. This processor supports seven clock frequency levels, from 1.6 to 2.4 GHz.

This considered server system may appear too small to represent typical servers in datacenters. A common myth is that the datacenters always consist of large-size servers, which have many processors. In fact, this is not true for all datacenters; the Google datacenter consists of clusters of inexpensive desktop-class machines [16], [17]. As another example, the Facebook datacenter is comprised of dual processor servers [18], [19]. There are a few reasons why datacenters consist of many small servers rather than fewer large-size servers [20]. First, resource management in many processor servers is a complex and challenging task, so actual performance may not high enough. Second, the license cost of resource management software for large servers is high. Third, it is tricky to properly handle a failure of individual component, that is, failure of one processor in a large server may cause the whole server system to fail, taking out a big chunk of computing resources within a datacenter. Hence, our setup is realistic and representative of typical server systems found in some datacenters.

A power analyzer tool measures the total platform (system) power dissipation, which includes total power consumed by all components, e.g., processor, HDD, DRAM, fan, and so on. None of the components other than CPU is optimized to achieve any power savings. For example, cooling fans are running at highest speed all the time and high-performance HDDs are used all the time in order to avoid any risk of performance degradation. Hence, the system power dissipation is very high even when the system is idle (we call this quantity the standby power from now on). In order to compensate potential power inefficiency of other system components, we calculate and report power dissipation as the difference between the total system power and the standby power

$$\text{power} = \text{power}_{\text{measured}} - \text{power}_{\text{standby}}. \quad (13)$$

The reported power value thus accounts for dynamic power consumption of all system components. The standby power of our system is 98.1 W. When the system is fully loaded, the system power is ~ 160 W; that is, we report 61.9 W as the power consumption. Consolidation is needed only when the system is underutilized. If the average core utilization is 50%, calculated power consumption is ~ 30 W. If consolidation reduces power dissipation by 15 W, then we report 50% power savings. On the other hand, the power savings would have to be reported as only 12% if we had used the total system power for the calculation. We believe reporting 50% total dynamic power saving is more indicative of the actual effect of consolidation that reporting 12% saving in the total platform power. All power dissipation numbers reported in Section VI are calculated using the above equation unless there is a specific description. A photo of the system under test is shown in Fig. 3.

We have built the virtualized system using XEN (version 4.0.1), which is an open source hypervisor-based virtualization product and provides the application programming

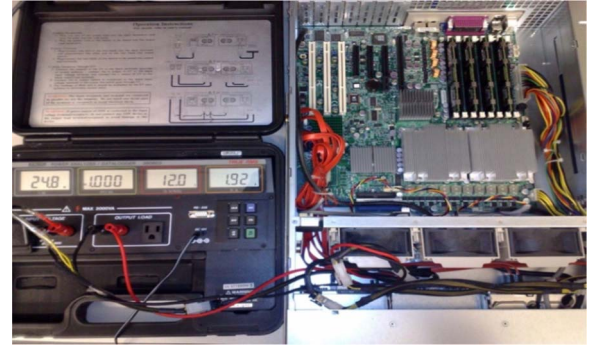


Fig. 3. Server system along with a power analyzer.

TABLE I
C-STATE LIMIT AND HARDWARE-REPORTED INFORMATION

C-state limit	Core C-state				Processor C-state			
	T_{CC_0}	T_{CC_1}	T_{CC_3}	T_{CC_6}	T_{PC_0}	T_{PC_1}	T_{PC_3}	T_{PC_6}
$C1$	available but not reported		n/a	n/a	available but not reported		n/a	n/a
$C2$			OK	n/a			OK	n/a
$C3$			OK	OK			OK	OK

interface (API) for changing VM configurations: 1) the number of virtual CPUs (vCPUs); 2) clock frequencies; and 3) the set of active CPUs. We change these configurations by calling the XEN built-in functions.

B. Benchmarks—PARSEC and SPECWeb2009

For this paper, two different benchmark suits are used: 1) the PARSEC [21] and 2) SPECWeb2009. PARSEC consists of 13 multithreaded and shared-memory programs, which represent next-generation programs for CMP. All these programs are designed and developed for real applications. Characteristics of these programs are very different from one another, and they represent wide range of applications. Therefore, we can make strong conclusion using PARSEC benchmark. Note that there are a total of 13 programs provided, but we use 11 programs. It is because facesim and ferret programs are very instable and often crashed in our setup. The PARSEC benchmark does not provide an I/O-bound program, so SPECWeb2009 was used as an I/O-bound one.

For PARSEC, we present improvement of both metrics: E/task and ED/task. On the other hand, for SPECWeb2009, we only present ED/task. Delay, which is defined as turn-around time for SPECWeb2009, is very important for Web service, so E/task is not an appropriate metric for SPECWeb2009.

VI. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, experimental results of PARSEC and SPECWeb2009 running on the system under test are presented and discussed below. We start from presenting a detailed power model as a function of CC- and PC-states. We also investigate the consolidation overhead² and suggest that the number of vCPUs has to be dynamically changed to reduce

²The DVFS overhead has been extensively studied in [22].

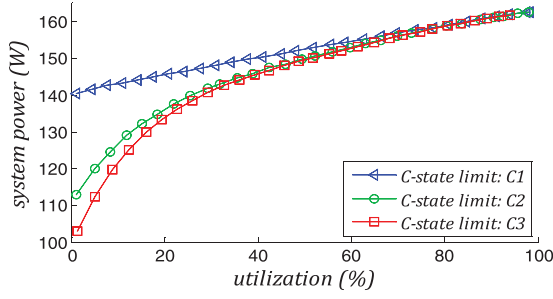


Fig. 4. Power dissipation versus utilization for C-state limits.

the overhead. It is also important to find out which set of CPUs should be active in order to maximize energy efficiency. Next, we report the E/task and ED/task improvement of PARSEC using three techniques: 1) DVFS; 2) consolidation; and 3) combined. Finally, we present a highly effective, yet simple, online consolidation algorithms for SPECWeb2009 and report energy efficiency improvement.

A. Power Model Derivation and Verification

This section presents a full platform-level power dissipation model, accounting for the power consumed by the core and uncore components within the target server system. As will be seen, this model is more detailed than the generic one that was described in Section III-A.

Our system allows limiting the deepest C-state, and we can set the limit to C1, C2, or C3 by using the `xenpm` [23] of the XEN hypervisor. The hardware-reported information for each C-state limit is shown in Table I. As shown in Table I, not all information is available; percentage of times spent in the CC_0 , CC_1 , PC_0 , and PC_1 states are not reported, and hence, these unreported times will be estimated. Our goal is to estimate the power dissipation when all C-states are available, i.e., the C-state limit is C3, but this is a difficult undertaking. Therefore, we start from the simplest case when the C-state limit is C1. Subsequently, we go over the second case when the C-state limit is C2. Finally, we will derive the power equation when the C-state limit is C3.

Power dissipation is dependent on the C-state limit, as shown in Fig. 4. For the higher C-state limit, the power dissipation is lower. Note that utilization and system power reported in Fig. 4 are all measurements. In particular, power is measured using the power analyzer tool, whereas the utilization is reported by `xentop`. The power difference among different C-state limits is greater when the utilization is lower. This is because cores stay in the C_0 state most of the time when utilization is high.

We do not provide details about how we derive the power dissipation equations for the three C-state limits, since the derivations are involved and do not fit in the page limit that we have. Full derivations can be found in the Appendix of a USC CENG technical report [24]. The key idea behind the derivation is to start with (2) and (3), and then use a combination of analytical expansion of terms, lookups from hardware-reported information (Table I), and regression analysis to derive the appropriate power macromodels, as shown

TABLE II
POWER MACROMODELS FOR THE SERVER SYSTEM UNDER TEST
(COMPRISING TWO INTEL XEON E5620
PROCESSOR PACKAGES) [24]

C-state limit	Power equation
C1	$p_{est}^{total} = 21.88T_{active} + 141.12$
C2	$p_{est}^{total} = 22.48T_{active} - 5.76T_{CC_3} - 31.16T_{PC_3} + 140.7$
C3	$p_{est}^{total} = 22.48T_{active} - 5.76T_{CC_3} - 8.56T_{CC_6} - 31.16T_{PC_3} - 42.55T_{PC_6} + 140.7$

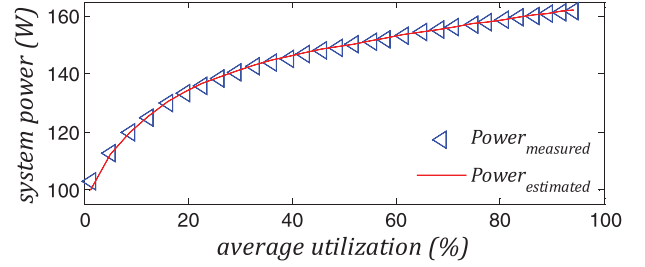


Fig. 5. Power estimation versus measurements when the C-state limit is C3.

in Table II. Note that the time spent in power states of a core is almost identical to one another because a CPU scheduler evenly distributes tasks. Hence, these times in Table II are core-independent terms.

The power models presented in Table II are highly accurate. Fig. 5 shows a comparison between measurements and model predictions for the case that C-state limit is set to C3; estimation is very close to measurements.

The first coefficient (for active state) in Table II is application-dependent. The main point here is not finding very accurate parameters of the power model but showing that power dissipation can be well estimated using CC_0 / PC_0 -states. Therefore, we will see how those states are changed by consolidation in order to understand how consolidation improves energy efficiency.

B. Package-Level Consolidation

As shown in (3), uncore power is a function of PC -states. If we have more than one package in a system, further power savings may be achieved by package-level consolidation: select CPUs from the minimum number of packages and put other packages in the deepest power state.

Package consolidation can reduce the total time spent in the active state (PC_0), which is obtained by summing over all packages the time that each package spends in its PC_0 state; thus, the uncore power dissipation decreases. In particular, package consolidation utilizes as few packages in a server as possible, so the amount of time when multiple CPUs in the same package is in the CC_0 state at the same time increases, i.e., the CC_0 state overlap time increases and, therefore, the total time spent in the PC_0 state (which is equal to $\sum_i T_{PC_0}^{core_i}$) decreases. When we consider an extreme case, this point becomes more obvious. Let us say one CPU is chosen from each package to remain active. Then, the total time spent in PC_0 state will be greater than or equal to times spent in the

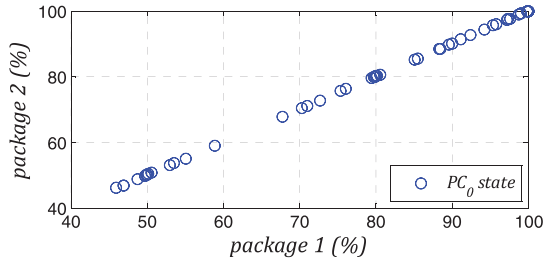


Fig. 6. Relationship between PC0 states of two packages of the target server when there are four active CPUs in exactly one of the packages.

CC₀ state of each CPU because there is no possibility for CC₀ state overlap. In comparison, if the two CPUs are chosen from the same package, then only one of the packages will be active and even then the time spent in PC₀ for that package is less than or equal to times spent in the CC₀ state of each CPU because there is CC₀ state overlap.

The above discussion is based on a key assumption, i.e., the PC-state of a package is independent of that of other packages. In practice, this assumption is far from the truth. Fig. 6 shows PC₀ state of two packages, which reveals the opposite. The data presented in Fig. 6 is for a case when all active CPUs are chosen from exactly one package, and therefore, all CPUs in the other package are idle. Fig. 6 shows that PC₀ states of both packages are nearly identical. Indeed, the same behavior is observed for other PC-states (PC₃ and PC₆), although not shown here. This implies that all packages should stay in the active state (PC₀) when any CPU, which may in fact reside in another package, is active. This is because an active CPU may need data from a remote DRAM, so not only the package where the active CPU is located but also all the other packages should remain in active states to provide the requested data (details of Intel QPI architecture have already been discussed in Section II-D) in order to avoid a significant additional latency.

As discussed above, PC-states of all packages are nearly the same due to Intel QPI, and hence, little or no energy savings are expected from package-level consolidation. On the other hand, if a task accesses memory infrequently (such as in the case of CPU bound tasks), package consolidation may save further energy. In other words, package-level consolidation may or may not improve energy efficiency depending on the characteristics of tasks. We discuss later whether or not package-level consolidation results in any energy savings.

There are eight CPUs in the system, so only when all these CPUs are inactive, both packages can be switched to inactive states (PC₃ or PC₆). Hence, we expect that the two packages are active most of the time even when the total utilization is low. As shown in Fig. 7, PC₀ state is 100% when total utilization is >150% out of 800%. This implies that there is a very small room for uncore power reduction.

C. Consolidation Overhead—vCPU Count

The number of vCPUs, called vCPU count, is an important parameter of a VM because this count limits the performance of the VM. For example, a VM with two vCPUs is capable

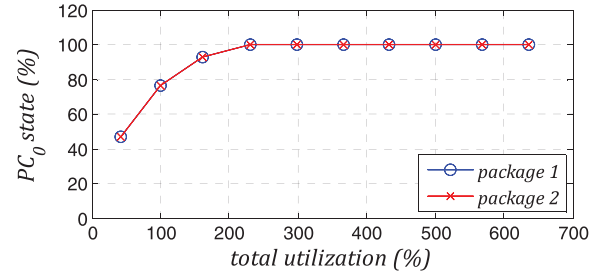


Fig. 7. Percentage of the time that each package in the target server is in the PC0 state as a function of the total utilization.

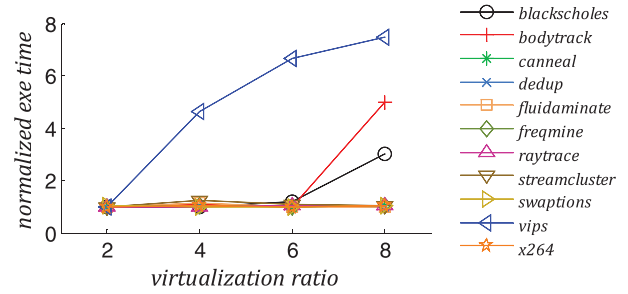


Fig. 8. Consolidation overhead, i.e., execution time as a function of the virtualization ratio.

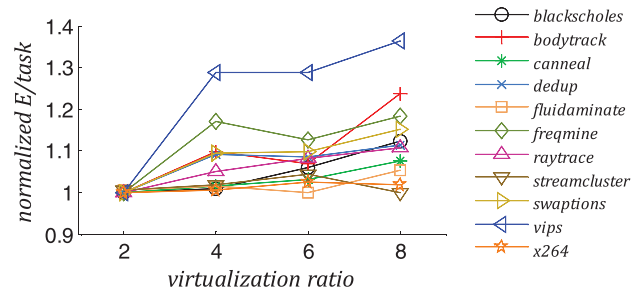
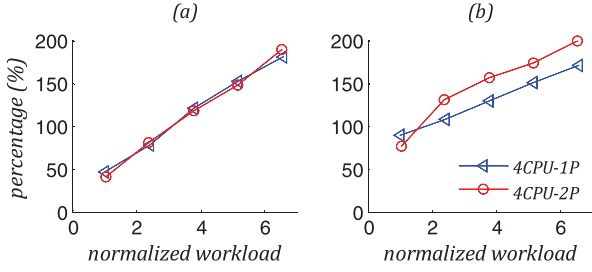


Fig. 9. Consolidation overhead, i.e., E/task as a function of the virtualization ratio.

of utilizing up to two CPUs at a time, so the maximum total CPU utilization of the VM is 200%. However, managing each vCPU causes additional overheads; thus, it could hurt both performance and energy efficiency if VMs have unnecessarily too many vCPUs. We use the ratio of vCPU to active CPU counts (called virtualization ratio) as an indicator of this overhead.

Experimental results of PARSEC benchmark programs with different virtualization ratios are reported in Figs. 8 and 9. The active CPU count is four; that is, the total CPU utilization is always equal to or less than 400%. The same experiments are repeated for four different vCPU counts: 8, 16, 24, and 32 (corresponding to virtualization ratios of 2, 4, 6, and 8, respectively). Except one program, i.e., vips, execution time remains the same when the ratio is 6 or less, while E/task of many programs increases noticeably even when the ratio is 4 (Fig. 9). This is due to higher overheads of vCPUs management. We suggest keeping the virtualization ratio to be less than or equal to 3.

Fig. 10. (a) CC₀ and (b) PC₀ state residencies for the bodytrack program.

D. CPU Selection Policy

The basic idea of the consolidation is to have as fewer active CPUs at any time. In addition to the active CPU count, the CPU selection policy can be important for multicore/processor systems, e.g., choosing CPUs from a minimum number of packages or selecting CPUs uniformly from all packages. The system under test has two packages, so there are two possible selection policies.

- 1) Select all CPUs from one package first and take additional CPUs from the other package if necessary.
- 2) Select equal number of CPUs (modulo plus/minus one) from each package.

CC₀ and PC₀ states of the bodytrack program are shown in Fig. 10. Each plot compares two selection policies: 1) all four CPUs are selected from one package (4CPU-1P) and 2) two CPUs are chosen from each package (4CPU-2P). In our experimental results, core and package may reside in either active (CC₀ and PC₀) or deepest sleep (CC₆ and PC₆) states. Hence, we only present statistics of active states. The total time spent in each state (called state residency) is calculated as the sum of all times spent in the corresponding state by all active cores (for CC₀) or packages (for PC₀), so these times can be >100%. Normalized workloads are calculated as ratios of actual workloads over workloads that result in 50% total core utilization. As shown in Fig. 10, CC₀ states of the policies close to each other, which is reasonable and expected. On the other hand, the total time spent in the PC₀ state for the first policy (4CPU-1P) is smaller than that for the other (4CPU-2P). That is, the time spent in the PC₆ state under 4CPU-1P is greater; therefore, the uncore power dissipation of 4CPU-1P is smaller. According to the discussion in Section VI-B, bodytrack is considered to be a nonmemory-intensive task.

Result of another program, canneal, is shown in Fig. 11. Here, there is a significant difference in the CC₀ state residency between the two policies. Canneal is a program to find a chip design with minimum routing cost. It uses cache-aware simulated annealing which creates intensive memory read/write activity. If we use all four CPUs in the same package, then we only use half of the L3 cache compared with the other case (4CPU-2P). This causes much higher cache misses, so both the time spent in the CC₀ state and the application execution time increase. There is negligible difference in the PC₀ state residency. From this result, package-level consolidation is not a good idea for applications requiring extensive data transfers to and from the main memory.

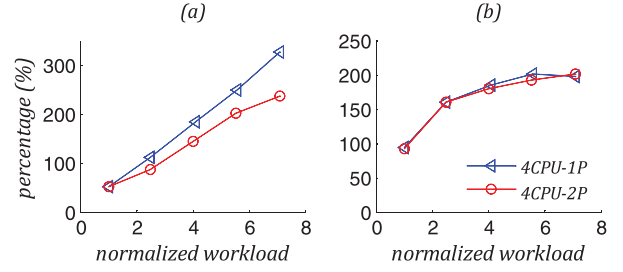
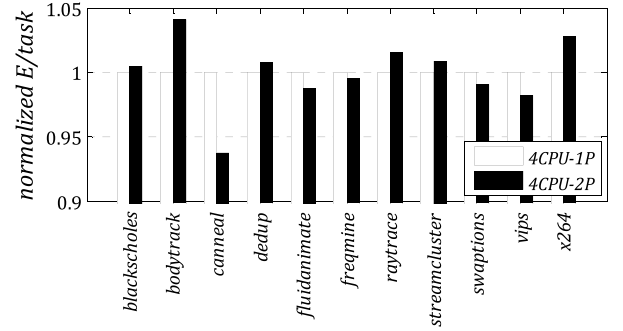
Fig. 11. (a) CC₀ and (b) PC₀ state residencies for the canneal program.

Fig. 12. Effect of simple CPU selection policies on energy consumption per task of various PARSEC programs.

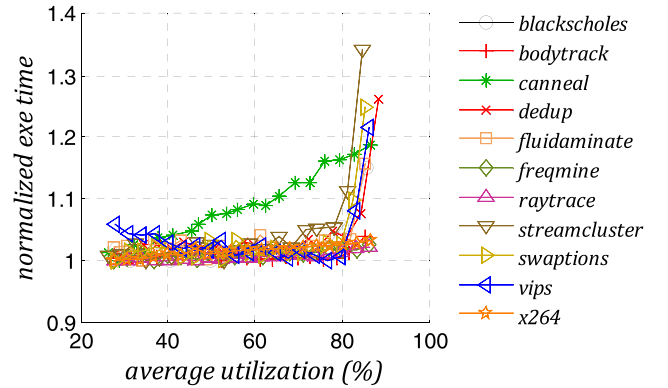


Fig. 13. Execution time of PARSEC benchmark programs as a function of the average utilization per core.

Normalized E/task comparisons for all PARSEC programs are reported in Fig. 12. Due to run-to-run variations, the average of 15 measurements is presented. E/task difference is <3% for most of programs except bodytrack, canneal, and x264. Most significant difference (~6%) is observed for canneal. Later, in this paper, we will present a more sophisticated CPU selection policy to minimize the E/task.

E. Execution Time

According to (10), the delay (execution time) of a task increases as the average utilization per core increases. The marginal rate of increase at high utilization is greater than that at low utilization. Hence, if we keep the average utilization lower than a certain threshold, then a delay increase by consolidation can be made small. The normalized execution times of PARSEC benchmark programs at various average utilizations are shown in Fig. 13.

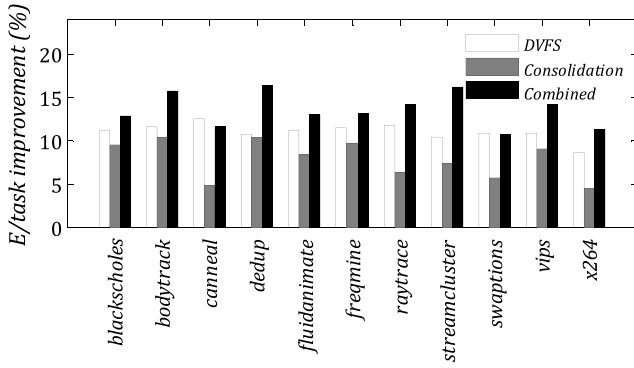


Fig. 14. E/task improvement.

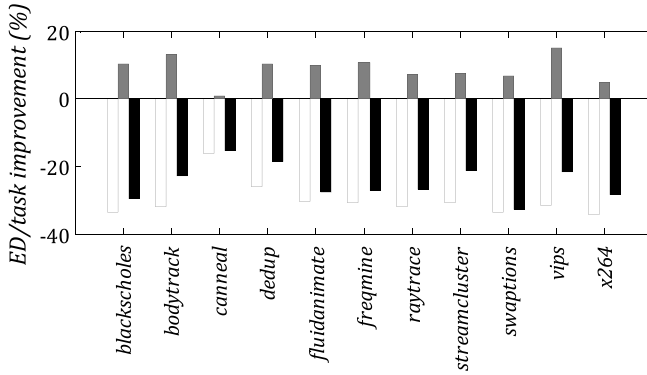


Fig. 15. ED/task improvement.

Except one program, i.e., canneal, the execution time increase is $<5\%$ at average utilizations as high as 70%. In other words, if we keep the average utilization below 70%, the maximum execution time increase will be $<5\%$ for most applications. We use this threshold to decide about the degree of consolidation that we do. For example, let us say that the average CPU utilization is 40% with eight active CPUs. If we consolidate the workloads to four CPUs, then the new average utilization will be $\sim 80\%$, which is greater than our threshold (70%). This implies that there should be at least five active CPUs in order to avoid a considerable increase in the average execution time of tasks. As discussed in Section VI-D, canneal generates extensive memory read/write requests, so the cache miss rate is expected to rise rapidly as the CPU utilization increases. In other words, the execution time of canneal increases monotonically with CPU consolidation. More generally speaking, when we have an execution time target (upper bound), aggressive CPU consolidation can result in significant SLA violation for memory-intensive applications.

F. E/Task and ED/Task Improvements for PARSEC

In this section, we present E/task and ED/task for PARSEC benchmark programs. The average utilization is kept to be $<70\%$ and the virtualization ratio not to be greater than 3 as discussed in Section VI-C.

The first interesting metric is E/task, which is reported in Fig. 14. The white, gray, and black bars in the plot show improvements in E/task achieved by DVFS,

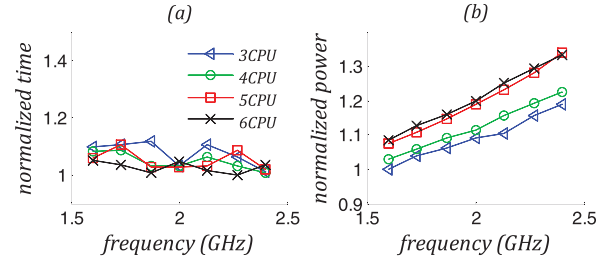


Fig. 16. (a) Response time. (b) Power dissipation.

consolidation, and both DVFS and consolidation, respectively. For all programs, improvement by DVFS is always greater than that by consolidation. Recall that the processor under test supports seven frequencies from 1.6 to 2.4 GHz. DVFS can thus effectively reduce E/task by slowing down the clock frequency from 2.4 GHz all the way down to 1.6 GHz (and accordingly lowering the supply voltage level). Evidently, this action increases the average execution time of tasks; however, this execution time increase does not affect the E/task metric much (this is because energy consumption of the server is dominated by dynamic power and not leakage power). The maximum E/task improvement achieved by consolidation is $\sim 10\%$. Another observation is that the effects are somewhat additive, that is, when we apply both DVFS and consolidation (see the combined results in Fig. 14), the improvement is greater than the other two cases for most programs with the exception of canneal. The maximum improvement of the combined technique is $>15\%$ (achieved for dedup).

Surprisingly, we observe very different results for the ED/task metric, as shown in Fig. 15. The ED/task is worsened by DVFS because the task execution time increases significantly as a result of reducing the CPU clock frequency. On the other hand, consolidation maintains its relative energy savings except for the case of canneal. This is because the execution time of canneal increases monotonically even when the average CPU utilization is kept below 70%. Therefore, the ED/task improvement of consolidation for canneal is much smaller than all other programs. From this result, we can conclude that consolidation is a much more effective solution for delay sensitive applications compared with DVFS (although it loses much of its advantage in memory-bound applications).

G. CPU Consolidation for SPECWeb2009 Benchmarks

In Section VI-F, the relative effectiveness of the CPU consolidation and DVFS was studied for the PARSEC benchmark suite. In this section, results for the SPECWeb2009 are presented. This benchmark suite comprises of I/O bound application programs whose characteristics are very different from those of the PARSEC programs. SPECWeb2009 is a very well developed benchmark suite, and its main purpose is to evaluate a Web server (I/O-bound application); hence, we can see how consolidation affects the delay and energy efficiency of I/O-bound applications from SPECWeb2009 results.

The energy efficiency is quantified as ED/packet because delay (i.e., response time) is a critical performance metric

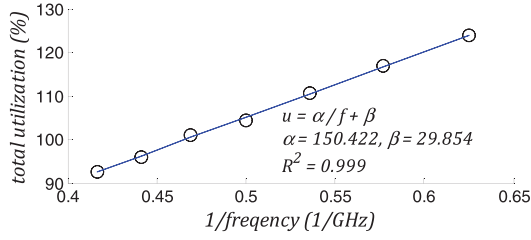


Fig. 17. Frequency versus total utilization (SPECWeb).

in these applications. SPECWeb2009 requires a simultaneous user session (SUS) count as an input, which is another way of specifying the workload intensity. The SUS count specifies only the average workload intensity (the instantaneous workload intensity fluctuates a lot). Hence, an online method, which dynamically finds optimal settings for consolidation, is needed. In this section, we start from analyzing the characteristics of the SPECWeb2009. After that, four online consolidation algorithms are presented, and results of those algorithms are reported and analyzed.

Web applications are in general not compute-intensive [25]; hence, the average response time is less dependent on CPU clock frequencies, as shown in Fig. 16(a). This is because the response time of Web servers is closely related to the I/O processes, such as network and disk access. Likewise, the response time is almost independent of the active CPU count when a sufficiently large number of CPUs are active. The relationship between the power dissipation and clock frequency/active CPU count is shown in Fig. 16(b). The power dissipation declines as the frequency decreases and/or the active CPU count is reduced. This result implies that both DVFS and the CPU consolidation improve the energy efficiency without any significant performance degradation. In addition, we expect higher power efficiency gains when both techniques are applied at the same time.

When the OS changes the CPU clock frequency, the CPU utilization also changes under the same workload. Therefore, before changing the CPU clock frequency, the corresponding CPU utilization must be estimated in order to prevent the undesirable situation whereby active CPUs are overloaded because the chosen frequency is too low for the given workload. The relationship between the total CPU utilization and CPU clock frequency is shown in Fig. 17. Note that utilization is the percentage of time that a CPU spends executing user and system space codes. When a task is waiting for an I/O operation to be completed, the task is suspended and CPU does nothing. Hence, this suspension time is not included in the utilization.

According to the R^2 value, a linear equation is a nearly perfect fit the data points in Fig. 17. The relationship is then as follows:

$$(u - \beta)f = \alpha \quad (14)$$

where $\alpha = 150.4$, $\beta = 29.9$, and $0 \leq u \leq 800$ (i.e., there are eight CPUs). Since coefficient β is relatively small, it can be ignored to simplify the relationship. Hence, the equation may

$\begin{aligned} &\text{Function } \min_cpu(u_i, f_i, c_i) \{ \\ &\quad x_c = \left\lceil \frac{u_i f_i}{u_{mid} f_{max}} c_i \right\rceil; \\ &\quad x_f = \left\lceil \frac{u_i c_i}{u_{mid} x_c} f_i \right\rceil; \\ &\quad \text{return } (x_c, x_f); \\ &\} \end{aligned}$	$\begin{aligned} &\text{Function } \min_freq(u_i, f_i, c_i) \{ \\ &\quad x_f = \left\lceil \frac{u_i c_i}{u_{mid} c_{max}} f_i \right\rceil; \\ &\quad x_c = \left\lceil \frac{u_i f_i}{u_{mid} x_f} c_i \right\rceil; \\ &\quad \text{return } (x_c, x_f); \\ &\} \end{aligned}$
---	--

Fig. 18. Psuedocodes for $\min_cpu()$ and $\min_freq()$.

be written as follows:

$$f_i u_i = f_j u_j = \alpha. \quad (15)$$

H. Online CPU Consolidation Algorithms

As shown in Section VI-F, both the clock frequency and the active CPU count affect the E/task and ED/task. In this section, we present online algorithms, which perform voltage/frequency setting and consolidation simultaneously. These algorithms monitor the CPU utilization, and change the frequency setting and/or the active CPU count depending on the current workloads. The main idea of these algorithms is to utilize as few CPUs at low frequencies as possible (while meeting the performance constraints); the decision is made by considering the current CPU utilization levels. This approach is reasonable for I/O bound applications because performance degradation is not significant unless the CPU is very highly utilized [10]. To avoid energy and delay overheads associated with frequent state changes, the proposed algorithms change the system configuration conservatively, that is, if the system is overloaded, these algorithms will immediately increase the frequency and/or the number of active CPUs. If, however, the system is underutilized, they will apply a state change (reduce frequency and/or turn OFF some CPUs) only if the situation persists for at least some time. We achieve this goal by introducing two different thresholds with hysteresis as described below.

We present four algorithms whose main ideas are quite similar to each other. If the average utilization (u_i) of a CPU is greater than an upper threshold (u_{high}), these algorithms deploy more computing resources by increasing the clock frequency of the active CPUs and/or by adding to the number of active CPUs. On the other hand, if the average utilization is less than a lower threshold (u_{low}), then they will release some computing resources by decreasing the CPU frequency and/or reducing the number of active CPUs. It is necessary to estimate the new utilization level under the new frequency and active CPU count to avoid any performance degradation. Equation (15) does not account for the number of active CPUs (c_i) in the system, and hence, it is modified to apply to this new situation

$$c_i f_i u_i = c_j f_j u_j. \quad (16)$$

Because we can change both the CPU frequency and the active CPU count (when needed), we must decide which strategy must be given higher priority: 1) changing the clock frequency first and the CPU count next and 2) changing the CPU count

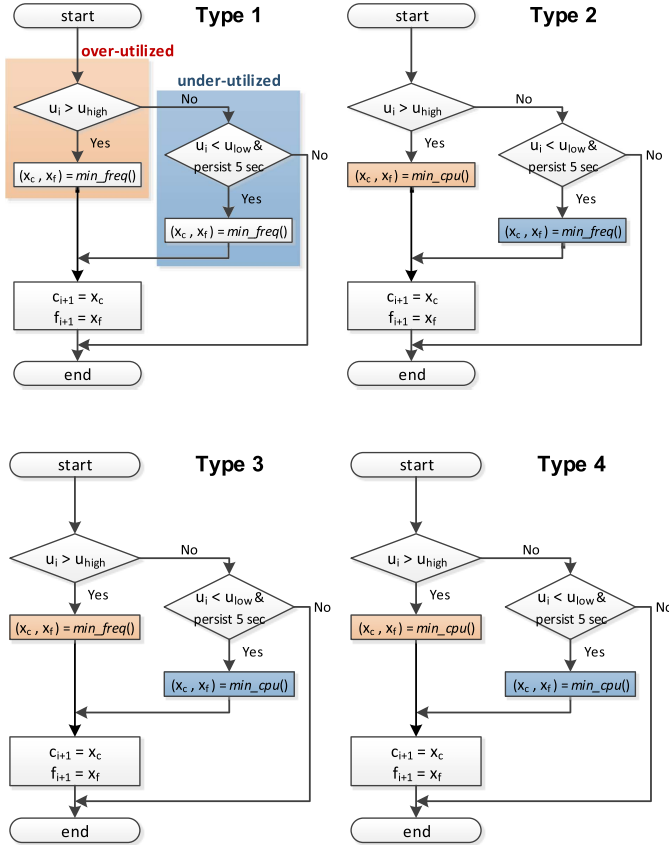


Fig. 19. Four online consolidation algorithms.

first and the clock frequency next. Two pseudocodes are presented in Fig. 18. The first function $min_cpu()$ finds the minimum CPU count (x_c) without any performance degradation. After finding the minimum CPU count, it determines the slowest frequency (x_f) with the new CPU count that would still avoid any performance degradation. This function tries to achieve a new CPU utilization close to u_{mid} , which is the median of high/low thresholds and is calculated as follows:

$$u_{mid} = \frac{u_{high} + u_{low}}{2}, \quad u_{high} = 85\%, \text{ and } u_{low} = 65\%. \quad (17)$$

The second function $min_freq()$ finds the slowest frequency first, and then finds the minimum CPU count with the new frequency. Again, no performance penalty is allowed. The two functions are called when the system is underutilized (i.e., the current utilization is smaller than u_{low}) or overutilized (i.e., the current utilization is greater than u_{high}). For each case, we can choose which function is called, i.e., $min_cpu()$ or $min_freq()$. Therefore, there are a total of four online algorithms, which are shown in Fig. 19.

The first algorithm (Type1) calls $min_cpu()$ function for both the underutilized and overutilized CPU cases. The Type2 algorithm calls $min_cpu()$ when a CPU is overutilized and $min_freq()$ if it is underutilized. The Type3 algorithm calls $min_freq()$ when a CPU is overutilized and $min_cpu()$ if it is underutilized. The last algorithm (Type4) calls $min_freq()$ for both overutilized and underutilized CPU cases.

We do experiments for three different SUS counts and compare the ED/packet and the quality of service (QoS) for

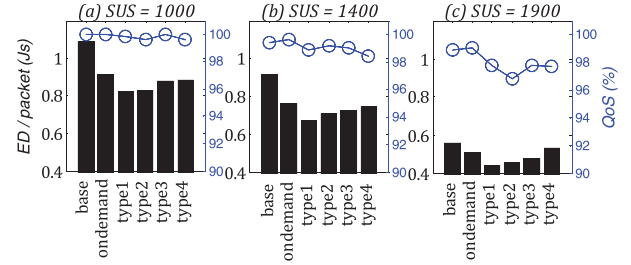


Fig. 20. ED/packet and QoS comparisons.

TABLE III
COEFFICIENTS OF THE DELAY MODEL

SUS	ED/packet (Js)		$\Delta ED/packet(\%)$
	ondemand	Type1	
1000	0.91	0.82	9.44
1400	0.76	0.67	11.83
1900	0.51	0.44	13.65

the aforesaid four consolidation algorithms and two more algorithms (read below). The QoS refers to the percentage of packets whose response time (latency) is less than the predefined threshold. This QoS is reported by SPECWeb2009 benchmark suite. In addition to the four proposed algorithms, we provide results for two other algorithms: base and ondemand. The base algorithm means there is no dynamic adjustment of the active CPU count and clock frequency, i.e., all CPUs are active and running at the maximum allowed clock frequency. The ondemand algorithm is the default DVFS method used in Linux, which does not change the active CPU count but changes the CPU frequency.

Experimental results are reported in Fig. 20. Regardless of the SUS count, the proposed algorithms always result in smaller ED/packet compared with the base and ondemand algorithms. Among the four proposed algorithms, Type1 algorithm is the best one in terms of the ED/packet. As the SUS count increases, QoS of all algorithms decreases, but QoS remains greater than 95%; hence, there are no appreciable performance degradation concerns. Note that the magnitude of ED/packet metric also decreases as the SUS count increases, which implies that the system consumes less energy for executing a packet. This is because of the energy nonproportionality of the existing server systems (including the one used in this paper). From these results, we can state that the Type1 consolidation algorithm is the best. This implies that, at least for the system under experiment, adjusting the CPU frequency has higher impact on the ED/packet metric than changing the CPU count. We compare ED/packet of the ondemand and Type1 algorithm in Table III. For three SUS settings, ED/packet of Type1 algorithm is always smaller than that of ondemand. In addition, the difference between the two algorithms increases for larger number of user sessions.

VII. CONCLUSION

DVFS has been a promising method for reducing the energy consumption, but the energy saving leverage of DVFS reduces as the supply voltage level decreases with CMOS scaling. In this paper, CPU consolidation is considered

as a substitute, or better stated, as a complement. The idea looks simple; however, we need to investigate CPU consolidation under realistic setup to maximize the energy efficiency. The effectiveness of CPU consolidation was thus investigated for different configurations: 1) types of applications; 2) the vCPU count; 3) the active CPU count; and 4) the active CPU set. From the investigation, we learn a few useful lessons. First, unnecessarily a large number of vCPUs cause significant performance degradation; hence, the vCPU count must be dynamically adjusted. Second, we need to choose a different CPU selection policy depending on applications. Third, DVFS outperforms consolidation in terms of E/task improvement. On the other hand, DVFS do not improve ED/task of PARSEC while consolidation does. Fourth, the maximum improvement of ED/task for SPECWeb2009 is also achieved when both DVFS and the consolidation are applied. Similarly, biggest E/task improvement of PARSEC is achieved when both techniques are used.

ACKNOWLEDGMENT

The authors would like to thank Dr. T. Kam of Intel Corporation, who contributed to the conference version of this paper and who provided helpful feedback and advice regarding the experimental setup, results analysis, and discussions.

REFERENCES

- [1] L. A. Barroso and U. Hölzle, "The case for energy-proportional computing," *Computer*, vol. 40, no. 12, pp. 33–37, 2007.
- [2] O. Bilgir, M. Martonosi, and Q. Wu, "Exploring the potential of CMP core count management on data center energy savings," in *Proc. Workshop Energy Efficient Design (WEED)*, Jun. 2011.
- [3] G. Dhiman, G. Marchetti, and T. Rosing, "vGreen: A system for energy efficient computing in virtualized environments," in *Proc. 14th ACM/IEEE Int. Symp. Low Power Electron. Design*, San Francisco, CA, USA, Aug. 2009, pp. 243–248.
- [4] R. Nathuji and K. Schwan, "VirtualPower: Coordinated power management in virtualized enterprise systems," in *Proc. 21st ACM SIGOPS Symp. Oper. Syst. Principles*, Stevenson, WA, USA, 2007, pp. 265–278.
- [5] N. Bobroff, A. Kochut, and K. Beaty, "Dynamic placement of virtual machines for managing SLA violations," in *Proc. 10th IFIP/IEEE Int. Symp. Integr. Netw. Manage.*, May 2007, pp. 119–128.
- [6] H. N. Van, F. D. Tran, and J.-M. Menaud, "Autonomic virtual resource management for service hosting platforms," in *Proc. ICSE Workshop Softw. Eng. Challenges Cloud Comput.*, 2009, pp. 1–8.
- [7] C. Clark *et al.*, "Live migration of virtual machines," in *Proc. 2nd Conf. Symp. Netw. Syst. Design Implement.*, vol. 2, 2005, pp. 273–286.
- [8] G. von Laszewski, L. Wang, A. J. Younge, and X. He, "Power-aware scheduling of virtual machines in DVFS-enabled clusters," in *Proc. IEEE Int. Conf. Cluster Comput. Workshops*, Aug./Sep. 2009, pp. 1–10.
- [9] P. Pillai and K. G. Shin, "Real-time dynamic voltage scaling for low-power embedded operating systems," in *Proc. 18th ACM Symp. Oper. Syst. Principles*, Banff, Alberta, AB, Canada, 2001, pp. 89–102.
- [10] M. Pedram and I. Hwang, "Power and performance modeling in a virtualized server system," in *Proc. 39th Int. Conf. Parallel Process. Workshops*, Sep. 2010, pp. 520–526.
- [11] J. Leverich, M. Monchiero, V. Talwar, P. Ranganathan, and C. Kozyrakis, "Power management of datacenter workloads using per-core power gating," *Comput. Archit. Lett.*, vol. 8, no. 2, pp. 48–51, Feb. 2009.
- [12] I. Hwang, T. Kam, and M. Pedram, "A study of the effectiveness of CPU consolidation in a virtualized multi-core server system," in *Proc. ACM/IEEE Int. Symp. Low Power Electron. Design*, Redondo Beach, CA, USA, 2012, pp. 339–344.
- [13] H. Goudarzi and M. Pedram, "Multi-dimensional SLA-based resource allocation for multi-tier cloud computing systems," in *Proc. IEEE 4th Int. Conf. Cloud Comput.*, Jul. 2011, pp. 324–331.
- [14] H. Goudarzi, M. Ghasemazar, and M. Pedram, "SLA-based optimization of power and migration cost in cloud computing," in *Proc. 12th IEEE/ACM Int. Symp. Cluster, Cloud Grid Comput.*, May 2012, pp. 172–179.
- [15] M. Pedram, "Energy-efficient datacenters," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 31, no. 10, pp. 1465–1484, Oct. 2012.
- [16] L. A. Barroso, J. Dean, and U. Holzle, "Web search for a planet: The Google cluster architecture," *IEEE Micro*, vol. 23, no. 2, pp. 22–28, Mar./Apr. 2003.
- [17] S. Shankland, *Google Uncloaks Once-Secret Server*. [Online]. Available: <http://www.cnet.com/news/google-uncloaks-once-secret-server-10209580/>, accessed Dec. 11, 2009.
- [18] *Server/SpecsAndDesigns*. [Online]. Available: <http://www.opencompute.org/wiki/Motherboard/SpecsAndDesigns>, accessed Nov. 17, 2015.
- [19] A. Andreyev, *Introducing Data Center Fabric, the Next-Generation Facebook Data Center Network*. [Online]. Available: <https://code.facebook.com/posts/360346274145943/introducing-data-center-fabric-the-next-generation-facebook-data-center-network/>, accessed Nov. 14, 2014.
- [20] L. A. Barroso, J. Clidaras, and U. Hölzle, *The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines* (Synthesis Lectures on Computer), vol. 8, 2nd ed. San Rafael, CA, USA: Morgan & Claypool, Jul. 2013, pp. 33–45.
- [21] C. Bienia, "Benchmarking modern multiprocessors," Ph.D. thesis, Princeton Univ., Princeton, NJ, USA, Jan. 2011.
- [22] S. Park *et al.*, "Accurate modeling of the delay and energy overhead of dynamic voltage and frequency scaling in modern microprocessors," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 32, no. 5, pp. 695–708, May 2013.
- [23] *Xen Power Management*. [Online]. Available: http://wiki.xen.org/wiki/Xen_power_management, accessed Feb. 12, 2015.
- [24] I. Hwang and M. Pedram, "A comparative study of the effectiveness of CPU consolidation versus dynamic voltage and frequency scaling in a virtualized multi-core server," USC CENG Tech. Rep., 2013.
- [25] D. Meisner, C. M. Sadler, L. A. Barroso, W.-D. Weber, and T. F. Wenisch, "Power management of online data-intensive services," in *Proc. 38th Annu. Int. Symp. Comput. Archit.*, Jun. 2011, pp. 319–330.



Inkwon Hwang (S'10) received the B.S. degree in electrical engineering from Seoul National University, Seoul, Korea, in 2006, and the M.S. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 2008, where he is currently pursuing the Ph.D. degree in electrical engineering.

He has been involved in power management of cloud computing systems and CPU consolidation.



Massoud Pedram (F'01) received the Ph.D. degree in electrical engineering and computer sciences from the University of California at Berkeley, Berkeley, CA, USA, in 1991.

He is currently the Stephen and Etta Varra Professor with the Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA, USA. He holds ten U.S. patents, and has authored four books, 13 book chapters, and more than 140 archival and 380 conference papers. His current research interests include low-power electronics, energy-efficient processing, and cloud computing to photovoltaic cell power generation, energy storage, and power conversion, and from RT-level optimization of VLSI circuits to synthesis and physical design of quantum circuits.

Dr. Pedram was a recipient of the 1996 Presidential Early Career Award for Scientists and Engineers. He and his students have received seven conference and two IEEE TRANSACTIONS Best Paper Awards. He has also served on the Technical Program Committee of a number of premiere conferences in his field and he was the Founding Technical Program Co-Chair of the 1996 International Symposium on Low Power Electronics and Design and the Technical Program Chair of the 2002 International Symposium on Physical Design. He is an ACM Distinguished Scientist, and serves as the Editor-in-Chief of the *ACM Transactions on Design Automation of Electronic Systems* and the *IEEE JOURNAL ON EMERGING AND SELECTED TOPICS IN CIRCUITS AND SYSTEMS*.