Simulation Modelling Practice and Theory 39 (2013) 41–55

Contents lists available at SciVerse ScienceDirect

Simulation Modelling Practice and Theory

journal homepage: www.elsevier.com/locate/simpat

A high-level energy consumption model for heterogeneous data centers

Xiao Zhang [a,*], Jian-Jun Lu [a], Xiao Qin [b], Xiao-Nan Zhao [a]

[a] School of Computer Science, Northwestern Polytechnical University, 127 Youyi xi Road, Xi'an, Shaanxi Province, China
[b] Department of Computer Science and Software Engineering, Auburn University, AL 36849-5347, United States

ARTICLE INFO

Article history:
Received 30 November 2012
Received in revised form 14 May 2013
Accepted 16 May 2013
Available online 12 June 2013

Keywords:
Energy consumption model
Heterogeneous data centers
Performance event counters

ABSTRACT

Data centers consume anywhere between 1.7% and 2.2% of the United States' power. A handful of studies focused on ways of predicting power consumption of computing platforms based on performance events counters. Most of existing power-consumption models retrieve performance counters from hardware, which offer accurate measurement of energy dissipation. Although these models were verified on several machines with specific CPU chips, it is difficult to deploy these models into data centers equipped by heterogeneous computing platforms. While models based on resource utilization via OS monitoring tools can be used in heterogeneous data centers, most of these models were linear model. In this paper, we analyze the accuracy of linear models with the SPECpower benchmark results, which is a widely adopted benchmark to evaluate the power and performance characteristics of servers. There are 392 published results until October 2012; these servers represent most servers in heterogeneous data centers. We use *R*-squared, RMSE (Root Mean Square Error) and average error to validate the accuracy of the linear model. The results show that not all servers fit the linear model very well. 6.5% of *R*-squared values are less than 0.95, which means linear regression does not fit the data very well. 12.5% of RMSE values are greater than 20, which means there is still big difference between modeled and real power consumption. We extend the linear model to high degree polynomial models. We found the cubic polynomial model can get better results than the linear model. We also apply the linear model and the cubic model to estimate real-time energy consumption on two different servers. The results show that linear model can get accurate prediction value when server energy consumption swing in a small range. The cubic model can get better results for servers with small and wide range.

1. Introduction

The power requirements of today data centers range from 75 W/ft^2 to 150–200 W/ft^2 and will increase to 200–300 W/ft^2 in the nearest future. Energy cost becomes a major part of data center operational cost. To reduce the operational cost in large-scale data centers, researchers developed a wide range of energy-saving and thermal management techniques (e.g., workload consolidation, live migration, CPU throttling solutions).

Workload consolidation is one of the most effective ways of conserving power by turning off spare servers. In many cases, the workload consolidation technique is incorporated with virtual machines, which are migrated from many physical

* Corresponding author. Tel.: +86 15902978022.
E-mail addresses: zhangxiao@nwpu.edu.cn (X. Zhang), lujianjun@mail.nwpu.edu.cn (J.-J. Lu), xqin@auburn.edu (X. Qin), zhaoxn@nwpu.edu.cn (X.-N. Zhao).

1569-190X/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.simpat.2013.05.006

machines into a smaller number of physical machines. This approach can save energy by reducing the number of active servers. Migration policies depend on both performance requirements and power consumption of different workloads. Power consumption models fall into two categories. Models in the first one can predict the power consumption of each physical machine; models in the second group aim to determine the impact of each virtual machine's load on energy consumption. Previous studies show a strong correlation between performance events and the power consumption. Existing models rely on various performance events (e.g., CPU, IO, memory, and cache) to estimate energy consumption of of sub-components of a computing system; sub-components can be classified as either CPU-bound (e.g., CPU and cache) or IO-Bound (e.g., DRAM, HDD).

The existing power consumption models are practical for specific machines under certain conditions. These models are reasonably accurate for data centers equipped by homogeneous computing platforms. However, the heterogeneous and dynamic characteristics of modern data centers make these models less accurate and trustworthy. To apply the traditional models in a heterogeneous data center, system administrators must verify the models on different types of servers. When it comes to data centers containing a large number of heterogeneous computing components, it is unproductive and time consuming to adopt and validate a single model to predict energy consumption of a wide variety of servers.

Early energy consumption model use CPU utilization as only parameter [1]. Some studies try to monitor several performance counters related with CPU to estimate the power consumption [2]. Past studies use multiple performance counters to calculate the energy consumption including CPU, memory, disks and network [3,4]. They extended the model with more parameters to get more accurate power consumptions. All of these models use linear model. Most of these models were only validated on several servers. Fan et al. validate the linear model in Google data center with thousands of servers, but they did not mention how many kinds of servers in their data center.

To address the aforementioned problem, we validate the accuracy of the linear model by comparing modeling results against available data obtained from the SPECPower_ssj2008 benchmark (or SPECpower for short), which is a widely adopted benchmark used to evaluate the power and performance characteristics of servers. Until October 2012, 392 servers have been evaluated by SPECpower. These servers came from 26 different vendors through past 6 years, these servers can represent the heterogeneous servers in the market. To our knowledge, this is the first power usage study of so many kinds of different servers. Some of our key findings and contributions are:

- A wide validation of heterogeneous servers. We validated the linear energy model by 392 published results tested by different kinds of servers. We analyzed the accuracy though *R*-squared, RMSE and average error for different kinds of servers.
- We found that not all servers fit the linear model well. 6.5% (25 kinds of servers) of *R*-squared values are less than 0.95, which means CPU utilization is not significantly correlation with power consumption.
- Although the average RMSE of all servers is 14.86, there are 118 kinds of servers' RMSE values bigger than 10 and 49 kinds are bigger than 20. Which means for these servers, even though *R*-squared value shows the linear model fit well, there is still big difference between modeled and real power consumption.
- We also illustrate the different power consumptions of servers with same CPU. We find that different servers have different power consumption characters even with same CPU and similar workloads.
- We found many servers swing in a wide range (max power consumption minus idle power consumption) does not fit the linear model well. We use high degree polynomial models to fit these data. We find cubic polynomial can get better results.
- We apply the linear model and the cubic model on two different servers to validate our conclusion. One server has a small range and another has a big range. We illustrate several estimation results under different applications and benchmarks. The results show that the server with big range has bigger error.

The rest of this paper is organized as follows. Section 2 introduces the existing energy consumption models and the SPECpower benchmark. Section 3 presents our energy consumption model using high level performance counters. Section 4 shows validation results of SPECpower results. Section 5 validate the accuracy of models in real-time manner. The last section discusses the future work of this study.

2. Related work

2.1. Energy consumption model

Power management is becoming an important issue to be addressed in data centers. Managers have to reduce energy costs of servers and cooling systems in order to offer competitive services. It is straightforward to measure the energy consumption of an entire data center [5]. To schedule jobs or workload consolidation in an energy-efficient way, one has to estimate the energy consumed by each computer node in a data center.

Power meters retrieve system power usage accurately in real-time manner. This method can not adapt to dynamic computing environments like computing clouds, because there are excessive number of computers node to be measured in real-time. The second approach is to estimate the power usage based on functional units, which sum up nominal power of each component. Values measured by this approach are constant under certain configurations, and these measured values are

always larger than actual energy usage. Ludashi applied this method to estimate the energy cost of personal computers [6]. Ludashi's approach checks all components of a PC, including main board, CPU, hard disk, and memory. The measurement results are not accurate under normal usage conditions. For example, the estimated value of a laptop (i.e., ASUS U35JC) is 107 W, but the accurate energy consumption is 47 W. CPUs cost most of the energy of computers; the power consumed by CPUs can be measured by thermal design power (TDP), which is measured under normal load [7]. Energy consumption varies sharply when workload and configuration significantly changes, thereby making a large discrepancy between estimated energy consumption and actual energy usage.

The third method is to model power usage based on performance counters. Which can be divided into two classes: chip-level performance counter queried from chips and system-level counter got from OS. Bricher et al. designed a system power model using hardware performance counters for vital system subcomponents. Their model relies on microprocessor performance counters to measure an entire system power consumption [8]. They used two different sets of performance counter on a quad-socket Intel CPU and AMD dual-core CPU. They selected nine events to model power consumption of Intel Quad-socket CPU and use thirteen different events to model power of AMD CPU [9]. Singh et al. built some model for AMD Phenom [10].

The above existing hardware performance counter solutions have the following drawbacks. First, most of the models were tailored for special processors or computer architectures. Different CPUs have different performance counters [9], which make it impractical to apply these models to heterogeneous computing environments. Intel Pentium 4 processor has 18 performance counters that can be programmed to monitor up to 59 event classes, but there are only 15 events available for Intel XScale processors [11]. Second, there are new emerging technologies (e.g., dynamic voltage scaling or replace a SCSI disk to a SSD disk) to save energy consumption in computers. The existing static estimation methods are unable to address the dynamic features of computing environments.

Some models estimate power usage based on resource utilization. Fan et al. implemented models based on CPU utilization, and estimate energy consumption of each Rack, PDU and cluster. They focus on critical power and power usage of entire clusters with several thousand servers [5]. Heath et al. extended the models by using OS-reported CPU, memory and disk utilization [4].

Some other studies try to find energy consumption of each process or each virtual machine. Snowdon et al. discussed approaches to monitoring power for applications [12]. This method involves collecting information in real time about resources consumed by each application. Snowdon's work assumes that the energy usage of an application is directly related to the amount of CPU time. The downside of this approach is that it does not take into account loads handled by I/O devices and multi-core processing. Bohra et al. proposed a model called "VMeter", which predicts instantaneous power consumption of an individual virtual machine hosted on a physical node [13]. To predict energy consumption of virtual machines in cloud computing, Karan et al. proposed Joulemeter that uses only power models to accurately infer the power consumption of a virtual machine [10].

Models based on resource utilization can be adapted to heterogeneous cluster systems. Real-life servers are characterized by different configuration, performance and workloads. Most of previous models were only validated on several different kinds of server. Fan's models were validated on several thousand servers in Google, but they did not mention how many different kinds of servers in their data center. Heath et al. tested their models on 4 blade servers and 4 PCs [4].

2.2. The SPECpower benchmark

SPECpower is the first industry-standard SPEC benchmark that evaluates the power and performance characteristics of volume server class and multi-node class computers [14]. The benchmark was created to compare energy efficiency among different servers. Currently, many vendors provide energy efficiency evaluations, but the vendor's evaluation results are not comparable due to different workloads, configurations, and test environments. The benchmark helps to measure the power of computing systems under various workloads.

The newest version of the SPECpower benchmark was released on July 26, 2012. The current version exercises CPUs, caches, memory hierarchy and the scalability of shared memory processors (a.k.a., SMP) as well as the implementations of the Java Virtual Machine or JVM, the Just-In-Time compiler or JIT, garbage collections, threads and some aspects of operating systems.

The benchmark runs on a system under test (SUT) and a controller machine, which controls workloads of SUT and collects power data from a power meter connected to SUT. Fig. 1 shows our testbed that makes use of the SPECpower benchmark to evaluate energy efficiency of computing systems.[1]

The SPECpower benchmark is comprised of processes, where CPU utilization varies from 100% to idle with an increment of 10%. SPECpower uses very little network I/O; neither does SPECpower write measured data to disks during each test. Nevertheless, SPECpower issues reads to tested disks. Previous studies show that disk subsystems have nearly a constant power consumption during the entire range of workloads [8]. CPU activities are recorded when SPECpower is running a testbed; which keeps track of CPU usage, operation per second, and power consumption.

[1] http://www.spec.org/power_ssj2008.

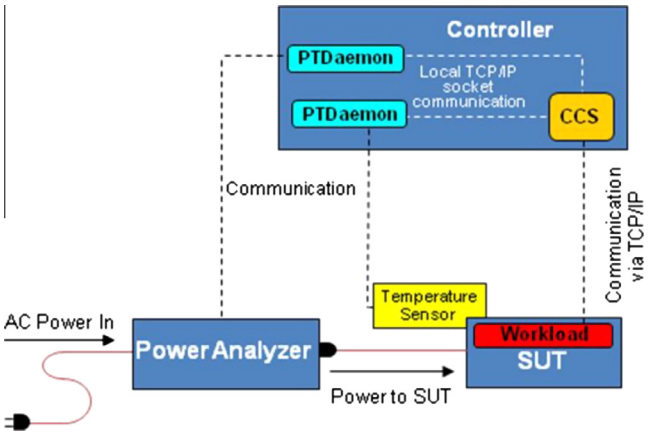


Fig. 1. Typical test environment of SPECPower_ssj2008.

3. High level energy consumption model

We start this section by offering an overview of the energy consumption model. Then, we describe how to build the model using the measurement results of SPECpower. We also will demonstrate how to apply the proposed model in heterogeneous data centers.

3.1. Overview

The first step forward to modeling energy consumption is to break down an entire computing system into several components. Recent studies show that workload has a linear correlation with energy consumption with respect to each component [8,15,16]. Fig. 2 illustrates the process of constructing, verifying, and deploying the model. In the initial phase, we build energy consumption model by analyzing the power consumption results of running the SPECpower benchmark. During the analysis of the measurements, we gather the workloads of SPECpower and map the test results into traditional workload metric, from which the energy consumption model is established. We verify the model by running the benchmark and other real-world applications on various servers. After completing the analysis and verification phases, we configure model parameters for each tested server. Next, we can integrate the model with performance monitor tools, including ganglia and performance co-pilot. Since the model is able to estimate energy consumption of all the tested servers, we can integrate the model into schedule management tools, which use energy consumption data as an input of schedule policies for heterogeneous data center.

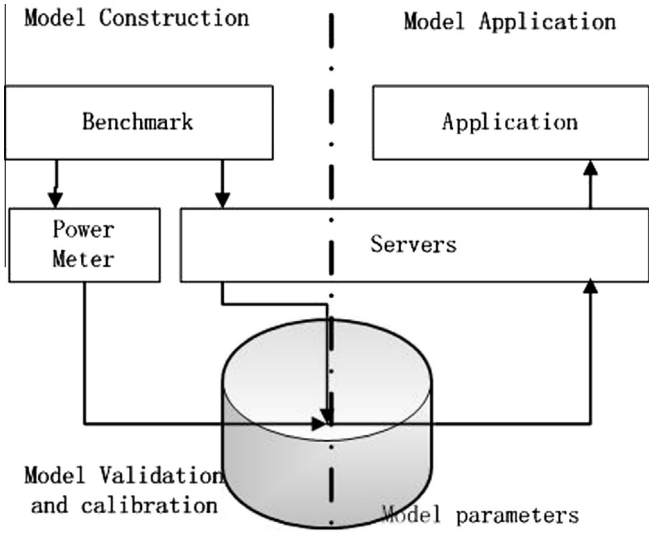


Fig. 2. Energy consumption model overview.

3.2. Model construction

Early energy consumption model use CPU utilization as only parameter [1]. Some studies try to monitor several performance counters related with CPU to estimate the power consumption [2]. Past studies use multiple performance counters to calculate the energy consumption including CPU, memory, disks and network [3,4]. They extended the model with more parameters to get more accurate power consumptions.

The energy consumption of each component can be calculated as $P = C + R \times P$, where C is a constant power consumption when the component is idle. R is the usage ratio of the component and P is the increment of energy consumption when the usage ratio goes up. Hence, the total power consumption P_{total} can be expressed as Eq. (1):

$$P_{total} = Rate \times Power = |1, R_{cpu}, R_{mem}, R_{disk}, R_{net}| \times \begin{vmatrix} P_{misc} \\ P_{cpu} \\ P_{mem} \\ P_{disk} \\ P_{net} \end{vmatrix} \tag{1}$$

where P_{misc} is a constant power consumption when the system is sitting idle. P_{misc} incorporates the power dissipation in all the components, including chassis, power supply, and peripheral devices. P_{cpu}, P_{mem}, P_{disk}, P_{net} are power consumption of CPU, memory, disk, and network interconnect, respectively. R_{cpu}, R_{mem}, R_{disk}, R_{net} are the utilization or usage ratio of the four types of resources.

Linux provides a lavish method to assess usage ratios at different levels. For example, hardware performance counters can be collected by the perfctl and perfmon driver programs [17]; and virtualized (per-process) counters also can be monitored in our testbed. The Linux servers offer utility programs (e.g., top, free and iostat) to measure usage ratios of memory, disks, network I/Os. SAR is applied to monitor the load of CPU load, disk, and memory at given intervals [18]. The challenge is that there are an excessive number of performance events; users have to pick a small set of representative events from the 40 detectable performance events provided by Pentium IV. Bircher and John studied the events of processor and selected representative events according to the examined architectures [8]. Their selections are determined by average error rates and a qualitative comparison of measurement and modeling results. Economou et al. collected data from SAR and perfctl, in their approach a single counter for each subsystem is used. The results show that most of average error of linear prediction models is less than 5% [15].

We use CPU utilization as the only parameter to estimate energy consumption. There are several advantages to take CPU utilization as the only parameter. Firstly, processors and memory are two major contributors to the power consumption of computing systems. Several past studies show that disk and network resources have almost constant power consumption [3,4]. Our findings also confirm that the energy consumption of the I/O devices and network do not noticeably change. Second, it will cause performance slowdown to capture and handle multiple performance events, especially sampling at a small interval. Data centers in Google also use CPU utilization to estimate the power usage [5]. We use the most popular linear mode to fit the data of the energy consumption of each servers (Eq. (2a). We find that linear mode does not fit well for some servers, we extend the model to higher degree polynomial (Eq. (2). The analysis show that cubic polynomial model get better prediction for these servers.

The construction process of an energy model is shown as Fig. 3. We gather benchmark results and workloads when the benchmark runs, then we can get the model parameters by regression. The parameters are generated by fitting the results of SPECpower with each model. The process is similar with "training" in machine learning, but it is different. Because there is not noise or random error in the data set and the benchmark results are steady. We use one set of parameter for one kind of servers in the cluster. The servers have same type means that they have same CPU/Chassis/Memory/Disks. The more strict definition is that servers with same model and same enclosures.

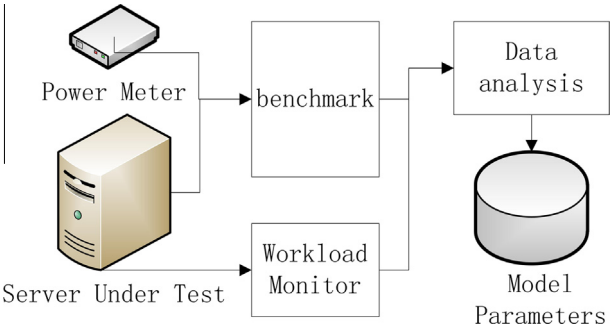


Fig. 3. Energy consumption model construction.

$$P_{total} = a + b \times R_{cpu} \tag{2a}$$

$$P_{total} = a + b \times R_{cpu} + c \times R_{cpu}^2 \tag{2b}$$

$$P_{total} = a + b \times R_{cpu} + c \times R_{cpu}^2 + d \times R_{cpu}^3 \tag{2c}$$

3.3. Model verification and analysis

We validate the accuracy of our energy model using *R*-squared, RMSE and mean error (Eq. (5)), which calculates the mean error for each combination of modeled power consumption and measured consumption. The mean error is used in many past studies [3,9].

The *R*-squared value is used to describe how well a prediction model fits a set of data (Eq. (3)). An *R*-squared near 1 indicates that the model fits the data well, while *R*-squared close to 0 indicates the model does not fit the data very well. Usually a value greater than 0.95 means the model is acceptable. A data set has values y_i, each of which has an associated modeled value f_i *R*-squared value is the value of total sum of residuals to total sum of squares.

$$SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{3a}$$

$$SS_{err} = \sum_{i=1}^{n} (y_i - f_i)^2 \tag{3b}$$

$$R^2 = 1 - SS_{err}/SS_{tot} \tag{3c}$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{3d}$$

RMSE (Root Mean Square Error) is a frequently used measure of the differences between values predicted by a model (Eq. (4)). The RMSE serves to aggregate the magnitudes of the errors in predictions for various times into a single measure of predictive power.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - f_i)^2}{n}} \tag{4}$$

The mean error is used in many past studies [3,9]. It reflects the average error from different points (Eq. (5)).

$$\mathrm{MeanError} = \frac{\sum_{i=1}^{n} \frac{|f_i - y_i|}{y_i} \times 100\%}{n} \tag{5}$$

The SPECpower benchmark tests the energy consumption of the servers under various workloads. Each result is comprised of eleven points, each of which is a set of energy consumption, CPU load, and operations per second. We verify all the 392 published results of different servers. Our results show that the linear model does not fit all condition well.

Another method to verify accuracy of Energy consumption is to compare modeled value and measure value together as shown in Fig. 4. In addition to the SPECpower benchmark, a couple of real-world applications were used to validate the energy consumption model.

Overfitting occurs when a statistical model describes random error or noise instead of the underlying relationship. When fitting a data set to different orders of polynomials, higher order function can get smaller errors but it has high varies in validation stage. We use the results of SPECpower benchmark to get the parameters of different model. The random error and

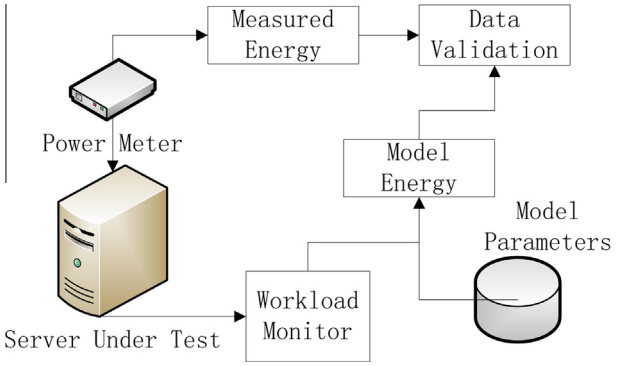


Fig. 4. Energy consumption model verification.

noise has been discarded when run the benchmark. We validate the two models by following two methods to make sure the cubic model does not over fitted:

- The more CPU utilization, the more energy consumption servers need. If the energy estimation function is $f(x)$, then the derivative function of $f(x)$ should be greater than 0 (Eq. (6)). If value of derivative function less than 0, it means the high degree function has over fitted the data.

$$f'(x) = \frac{d}{dx} f(x) \quad \text{when } 0 \leqslant x \leqslant 100, \quad f'(x) > 0 \tag{6}$$

- Regularization is a process of introducing additional information in order to solve an ill-posed problem or to prevent over-fitting. This information is usually of the form of a penalty for complexity, such as restrictions for smoothness or bounds on the vector space norm. We use cost function (Eq. (7)) to evaluate the variance of two models.

$$G(p) = \frac{1}{2m}\left[\sum_{i=1}^{m}(f(x_i) - y_i)^2 + \lambda \times \sum_{j=1}^{n} p_j^2\right] \tag{7}$$

3.4. Applying the model

Our proposed model can be used independently to estimate the energy consumption of whole cluster. We use a set of parameters for servers on each "group" of same model and same enclosures. For each server in the group, we estimate the energy consumption by their utilization (different utilization get different power consumption). We can gather real-time energy consumption from each server or just collect the utilization of each server and calculate the energy consumption on another machine.

Our model can be readily integrated with any performance monitor and scheduler (see Fig. 5). Ganglia is a scalable distributed monitoring system for high-performance computing systems like clusters and Grids. Ganglia – with a hierarchical design targeting at federations of clusters – can collect basic metrics (e.g., system load and CPU utilization). Ganglia also can keep track of user-defined metrics through plugins (e.g., C/Python modules) [19]. To implement our model in a heterogeneous cluster, we enable the modeling module to retrieve workloads from ganglia. Then, the model can output energy consumption of each computing node in the cluster. Estimated power consumption data may be used by job schedulers and workload consolidation modules in the cluster.

4. Analysis of benchmark results

4.1. Comparison results of same CPU

Previous studies show that CPU load is a major contributor to energy consumption. Since a CPU comes with matched North Bridge chips, we assume servers with the same CPU have a similar energy consumption. We choose seven different servers equipped with the same type of CPU (i.e., Intel Xeon E5-2670 2.60 GHz 2 chips 32 threads); All of the tested servers were single node. Table 1 lists the configuration of the seven servers and their power consumption. Our preliminary results show that not all the power consumption results follow the trend of a linear model. We observe that energy consumptions of the servers are very similar when the CPU load is below 50%. However, when the CPU load is increasing, the energy

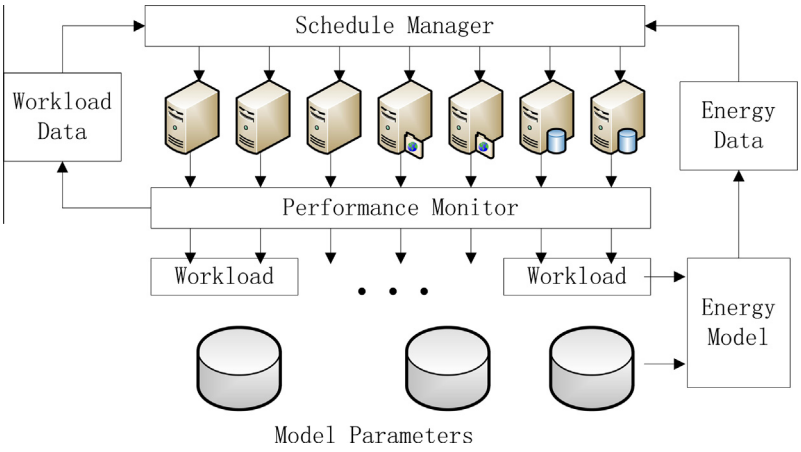


Fig. 5. Energy consumption model application.

Table 1
Power consumption of different servers with same CPU.

Hardware vendor	System enclosure	Mem. (GB)	Power information		
			Idle	Max.	Avg.
Dell	PowerEdge R620	24	54.1	243	139.5
Dell	PowerEdge R720	24	51.0	231	133.3
Dell	PowerEdge T620	24	50.2	227	131.2
Hitachi	HA8000/RS210-hHM	32	99.3	337	194.3
Hitachi	HA8000/RS220-hHM	32	104.0	325	190.9
Huawei	RH2288 V2 32G	32	55.6	282	146.1
Huawei	RH2288 V2 16G	16	54.8	279	142.9

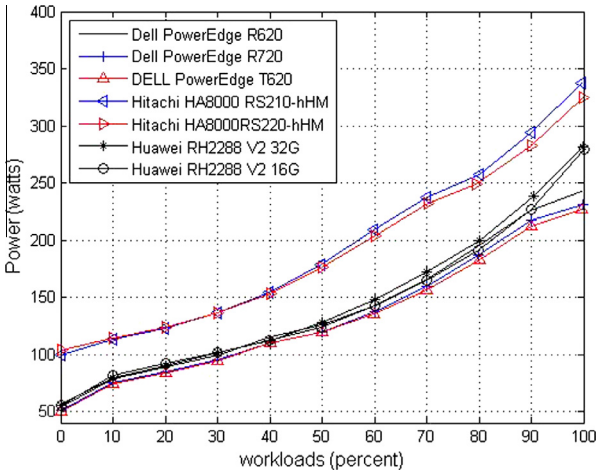


Fig. 6. Power consumption of different servers with same CPU.

consumption rates of the seven tested servers are very different from each other. This trend is especially true when the CPU load is approaching 100% (see Fig. 6).

4.2. Comparison of linear and cubic model

We validate the energy consumption of the servers running the SPECpower benchmarks. Each SPECpower result contains 11 pairs of workload and measured power consumption. We calculate the coefficients of a polynomial P (workloads) of linear mode that fits the data measured power best in a least-squares sense.

We find that the *R*-squared values of 15 kinds of servers are less than 0.95 (Fig. 7a). The minimum value is 0.8418, which comes from the server Colfax International CX2266-N2.[2] From Fig. 10b, we can find that the cubic model fits the data better than linear model.

From Fig. 7b, we can find that all *R*-squared value of cubic model is greater than 0.98. This means the model is acceptable. We use CDF (cumulative distribution function $F_x(x) = P(X \leqslant x)$) to show the distribution of three models in Fig. 7c. Eq. (8) shows the parameters of each model for CX2266, and we list measured and all modeled values for server CX2266-N2 in Table 2. We illustrate the goodness for each model in Table 3,the data shows that the cubic model fit the data best.

Fig. 8a shows that most of the mean errors of the linear model are below 8%. The average mean error is 2.74%. Interestingly, the model is more accurate before No. 300 than after. A majority of mean errors are under 4%. The first result of 2012 is No. 292, meaning that the model offers good power-consumption estimates for servers shipped before 2012. Fig. 8b shows that all of the mean errors of the cubic polynomial are below 3%. Most of the mean error values are below 1.5%. We use CDF function to show the distribution of each model in Fig. 8c. Our results show that Eq. (2c) is best for predicting the power usage by CPU utilization.

The RMSE serves to aggregate the magnitudes of the errors in predictions for various times into a single measure of predictive power. In this case, a value greater than 20 means each estimate value has a rough error of 20 W. One server will use about one more killowatt-hour every 2 days in these conditions. From Fig. 9c, we can find that 12.5% (49 servers) RMSE values are greater than 20, and several results get very large RMSE value. The average all RMSE is 14.86. The maximum value of

[2] http://www.spec.org/power_ssj2008/results/res2007q4/power_ssj2008-20071129-00018.html.

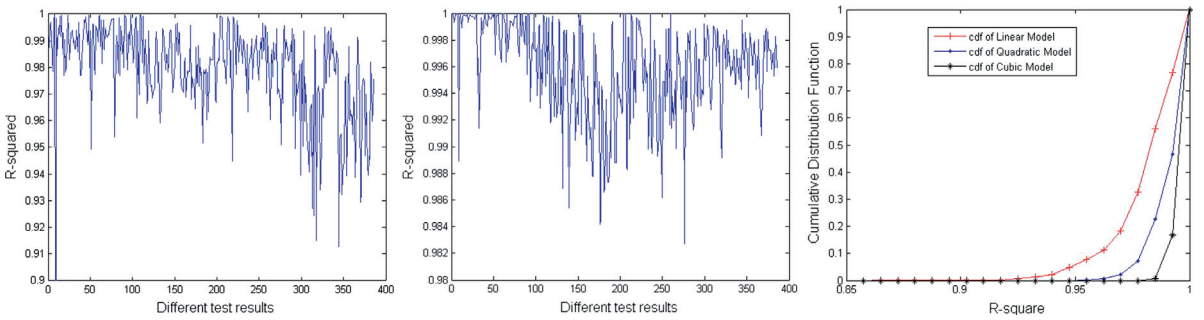


Fig. 7. *R*-squared value of different kinds of servers.

Table 2
Measured and modeled results for CX2266-N2.

No.	CPU utilization	Measured power	Modeled power		
			linear	quadratic	cubic
1	98.0	276	285.72	270.10	279.01
2	89.7	272	277.93	270.59	269.80
3	81.0	267	269.77	269.45	263.36
4	69.8	260	259.25	265.48	258.20
5	59.4	254	249.49	259.28	254.73
6	50.5	248	241.14	252.05	251.34
7	40.3	242	231.57	241.58	245.45
8	29.9	234	221.81	228.51	235.50
9	19.8	225	212.33	213.50	220.32
10	10.1	204	203.23	196.93	199.06
11	0.0	164	193.75	177.43	168.34

Table 3
Goodness of different models for CX2266-N2.

Model	Mean error	SSE	*R*-square	RMSE
Linear	0.0352	1509	0.8418	12.950
Quadratic	0.0204	510	0.9538	7.984
Cubic	0.0117	121	0.9889	3.319

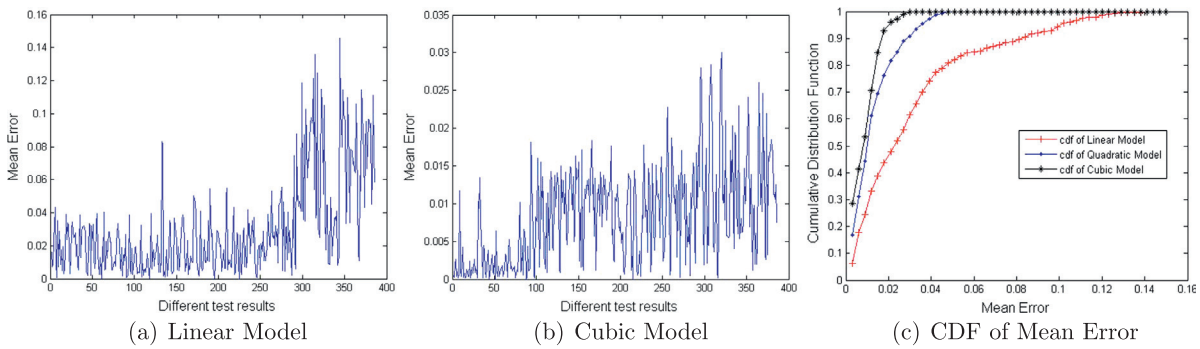
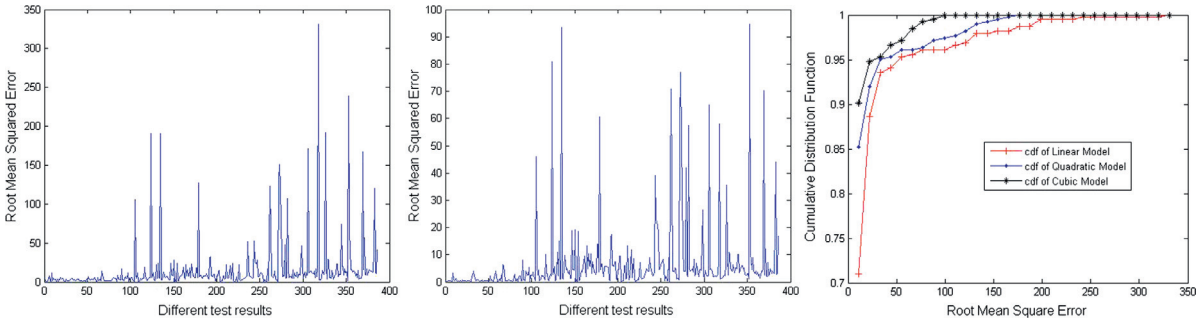


Fig. 8. Mean error of different models.

RMSE is 330.7, which comes from a high performance server (Fujitsu PRIMERGY BX920 S3),[3] which has a power consumption range from 1014 to 4965 W. The *R*-squared of the linear mode is 0.9217, which is less than 0.95. It does not fit linear model. But we find it fits cubic model well; the *R*-squared value of cubic model is 0.9976 and the RMSE is 72.63.

[3] http://www.spec.org/power_ssj2008/results/res2012q2/power_ssj2008-20120511-00459.html.



(a) Linear Model (b) Cubic Model (c) CDF of RMSE

Fig. 9. Root mean square error of different models.

Table 4
Goodness of different models for Fujitsu PRIMERGY BX920 S3.

Model	Mean error	SSE	*R*-square	RMSE
Linear	0.1245	1.20E+06	0.9218	330.7
Quadratic	0.0232	2.30E+05	0.9851	169.5
Cubic	0.0172	3.69E+04	0.9976	72.63

Fig. 9b shows that there are only 15 servers RMSE greater than 20 when using cubic model. The average of all RMSE is 6.40. The maximum value of RMSE is 94.66. The RMSE of server (Fujitsu PRIMERGY BX920 S3) is 57.94. Fig. 9c shows the CDF result of different models. Eq. (9) shows the parameters of each model for PRIMERGY BX920 S3., all modeled values can be calculated from Eq. (9). We list the goodness for each models in Table 4, the data shows that the cubic model fit the data best.

For the server in Fig. 10a, the derivative function of the cubic model is $F(x) = 0.168 * x^2 - 1.008x + 34.2$, which is greater than 0 when x is between 0 and 100. When λ is 100, $G(p)$ of linear model is 60557 and cubic model is 6996.2. For the server in Fig. 10b, the derivative function of the cubic model is $F(x) = 0.0009x^2 - 0.102x + 3.5294$, the function is also greater than 0 when x is between 0 and 100. When λ is 100, $G(p)$ of linear model is 72.5805 and cubic model is 62.1397. These results show that the cubic model does not over fit the data.

$$P_{total} = 193.7466 + 0.9385 \times R_{cpu} \tag{8a}$$

$$P_{total} = 177.4317 + 2.0432 \times R_{cpu} \pm 0.0112 \times R_{cpu}^2 \tag{8b}$$

$$P_{total} = 168.3358 + 3.5294 \times R_{cpu} \pm 0.0051 \times R_{cpu}^2 + 0.000271 \times R_{cpu}^3$$

$$P_{total} = 762.1 + 35.92 \times R_{cpu} \tag{9a}$$

$$P_{total} = 1268 + 2.156 \times R_{cpu} + 0.338 \times R_{cpu}^2 \tag{9b}$$

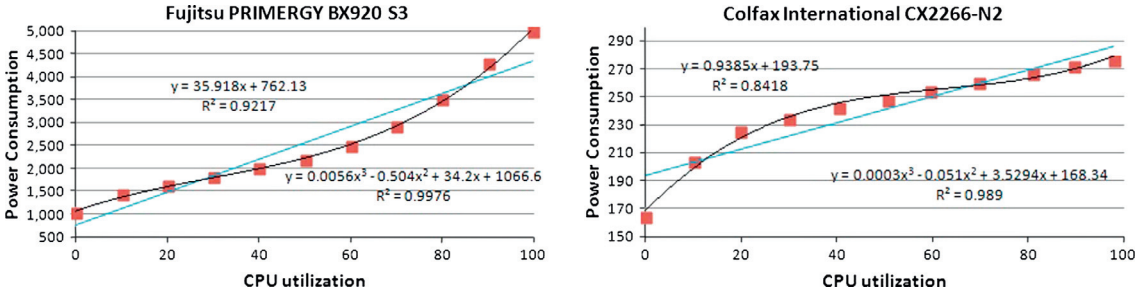
$$P_{total} = 1067 + 34.2 \times R_{cpu} - 0.504 \times R_{cpu}^2 + 0.005624 \times R_{cpu}^3$$

5. Real-time energy consumption

In addition to SPECpower results, real-time power estimation does not limit to several points. The power consumption model developed in Section 3.2 can derive from the results of SPECpower. In this section, we use the model to estimate the energy consumption of servers in a real-time manner. We validate the model to estimate the real-time energy consumption during SPECpower test processing. SPECpower benchmark only generate CPU and memory utilization, it does not generate too much IO and network workloads. We also validate the accuracy of model under complex workloads including IO.

5.1. Test environments

In our experiments, we collect performance counters from SAR and other utility programs in Linux. Sampling application also needs CPU and other resource. We collect performance counter at 10 s interval, it only use less than 2% CPU at the sampling time. We use a Chroma 66202 power meter to measure the total system power. The Chroma power meter – providing readings every 0.25–2 s – can measure power ranging from 1.5 W to 1000 W. Power meter does not use any CPU and other resource of system under test. We collect power consumption every second. The first server system is a highly integrated



(a) Energy consumption of Fujitsu PRIMERGY BX920 S3

(b) Energy consumption of Colfax International CX2266-N2

Fig. 10. Energy consumption of two servers.

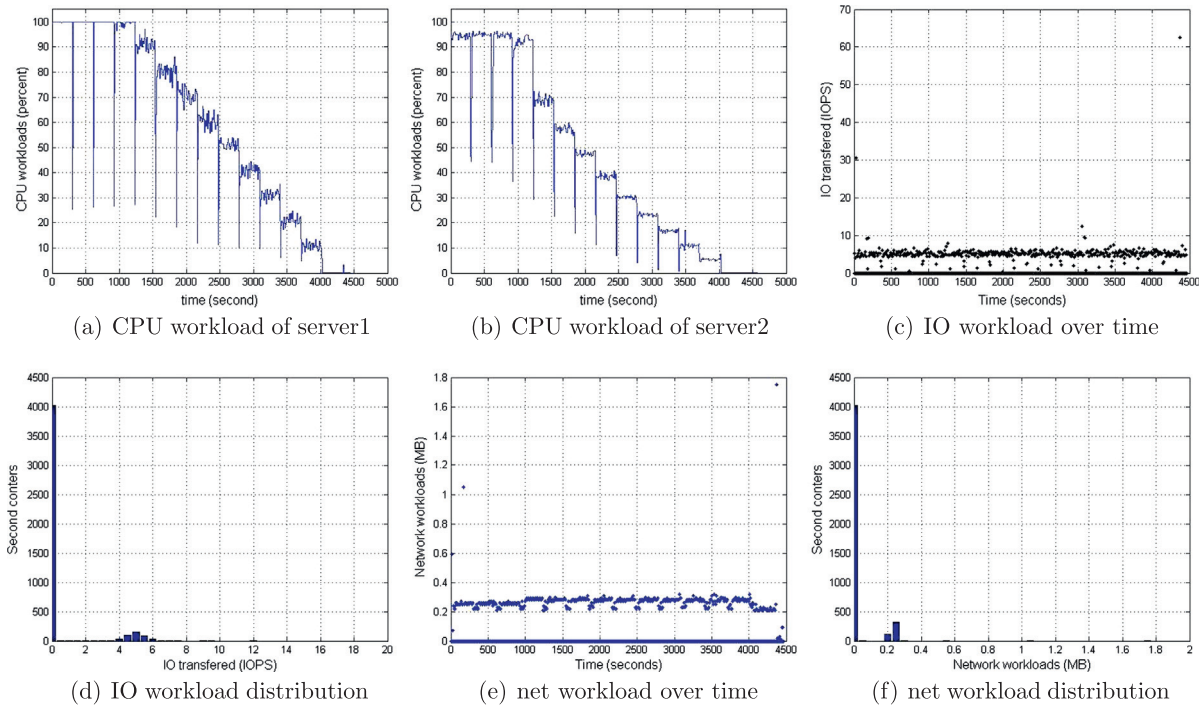
Table 5
Configurations of testbed.

Name	Hardware	Software
Server1 Inspur AS300N	1*Intel (R) Xeon (R) CPU E5502 1.87 GHz (4 core) 1*16GBytes of RAM 4*160 GBytes SATA disk (Hitachi HTS545016B9A300)	Centos5.3 Linux kernel 2.6.18
Server2 Dawning A840r-G	AMD Opteron (TM) 2.3 GHz 12 core 1*64GBytes of RAM 4*300GBytes SAS disk (Hitachi)	Redhat6.2 Linux kernel 2.6.32

blade server that includes an Intel (R) Xeon (R) processor. The second system is a high-end server containing four AMD Opteron (TM) processes. The last one is equipped with an AMD dual core processor. Table 5 summarizes the testbed used in these experiments.

5.2. Validation of real-time SPECpower benchmark power consumption

We run SPECpower on servers listed in Table 5. Firstly, we illustrate I/O and network workloads take little affect on energy consumption. Secondly, we show how to construct energy consumption model for each server. We use the results of SPEC-



(a) CPU workload of server1

(b) CPU workload of server2

(c) IO workload over time

(d) IO workload distribution

(e) net workload over time

(f) net workload distribution

Fig. 11. Workloads of servers under test.

power to get the parameter of linear model for each server. We monitor the workload and power consumption during the entire testing process.

Fig. 11a and b plots the workloads of tested benchmarks running on the servers. The benchmark generates network and disk I/O in a deterministic sequence. After comparing CPU loads imposed by the benchmark on the servers, we observe that the I/O and network workloads of the servers are very low (see Fig. 11). Regardless of the servers, the disk I/O load is close to zero in all the experiments. The entire testing procedure lasts anywhere between 4200 to 4500 s.

We use the linear model to estimate the energy consumption during the test in a real-time manner. Fig. 12 illustrates the measured and estimated power consumption of server1 and Fig. 13 is for server2.

From Fig. 12, we can find linear model get accurate estimate power consumption for server1. The error between modeled and measured power is between −5% and 3%. And the peak error appear when SPECpower switch to next stage, while the CPU utilization change dramatically. Most value is between ±2%. Fan et al. also claimed that linear model can estimate power consumption while error under 1% in Google data center [5].

Power estimate model for server2 does not fit measured power well as server1 (Fig. 13). The max error range from −50% to 179%. The peak error also occurs at switch phase. 79.13% modeled power value is between ±5%. Linear model is not suitable for server2 to estimate real-world energy consumption. We try to use high degree polynomial model to estimate power consumption. Fig. 14 show the all measured and modeled power of different CPU utilization. The results show that cubic model is the best fit model for all data. But quadratic model does better than cubic model when CPU utilization is between

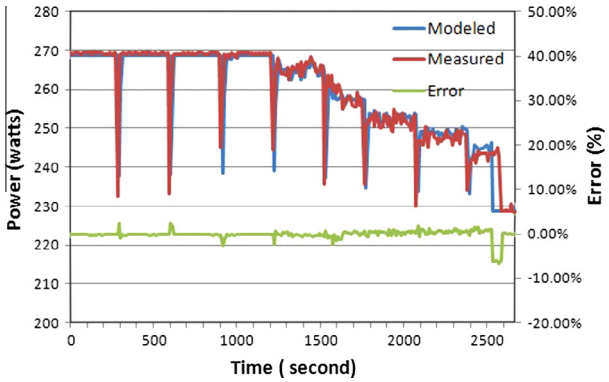


Fig. 12. Estimate power and measured power for server1.

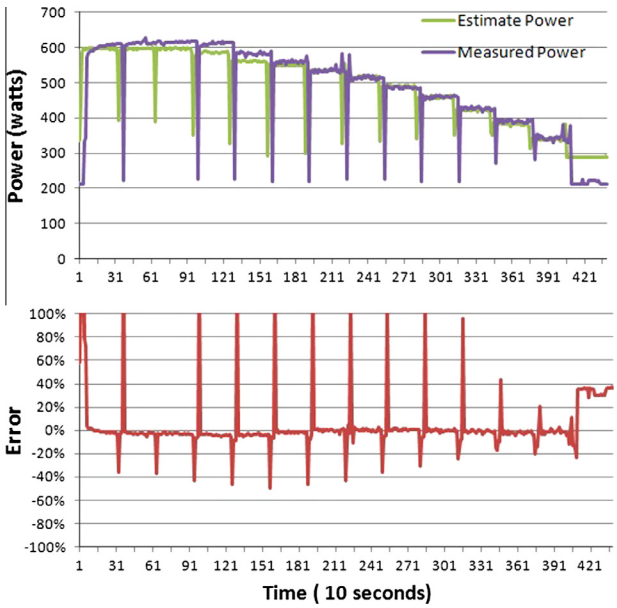


Fig. 13. Estimate power and measured power for server2.

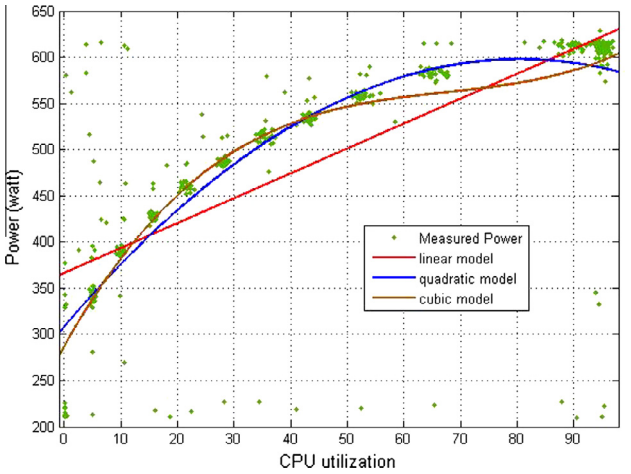


Fig. 14. Curve fitting result of models.

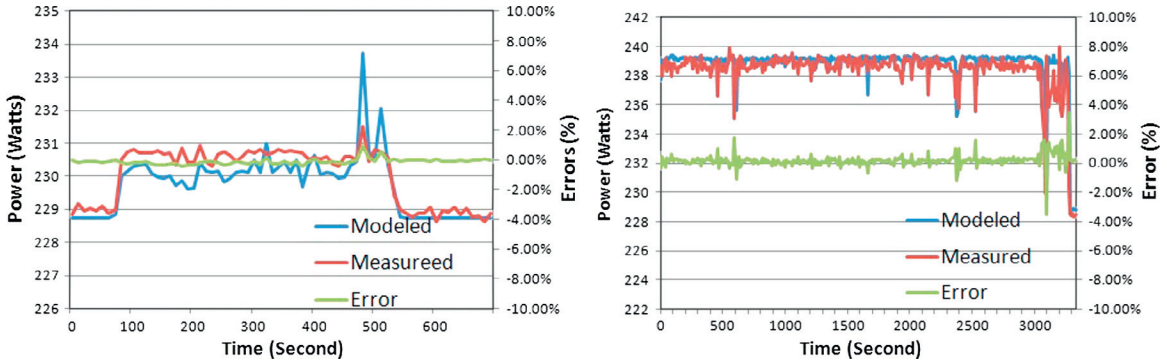
70% to 90%. The adjusted *R*-squared values are 0.5663 (linear), 0.6592 (quadratic) and 0.6703 (cubic). The RMSE are 81.54 (linear), 72.29 (quadratic) and 71.1 (cubic).

There are some points show it use extreme low power under different workloads. And there are also some points show that it use high power even CPU utilization is low. We think these are caused by sampling precision. We sampling CPU utilization every 10 s and power consumption every 1 s. If we sampling CPU and power at very small interval, it will decrease the deviation. Or if the CPU utilization keeps constant, it will get good modeled value. It is difficult to synchronize the time less than 1 s. If CPU utilization and power meter query data at different microsecond, it will cause the deviation.

5.3. Validation of real-time complex workload power

In addition to SPECpower, real applications (e.g., the gcc compiler and a Linux utility program) are running on the heterogeneous servers to validate our model. Each application represents a specific workload scenario. We collect CPU, memory, disk, and network utilization every 10 s. The evaluated workloads are CPU-bound; for example, the CPU utilization varies from 100% to 10% during a period of 40 min. We compile Linux kernel (3.2.32); the entire compilation time is 55 min. The workload of this compilation process is write-intensive (i.e., more writes than reads are issues to disks). 97% of the CPU utilization values are in the range between 23.12% and 25.52%. We use Find utility program to search keyword in the Linux kernel source codes. The utility program reads files in 12 min. This non-CPU-intensive workload has a high read I/O load.

Past studies shows that disk I/O utilization has little impact on the total energy consumption [8]. Fig. 15a and b plots the real-time results of power consumption of server1. We observe that the estimated energy consumption is slightly higher than the measured data. The error between estimated and modeling results is less than 3.5% in the case of the gcc compiler and less than 1% in the case of the find utility program. The mean error values for compile process are 0.01 (linear), 0.0091 (quadratic) and 0.0064 (cubic). Although cubic model can get better estimate results, the results of linear model is good enough for this process.



(a) Compile Linux kernel

(b) find keyword in Linux kernel source codes

Fig. 15. Estimate power and measured power with disk IO.

Table 6
Self-adaption mechanisms of each component.

No	Component name	Self-adaption mechanisms
1	CPU	dynamic frequency and voltage scaling (DVFS)
2	CPU	ACPI S3 "Sleep" state
3	DRAM	self-refresh
4	Storage	Solid State disk
5	Storage	massive array of idle disks (MAID)
6	FAN	Fan automatic speed

6. Future work

Comparing with those models relying on hardware performance event counters, our model has two compelling advantages over the hardware-event-counter-based models. First, our modeling approach can be easily applied to estimate the energy consumption of any types of servers. Second, the model can be readily incorporated in a data center equipped with a large number of heterogeneous servers.

Our model has been validated against the 392 measured results obtained from the previous studies. For most of the tested servers, the accuracy of our model is very high. However, the errors of the model for a few numbers of servers are higher than 10% (see Fig. 8). In the future, we plan to refine our model, thereby improving the accuracy of the model to offer better power-consumption prediction.

There is a trend to include power/energy sensors in the processors. Intel presented RAPL (Running Average Power Limit) technology to estimate energy by using various hardware performance counters in recent Intel CPUs [20]. The values are also exposed to users through PAPI (Performance API). Recent NVIDIA GPUs can report power usage via the NVIDIA Management Library (NVML). The nvmlDeviceGetPowerUsage ( ) function retrieves the power usage reading for the device, in milli-watts. This is the power draw for the entire board, including GPU, memory, etc. The reading is accurate to within a range of ±5 W [21]. These technologies have been validated to closely follow actual energy consumption [22]. But these models are limited to special sub components like CPU and GPU. And developers have to use different interface to get energy consumption of different sub components.

Some subsystems in a server automatically reduces their power usage after sitting idle for a period of time. Energy consumption depends on both workloads as well as previous and current power states. A linear model is unable to accurately predict changes in power states and; therefore, there is a large discrepancy between measured power consumption and that estimated by the linear model. Fig. 6 shows that energy consumption does not always have a linear relation with CPU load. This motivates us to extend our model to a non-linear one, where condition functions will be used to estimate energy consumption. For example, we may split data into ten sets of data points; we obtain ratio values from two adjacent data points.

Many energy-saving techniques (see the list in Table 6) have been developed for for laptop computers [23]. These laptop-oriented techniques may be deployed to servers to reduce energy cost of data centers. For example, the dynamic frequency and voltage scaling (DVFS) scheme decreases CPU voltage when workload is low. Evidence shows that relation between power consumption and CPU frequency is: $P = P_{fix} + P_f * f^3$, where P_{fix} is a constant CPU power consumption, P_f is the weight and f is CPU frequency. In addition to DVFS, the multi-core technique may affect energy consumption in servers.

The DRAM self-refresh solution not only isolates DRAM from memory controllers, but also autonomously refreshes DRAM content. With this technique in place, memory bus clock and PLL (Phase-locked loop) can be disabled, placing systems into a mode similar to standby. DRAM self-refresh allows systems to suspend operations of DRAM controllers to save power without losing data in DRAM.

A massive array of idle disks or MAID is an energy-efficient parallel disk system. MAID is designed for "Write once, Read Occasionally" (WORO) applications [24]. In MAID, each disk may shut down or run under low speed when no read/write operations occur. Compared to the RAID technology, MAID reduces power and cooling cost.

When it comes to non-I/O-intensive applications, disks have little impact on the power consumption of servers. Solid state disks (SSDs) costs very low power when they are idle. An SSD's is power consumption is 75% less than a typical hard drive [25]. A handful of studies show that compared with hard disks, SSDs might increase CPU load, which in turn can drive energy cost up.

Cooling subsystems have noticeable impact on energy efficiency of servers. Fan speed automatically vary according to CPU temperatures. A high CPU temperature causes the fan to consume more power. A few studies shows how to schedule CPU resources to decrease the temperatures of servers. In our future study, we will extend the model to take into account the power consumption of cooling systems.

7. Conclusion

In this study, we proposed an energy consumption model developed at the CPU utilization rather than CPU chips. This work was motivated by the fact that the existing models are inadequate for future data centers equipped with heterogeneous servers. Our modeling approach can be easily adopted to estimate energy consumption of a wide range of heterogeneous

servers, because our model is driven by CPU utilization retrieved from operating systems. We take a different approach to find the relationship between CPU utilization and energy consumption. We use published SPECpower benchmark results to find which model is the best. We find that the cubic polynomial model can get better fit results then the linear model. The accuracy of our model is validated by comparing modeling results against available data obtained from the SPECpower benchmark, which is a widely adopted benchmark used to evaluate the power and performance characteristics of servers. Experimental results of running two real-world applications on various servers confirm that the cubic model can get more accurate estimate results than the linear model. Our model can be integrated with performance monitoring tools deployed in data centers, thereby making job scheduling mechanisms more energy efficient.

Acknowledgments

This work was made possible thanks to the NPU Fundamental Research Foundation under Grant No. JC20110227, the National Science and Technology Ministry No. 2011BAH04B05 National High-tech R&D Program of China (863) under Grant No. 2013AA01A215, and the National Natural Science Foundation of China under Grant No. 61033007. And this work was supported by China Scholarship Council. Xiao Qin's research was supported by the U.S. National Science Foundation under Grants CCF-0845257 (CAREER), CNS-0917137 (CSR), CNS-0757778 (CSR), CCF-0742187 (CPA), CNS-0831502 (CyberTrust), CNS-0855251 (CRI), OCI-0753305 (CI-TEAM), DUE-0837341 (CCLI), and DUE-0830831 (SFS).

References

[1] C. Isci, M. Martonosi, Runtime power monitoring in high-end processors: methodology and empirical data, in: Proceedings of the 36th Annual IEEE/ACM International Symposium on Microarchitecture, IEEE Computer Society, 2003, p. 93.
[2] W.L. Bircher, L.K. John, Complete system power estimation: a trickle-down approach based on performance events, in: Performance Analysis of Systems & Software, 2007. ISPASS 2007. IEEE International Symposium on, IEEE, 2007, pp. 158–168.
[3] D. Economou, S. Rivoire, C. Kozyrakis, P. Ranganathan, Full-system power analysis and modeling for server environments, in: Proceedings of Workshop on Modeling, Benchmarking, and Simulation, 2006, pp. 70–77.
[4] T. Heath, B. Diniz, E. Carrera, W. Meira Jr., R. Bianchini, Energy conservation in heterogeneous server clusters, in: Proceedings of the 10th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, ACM, 2005, pp. 186–195.
[5] X. Fan, W. Weber, L. Barroso, Power provisioning for a warehouse-sized computer, ACM SIGARCH Computer Architecture News 35 (2) (2007) 13–23.
[6] A Software Predict Energy Consumption. <http://www.ludashi.com/>.
[7] Thermal Design Power List of Intel Xeon Processors. <http://en.wikipedia.org/wiki/List_of_Intel_Xeon_microprocessors>.
[8] W. Bircher, L. John, Complete system power estimation: a trickle-down approach based on performance events, in: IEEE International Symposium on Performance Analysis of Systems & Software, 2007. ISPASS 2007, IEEE, 2007, pp. 158–168.
[9] W. Bircher, L. John, Complete system power estimation using processor performance events, IEEE Transactions on Computers 61 (4) (2012) 563–577.
[10] K. Singh, M. Bhadauria, S. McKee, Real time power estimation and thread scheduling via performance counters, ACM SIGARCH Computer Architecture News 37 (2) (2009) 46–55.
[11] G. Contreras, M. Martonosi, Power prediction for intel XScale® processors using performance monitoring unit events, in: Low Power Electronics and Design, 2005. ISLPED'05. Proceedings of the 2005 International Symposium on, IEEE, 2005, pp. 221–226.
[12] D. Snowdon12, E. Le Sueur, S. Petters, G. Heiser, A platform for OS-level power management, in: The European Professional Society on Computer Systems 2009, ACM, 2009.
[13] A. Bohra, V. Chaudhary, VMeter: power modelling for virtualized clouds, in: Workshops and Phd Forum (IPDPSW), 2010 IEEE International Symposium on Parallel & Distributed Processing, IEEE, 2010, pp. 1–8.
[14] Standard Performance Evaluation Corporation. <http://www.spec.org>.
[15] D. Economou, S. Rivoire, C. Kozyrakis, P. Ranganathan, Full-system power analysis and modeling for server environments, in: Proceedings of Workshop on Modeling, Benchmarking, and Simulation, IEEE, 2006, pp. 70–77.
[16] D. Tsirogiannis, S. Harizopoulos, M. Shah, Analyzing the energy efficiency of a database server, in: Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data, ACM, 2010, pp. 231–242.
[17] Linux Hardware Performance Counter Collection. <http://sourceforge.net/projects/perfctr/>.
[18] Linux Performance Counter Collection-sar. <http://www.linuxjournal.com/content/sysadmins-toolbox-sar>.
[19] Ganglia Monitoring System. <http://ganglia.sourceforge.net/>.
[20] I. Intel, Intel 64 and IA-32 Architectures Software Developer's Manual. Volume 3b: System Programming Guide (Part 2) (2013) 14–19.
[21] N. NVIDIA, Nvml API Reference Manual, Version 3.295.45, 2012, p. 46.
[22] E. Rotem, A. Naveh, D. Rajwan, A. Ananthakrishnan, E. Weissmann, Power-management architecture of the intel microarchitecture code-named sandy bridge, IEEE Micro 32 (2) (2012) 20–27.
[23] L. Barroso, U. Holzle, The case for energy-proportional computing, Computer 40 (12) (2007) 33–37.
[24] D. Colarelli, D. Grunwald, Massive arrays of idle disks for storage archives, in: Proceedings of the 2002 ACM/IEEE Conference on Supercomputing, IEEE Computer Society Press, 2002, pp. 1–11.
[25] E. Elnozahy, M. Kistler, R. Rajamony, Energy-efficient server clusters, Power-Aware Computer Systems (2003) 179–197.