Applied Energy 265 (2020) 114806

Contents lists available at ScienceDirect

Applied Energy

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

A review of power consumption models of servers in data centers

Chaoqiang Jin, Xuelian Bai*, Chao Yang, Wangxin Mao, Xin Xu

Key Laboratory of the Three Gorges Reservoir Region's Eco-Environment, Ministry of Education, Chongqing University, Chongqing 400045, China
National Centre for International Research of Low-carbon and Green Buildings, Chongqing University, Chongqing 400045, China

HIGHLIGHTS

- Analyze the server power trends from generation to generation.
- Classify and compares existing power consumption models of servers.
- Summarize the application scenarios of the power consumption models.
- Discuss the model parameters for model construction.
- Identify the future research directions for improving data centers' energy efficiency.

ARTICLE INFO

Keywords:
Server
Power consumption model
Data center
Energy performance

ABSTRACT

This study provides an overview of power consumption models of servers in data centers. The server is the basic unit of both power and heat flow paths; therefore, its power consumption model can be used for both energy management and thermal management. Investigations of server power trends were carried out according to the data from the Standard Performance Evaluation Corporation (SPEC). It is found that a heavier workload can be handled without consuming more energy, and the difference between the peak power and idle power of the servers is not consistent from generation to generation. Furthermore, the existing power consumption models are categorized as additive models, baseline power + active power (BA) models, and other models based on calculation formula and other factors. Specifically, there are four forms of BA models: linear regression models, power function models, non-linear models and polynomial models. Besides, these models have been compared in terms of accuracy. It can be found that the polynomial model and the linear regression model perform better in terms of accuracy. Additionally, the model applications are summarized. Considering server architecture upgrades and technological innovation, the establishment of the new model and its application scenarios are discussed. Moreover, in-depth and accurate power consumption models must be extensively researched and applied to effectively improve data centers, including information technology (IT) equipment and cooling equipment, in terms of overall energy performance.

1. Introduction

As the digital transformation of industry continues to accelerate, the new generation of IT integration innovation (i.e., mobile Internet, Internet of things, cloud computing, big data, artificial intelligence, etc.) is increasingly active [1], which drives the growth of the economy and improves labor productivity [2]. According to the Synergy Research Group, there were nearly 400 large-scale data centers in the world in 2017. Of these data centers, 44% were in the United States, 20% were in China, Japan, and the United Kingdom, and 3–5% were in Australia, Germany, Canada, Singapore, India, and Brazil [3]. In China, the size of the big data industry in 2017 was 470 billion yuan, a year-over-year increase of 30%, the overall market size of cloud computing reached 94.61 billion yuan with a growth rate of 32.4%, and the overall energy consumption of data centers reached 120–130 billion kWh, which accounts for about 2% of the total Chinese electricity consumption [4]. In 2014, data centers in the U.S. consumed an estimated 70 billion kWh, representing approximately1.8% of the total U.S. electricity consumption [5].

Data centers have four major components: power equipment, cooling equipment, IT equipment, and miscellaneous components [6]; IT equipment and cooling equipment are the two main components, accounting for approximately 90% of the total energy consumption of the data center [7]. As shown in Fig. 1, servers consume energy and

* Corresponding author.
E-mail address: xuelianbai@cqu.edu.cn (X. Bai).

https://doi.org/10.1016/j.apenergy.2020.114806
Received 27 November 2019; Received in revised form 3 March 2020; Accepted 5 March 2020

Nomenclature

Abbreviations

BA	Baseline power + Active power
CFD	Computational fluid dynamic
CRAC	Computer room air conditioner
CRAH	Computer room air handler
DDR	Double data rate
DVFS	Dynamic voltage and frequency scaling
I/O	Input/output
ICT	Information and communications technology
IPO	The ratio of idle power to peak power
IRO	The ratio of idle power to rated power
IT	Information technology
OS	Operating system
PCM	Phase-change memory
PMC	Performance monitoring counters
PRO	The ratio of peak power to rated power
RCI	Rack cooling index
RHI	Return heat index
RTI	Return temperature index
SHI	Supply heat index
SLAs	Service-level agreements
SPEC	Standard Performance Evaluation Corporation
SSD	Solid-state drive
TDP	Thermal design power
UPS	Uninterrupted power supply
VM	Virtual machine
VRM	Voltage regulator modules

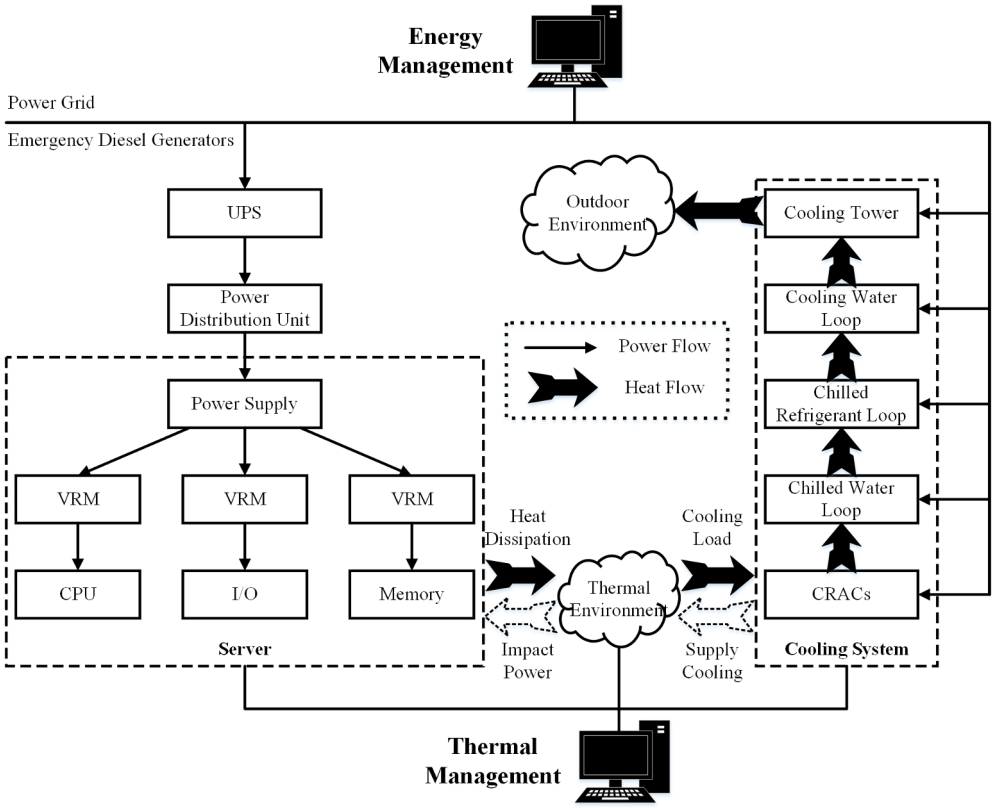

Fig. 1. The power flow and heat flow in general data centers.

dissipate heat to the thermal environment, which determines the cooling load of the cooling system and thereby influences the electricity consumption of the cooling equipment. In turn, changing the operating parameters of the cooling system affects the inlet air temperature and air flow of servers, which impact the energy consumption of the servers. Hence, the energy consumption of servers is the basic unit of power flow and heat flow in data centers, and IT equipment and cooling equipment are coupled due to the thermal environment. Moreover, with the continuous improvement of equipment efficiency, the energy savings of IT and cooling equipment will be maximized, and the energy conversation potential will be reflected in power and thermal management. To perform these two management functions in data centers, an accurate power consumption model of servers is necessary. Such a model helps in the identification of optimization opportunities and in the prediction of the consequences of the decisions and results in more effective management/control; consequently, energy savings can be maximized. Furthermore, such a model plays an important role in the cooling system design and power trend forecasting of data centers.

The purpose of this paper is to provide an overview of the recent power consumption models of servers and related studies. Previous reviews of data center thermal environments and energy efficiency are summarized. Subsequently, the development trend of server power consumption, which is one of the factors to be considered in the establishment of power consumption model, is analyzed. Then, the recent power consumption models of servers are classified, and their merits and demerits are compared. Furthermore, the application of these models is discussed. Finally, further studies of power consumption models are analyzed and discussed.

2. Related work

Although the modeling, prediction, and application of the energy

Table 1
Summary of review articles about thermal environment, energy efficiency and power models for data centers.

Reference	Major contents	Major contributions and conclusions
Lu et al. [8]	(1) Airflow distribution in data centers including ventilation configurations, underfloor plenum, and row and rack-based solutions (2) Airflow management and optimization in cold aisle containment and hot aisle containment systems	Geometry of plenum and inner obstructions are major factors for airflow distribution in underfloor data centers. Numerical simulations have shown great agreements with the experiment data. Different combinations of air distribution systems and airflow manage methods need to be explored to improve the thermal environment
Alkharabsheh et al. [9]	(1) Numerical modeling of raised-floor plenum, perforated tiles, rack, CRAC/CRAH, room level, and dynamic models (2) Experimental measurements (3) Recent technologies including containment systems, economizer cooling, hybrid-cooled systems, and device level liquid-cooling solutions	Numerical modeling with emphasis on CFD, experimental measurements, and recent cooling technologies (containment systems, economizer cooling, hybrid-cooling, and device level liquid cooling) are extensively reviewed. There are many challenges facing thermal management in data centers, such a workload variation, environment changes, and scaling issues (data centers vary in size, complexity, and business objective).
Chu and Wang [10]	(1) Methodologies in exploration the performance of data centers (2) Long-distance cooling including raised-floor and overhead air supply (3) Short-distance cooling and airflow management of rack-level cooling	CFD plays vital role in data center design and simulation. However, the simulation scale may change from sub-micro to hundreds meters. For long-distance cooling systems, the outside ambient air can be directly brought into data centers through filters which temperature is lower than that in data center. For short-distance cooling systems, the heat pipe or thermosyphon system may have better energy efficiency and cooling capacity.
Rambo and Joshi [11]	(1) Data center modeling objectives (2) Numerical modeling (3) Model validation (4) Rack-level compact modeling (5) Data center dynamics	Future trends in reduced or compact modeling of data center airflow and heat transfer are presented to serve as an overview of integrating rack-level compact models into full-scale facility level numerical computations. Compact models can be used to efficiently model data centers through varying model fidelity across length scales. Dynamic effects can be included to develop next-generation control schemes to maximize data center energy efficiency.
Wang and Khan [12]	(1) Green data center practice and benchmarking (2) Green performance metrics (3) Greenhouse gas emission (4) Humidity (5) Thermal metrics (6) Power/energy metrics (7) Multiple indicators (8) TCO: total cost of ownership	Performance metrics are important for designing, building and evaluating a data center. It also encourages the use of new technologies for next generation data centers. The key driving features to develop above metrics include: economic features and environment concern. Metrics for balancing purchasing cost and operation cost and for balancing compute performance and green performance need to be addressed.
Jin et al. [13]	(1) Factors influencing the thermal environment from room level to server level (2) Numerical study including simulation software and validation (3) Airflow performance metrics and thermal optimization	As for the room level, it is recommended that the CRACs locate at the ends of each rack row could be the optimal choice, and the height of plenum and ceiling is 0.79–0.91 m and 2.52 m, respectively. Besides, it is suggested that the open area of the perforated tiles is 25%, and the airflow outlet angle is 60°. Comparing existing simulation tools, 6SigmaDCX is recommended because of its specialty, easy to use. If engineers want to assess the magnitude of recirculation and mixing of hot and cold air, SHI or RHI combining with β index can be the best choice. If the engineers want to address the cooling efficiency, RCI and RTI are recommended.
Ge et al. [14]	(1) General power-saving strategies (2) Taxonomy of power-saving research in data center networks and content delivery networks (3) Power-saving schemes in data center network and content delivery network	Dynamic provisioning-based strategy is the most effective power-saving approach. Response time plays an important role for data center network and content delivery network to pursue better power-saving effects while keeping the power-performance tradeoff well balanced.
Mittal [15]	(1) The techniques for managing power consumption of embedded system (2) The need of power management (3) A classification of the techniques on similarities and differences	Power management techniques which ensure high-performance and low-costs are expected to become a crucial part of future enterprise architectures.
Mittal [16]	(1) DVFS based techniques (2) Server consolidation and power state transitioning based techniques (3) Workload scheduling based techniques (4) Thermal-aware power management techniques	The leakage energy consumption is increasing and hence the dynamic range of energy consumption that DVFS can utilize has reduced.
Orgerie et al. [17]	(1) The techniques for improving the energy efficiency of computing and networking resources (2) Studies and models for estimating the energy consumption of these resources (3) A classification of existing solutions and research work	Energy-efficiency approaches and technologies are often combined and applied in a coordinated way in large-scale distributed systems. Although many research directions have been studied to save energy, several key problems remain open.
Beloglazov et al. [18]	(1) Causes and problems of high power/energy consumption (2) A taxonomy of energy-efficient design of computing systems covering the hardware, operating system, virtualization and data center levels	Currently IT infrastructures contribute about 2% of total CO_2 footprints. And the power management problem becomes more complicated when considered from the data center level.
Bostoen et al. [19]	(1) Research on power-aware enterprise storage systems (2) Exiting power-reduction techniques	A majority of power-reduction techniques is based on dynamic power management. For every energy-conservation technique, the fundamental trade-offs between power, capacity, performance, and dependability are uncovered.
Wang et al. [20]	(1) Energy efficient computer servers in hardware-level, system-level, and application-level (2) Energy efficient cluster systems in virtualization, cluster-level optimization, heterogeneous energy strategy, and green cloud computing	For building more energy efficiency systems, we must construct a model about the relationship between energy and performance in a given hardware environment. The model must consider such factors as resource utilization and system performance requirements.
Hammadi and Mhamdi [21]	(1) Conventional tree-based data center network architecture (2) The data center architecture evolution	

(*continued on next page*)

Table 1 (*continued*)

Reference	Major contents	Major contributions and conclusions
	(3) Data center network architectures (4) The energy-saving approaches and technologies	A good architecture design should be tailored along with different algorithms to address not only power consumption but also network performance along with cost.
Darahmeh and Wang [22]	(1) Airside economizers in direct free cooling, indirect cooling and multi-stage evaporative cooling systems (2) Waterside economizers in integrated dry cooler-chiller system and cooling tower system (3) Cold energy storage systems and integrated system of mechanical refrigeration and thermosyphon	The free cooling technologies applicable for data centers were summarized, including airside economizers, waterside economizers and heat pipe application. The indirect airside coolers such as air-to-air heat exchanger system and heat wheel shows very high efficiency, yet the thermosyphon system reveals even more promising features.
Ni and Bai [23]	(1) Air conditioning summary including the indoor thermal guidelines, cooling methods and air distribution in data centers (2) Energy performance in mechanical cooling equipment, cooling distribution equipment and heat rejection equipment. (3) Energy efficiency strategies by applying economizer cycles, airflow optimization and energy management.	Analyzing energy consumptions for air-conditioning system from 100 data centers. Some currently available energy efficiency strategies like economizer, airflow optimization, energy management and simulations tools are reviewed and summarized. The collected data from articles and reports showed that the range of air-conditioning system energy usage was 21% for the most efficient system and 61% for the least efficient system.
Shuja et al. [24]	(1) Computing systems including server architectures, power distribution, and cooling (2) Storage systems including storage architectures, technologies, and DRAM architectures (3) Communication systems including data center network architecture, optical data centers, and data center routing	Energy-efficiency techniques conflict with the business objectives and call for reduced redundancy of cloud resources that may lead to violation of user SLAs. Energy-efficient hardware devices such as SSDs and optical interconnects do not fit into current cloud paradigm due to their high cost. Furthermore, energy-efficient hardware devices such as ARM processors and PCM drives do not provide comparable performance to prevalent hardware technology.
Bhattacherjee et al. [25]	(1) Existing resource management techniques including resource scheduling, load balancing, and migration (2) Energy efficient resource management	Prediction models can be used to predict the future behaviors of the physical machines and the virtual machines by recording and analyzing their past behaviors.
Atiewi et al. [26]	(1) Virtual machine scheduling (2) Scheduling model in cloud data center (3) Scheduling parameters (4) Existing energy-efficient task scheduling algorithms	Most algorithms perform scheduling based on one or two parameters. A better scheduling algorithm can be developed from existing methods by adding more metrics, which can result in good performance and outputs that can be deployed in a cloud environment in the future.
Vasques et al. [7]	(1) Energy efficiency strategies of CPU, memory, disk, network interface, virtualization framework, uninterrupted power supply, and cooling (2) Demand response in data centers (3) Small and medium data center perspective	Studies to further develop strategies in small and medium data centers are needed, including more case studies, simulation, planning of polices, and overall evaluation to provide a clearer view of how the aforementioned actors can as a matter of fact benefit from each other, thereby promoting effective transformations in the energy market.
Reda and Nowroz [27]	(1) Techniques in power modeling and characterization for three computing substrates (2) The basic principles that govern power consumption in digital circuits (3) Utilize these principles to describe high-level power modeling techniques	Power consumption is contributed by two components: dynamic power and static power. Power estimation is highly challenging due to many modeling complexities and unknown factors during design time. Popular methods to adjust performance and power consumption include dynamic frequency and voltage scaling, nap modes, and sleep modes.
Zhang et al. [28]	(1) Cooling infrastructure of data centers includes air cooling, liquid cooling, and immersion cooling (2) Techniques of ICT and cooling management from chip level management to room level management (3) Models of ICT and cooling management including ICT workload, ICT power, heat transfer, ambient temperature, and cooling power (4) Optimization of ICT and cooling management including ICT control, cooling control optimization (5) Algorithms of workload scheduling including heuristic, static cooling, and dynamic cooling scheduling algorithms	Based on the available techniques of data center operation and previous work on data center management (i.e., modeling, optimization, algorithms and testbeds), the learning-based approach can be a promising framework for the joint ICT and cooling management in the data center, from data profiling, learning, optimization to execution. Equipment safely and hardware reliability are important for the performance and operational cost in the data center. In addition, the hybrid solution of cooling technologies and the integration of renewable can also be the future for enhancing the energy efficiency.
Mobius et al. [29]	(1) Estimation models' essential steps: model inputs and training model with benchmarks (2) CPU models (3) Virtual machine models (4) Server models	All of the models use either OS-provided metrics, PMC, or a mixture of these. Most of the models are trained and tested with standard benchmarks. A feedback system or a belief revision mechanism is more realistic to implement as is done in probabilistic parameter estimation.
Dayarathna et al. [30]	(1) Data center energy consumption (2) Digital circuit level energy consumption modeling (3) Server energy models (4) Processor power models (5) Memory and storage power models (6) Data centers level energy consumption modeling (7) Software energy models (8) Energy consumption modeling using machine learning	More than 200 power models were examined in this survey. While there has been a large number of studies conducted on the energy consumption modeling at lower levels of the data center hierarchy, much less work has been done at the higher levels. The accuracy, generality, and practicality of the majority of the power consumption models remain open.

consumption of data centers have been researched, few reviews have been conducted in this area. Most published reviews of data centers can be classified into two categories: thermal environment and energy efficiency. Table 1 summarizes the existing reviews of thermal environment, energy efficiency and power models for data centers. The reviews of the thermal environment mainly focus on air flow distribution, thermal managements, numerical simulation, airflow performance, and thermal optimization. Lu et al. surveyed the airflow distribution in data centers via ventilation configurations, underfloor plenum, and row- and rack-based solutions [8]. Alkharabsheh et al. [9] and Chu and Wang [10] summarized the existing thermal management techniques by considering the solutions and the air supplies, respectively. Rambo and Joshi overviewed data center modeling by considering model validation, rack-level compact modeling and dynamic modeling [11]. The survey conducted by Wang and Khan focused on the green performance metrics, thermal metrics, power/energy metrics of the data centers

[12]. Jin et al. surveyed the factors influencing the thermal environment, from room level to server level, as well as simulation software, airflow performance metrics and thermal optimization [13]. These reviews cover thermal environment design, optimization, control, and evaluation, but they do not consider the power consumption model at the component, server, or data center level.

The reviews of energy efficiency mainly focus on energy-efficient techniques cover component-level, server-lever, and room-level. For example, Ge et al. [14] conducted a survey of the power-saving techniques used in data centers and content delivery networks. Mittal [15,16] surveyed the techniques for managing the power consumption of embedded systems and data centers, including dynamic voltage and frequency scaling (DVFS)-based techniques, power-aware scheduling techniques, thermal-aware power management techniques, the use of power modes, energy savings in specific components, and the use of unconventional cores. Orgerie et al. [17] reviewed the techniques and solutions that aim to improve the energy efficiency of computing and wired network resources and discussed techniques that operate at the infrastructure level by coordinating the energy-saving capabilities of individual hardware and software components. Beloglazov et al. [18] created a taxonomy of energy-efficient design at the hardware level, operating system level, virtualization level and data center level. Bostoen et al. [19] summarized the power-reduction techniques for data-center storage systems. Wang et al. [20] overviewed the energy-saving techniques developed at the hardware, system software, and application levels, with an emphasis on energy-efficient computer servers and cluster systems. Similar work was also conducted by Hammadi and Mhamdi for data center networks [21]. Furthermore, Refs. [22] and [23] discussed the applicability of free cooling, which is an energy-efficient technology for cooling systems. Additionally, energy efficiency strategies, workload balance and demand response are the main focus for cluster servers and cloud data centers. Shuja et al. [24] discussed the conflict between maximizing service quality and minimizing energy consumption. Bhattacherjee et al. [25] conducted a review on energy resource management strategies, including resource scheduling, load balancing, and cloud migration. Vasques et al. [7] reviewed the energy efficiency strategies used for the CPU, memory, disk, network interface card, virtualization framework, uninterrupted power supply (UPS), and cooling. The demand response in data centers was also discussed. User task scheduling is another factor that plays a significant role in improving services. Atiewi et al. [26] presented a review of various energy-efficient task scheduling methods for cloud data centers. However, these reviews do not focus on power modeling.

Recently, some surveys of the energy consumption models of data centers have been conducted. Reda and Nowroz [27] described the main research directions in pre-silicon power modeling and post-silicon power characterization, focusing on chip/processor level power models. Zhang et al. [28] presented the techniques for enabling information and communications technology (ICT) and cooling management and the coupling models between them in data centers. Although this review combines the power models of IT equipment and cooling systems, it does not survey power consumption models of servers. Mobius et al. [29] provided a comprehensive survey of the proposed power estimation models by considering single-core and multicore processors, virtual machines, and the entire server. Moreover, Dayarathna et al. [30] conducted an in-depth study of the existing literature on power modeling from the hardware level to the data center level, covering more than 200 models. Although these two reviews cover the server power consumption models, none of them only focuses on the power consumption models of servers considering their power trends, categories, and applications. As mentioned in introduction, the energy consumption of servers is the basic unit of power flow and heat flow in data centers, and the power consumption models of servers are the basic consideration for the energy consumption estimation of data centers or that of the whole region. Therefore, the discussion of the power consumption models of servers are of great significance to the energy savings and emission reductions of data centers. The major contributions of this review are as follows.

- The server power trends are analyzed by considering IT development.
- The existing power consumption models of servers are reviewed by considering the calculation formula and other factors, and these models are compared in terms of advantages and disadvantages.
- The power consumption models of servers applied for energy prediction and management are reviewed to realize energy savings for server and cooling systems.
- The model parameters for model construction and the application scenarios of the power consumption models of servers are discussed.
- The future research directions for improving the energy efficiency and reducing the emissions of data centers are identified.

The review method is briefly introduced as follows. Studies that are directly or indirectly related to the power consumption models of servers in data centers, have been selected. These studies include journal papers, conference papers, white papers, handbooks and reports. Through preliminary searches, we compared databases and chose two that generated especially relevant results: Science Direct and IEEE. Combined searches have been conducted based on two or more of the following keywords: data center, server, power consumption, model, and cloud computing. Closely related studies were also checked.

3. Classification of servers and power consumption models

3.1. Classification of servers and power trends

3.1.1. Classification of servers

Server is a broad term describing a specific piece of IT equipment that provides computing capability and runs software applications in an environment networked with other IT equipment, including other servers. Most servers contain the following major hardware building blocks: processors, memory, chipset, input/output (I/O) devices, storage, peripherals, voltage regulators, power supplies, and cooling systems [31]. Servers can typically be divided into rack type, blade type, and tower type according to their configurations. A rack server, which is a server that is designed in according to uniform standards in appearance, can be used with cabinets. A blade server has a blade unit that can be inserted in a standard-height rack chassis; such servers are high-density and low-cost servers. All blades can be connected to provide a high-speed network environment, share resources and serve the same user group. The appearance and structure of a tower server are similar to those of a common computer, but its main board is expandable. Rack servers and blade servers are widely employed in data centers. According to the application scenarios, the servers can also be classified as database servers, file servers, mail servers, print servers, web servers, game servers, and application servers [32]. Furthermore, servers can be used for scientific and analytic purposes, business processing, visualization and audio, communications/telecommunications, and storage, and in cloud/Internet portal data centers, according to the workload, and the definitions and subsystem utilization are given in Table 2 [33]. It can be seen that the servers have different usage percentages for different components according to their work characteristics, which will result in differences in the power consumption of servers and should be considered in the establishment of a power consumption model.

3.1.2. Server power trends

The SPEC is a non-profit corporation formed to establish, maintain and endorse standardized benchmarks and tools for evaluating the performance and energy efficiency of the newest generation of computing systems [34]. Thus, the performance and power consumption data of servers with different configurations can be obtained from

Table 2
The workload types and common subsystem utilization per workload type [33]

Workload Type	Definition/Examples	CPU	Memory	I/O	Storage
Scientific	Scientific-based application including biological sciences, geosciences, weather forecasting, engineering, simulation, design, defense, security, and training of deep machine learning applications (versus run-time)	High	Medium-High	Low	Low
Analytics	Discrete data warehousing, data analysis, big data analytics, and run-time deep machine learning applications	High	Medium	Medium-High	Medium
Business Processing	Enterprise wide line of business applications that manage transactional, operational, and customer databases	Medium-High	Medium-High	Medium-High	Medium-High
Cloud/Internet Portal Data Center	Wikis, portals, social media, video-sharing websites, search engines, and online auction websites	Medium	Medium	Medium	Low
Visualization & Audio	Data center visualization applications including video processing, remote visualization, and audio processing	Medium	Medium-Low	Medium-Low	Medium-Low
Communications/Telco	Wired and wireless networking applications: application, control, packet, and signal processing	Medium-Low	Medium-Low	Medium	Low
Storage	Dedicated storage infrastructure and services including backup, tiering, and deduplication	Low	Low	Medium	High

different manufacturers. As of March 2019, there were 641 samples in SPEC; the number of samples of each year is given in Fig. 2. It can be seen that most (407) samples were from between 2008 and 2012, and the number of samples has also increased since 2018. Furthermore, 1U and 2U servers account for 53.8% of the samples, and the number of servers for each height is shown in Table 3. Servers with 1, 2, 4 and 8 CPUs make up 89.2% of the samples, and servers with 2 CPUs make up 62.6%. To describe the changes in server power, blade servers and other samples without nameplate power are not included, and the total number of valid samples is 508. In addition, to present the trend over time, all the samples are sorted by time, and the power at 100% is regarded as the peak power.

As shown in Fig. 3, the rated power of the servers hardly increased over time, which could be attributed to the full-load power far from the rated power. Fig. 4 depicts the relationships among rated power, peak power and idle power. It can be seen that the ratio of peak power to rated power (PRO) increases slowly. The mean PRO increases from approximately 40% to 50%. Conversely, the ratios of idle power to rated power (IRO) and idle power to peak power (IPO) decrease. The mean IRO and IPO decrease from 20% to 5% and from 55% to 15%, respectively. Moreover, the workload operation per second (ops) presents cliff growth for 1, 2, 4, 6 and 8 CPUs from 2014 (see Fig. 5), and the corresponding mean value is six times the previous value. Then, it can be concluded that servers are becoming efficient, as a heavier workload can be handled without consuming more energy. These changes indicate that the difference between the peak power and idle power of the servers is not constant from generation to generation. Hence, these trends should be reflected in the power consumption models.

3.2. Power consumption models of the server

According to recent studies, power consumption models are established for the following purposes: (1) to estimate the potential of power management; (2) to predict the power consumption of servers or data centers; and (3) to balance the needs of the data center (i.e., reducing energy consumption and providing high reliability). The power consumption models cover physical machines and virtual machines (VMs), and they can be classified as additive models, BA models, and other models by considering the calculation formula and other factors. A total of 47 models and their factors and benchmark/ testing tools are listed in Table 4.

3.2.1. Additive models

Fig. 6 depicts the power consumption composition of a server. It can be found that CPU has the largest contributor to power consumption, followed by peripheral slots (including network, I/O devices), power supply unit, memory, motherboard, disk/storage and fan [7]. Hence, the most direct way to model power consumption is to add the power contributions of all components. The expression is as follows:

$$P_{total} = P_{CPU} + P_{memory} + P_{disk} + \cdots + P_{I/O} \quad (1)$$

where P_{CPU}, P_{memory}, P_{disk}, and $P_{I/O}$ represent the CPU, memory, disk and I/O power consumption, respectively. Moreover, each component has its own power consumption model. For example, the power consumption of the CPU is a function of voltage, frequency and utilization, the power consumption of memory is a function of its size, frequency, and usage, the power consumption of the hard disk is a function of read rate and write rate, and the power consumption of the fan is a function of width, depth, and rotation speed. However, the main energy-consuming components can be slightly different owning to the configurations of servers, and the proportion of power consumption of each component is not constant, as shown in Fig. 7 [35]. Some research [36–38] selected the CPU, memory, hard disk, network, mainboard, and fan as the main components of the power consumption model, while other studies considered the CPU and memory to be the main

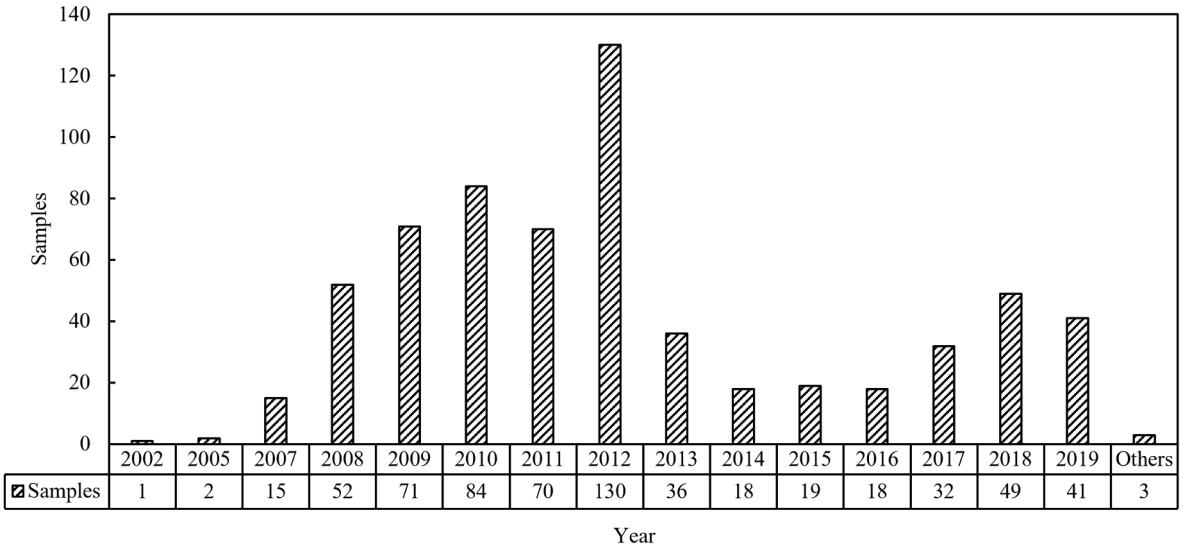

Fig. 2. The number of samples of each year.

Table 3
The number of each height of the servers for 641 samples.

Height	1U	2U	3U	4U	5U	7U	10U	12U	Blade	Others
Number	133	212	2	28	2	5	4	2	59	194

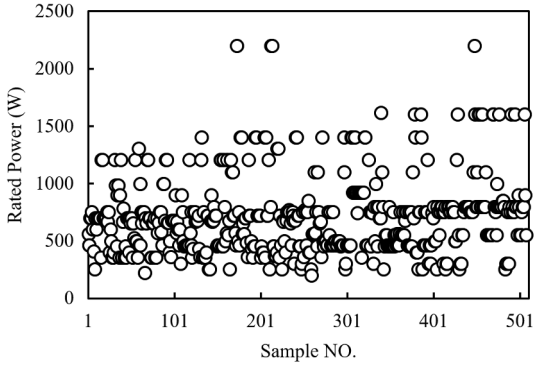

Fig. 3. The rated power of the servers.

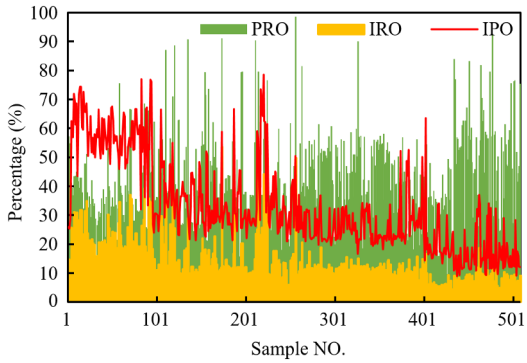

Fig. 4. The relationship among rated power, peak power, and idle power.

components, with the other components having a very narrow range or correlating well with CPU activity. Thus, the energy consumed by the server is a function of CPU, memory, and some other devices [39–42].

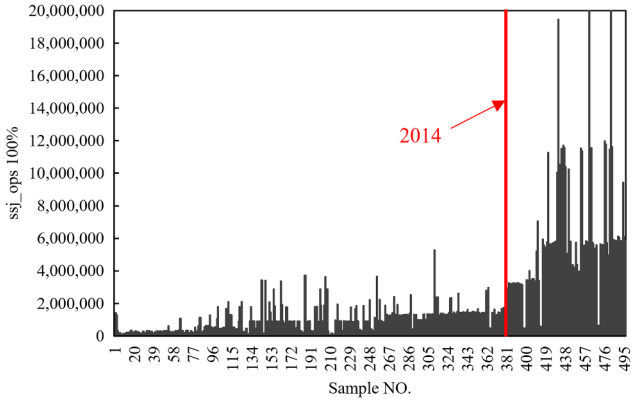

Fig. 5. The workload of servers for 1, 2, 4, 6 and 8 CPUs.

3.2.2. BA models

In data centers, the servers are not always active, as servers can be idle. Then, the power consumption of the counterparts can be divided into two parts. (1) Baseline power (P_{base}), which is the power consumption when the machine is idle, includes the power consumption of the fans, CPU, memory, I/O and other motherboard components in their idle state and is often considered as a fixed value. (2) Active power (P_{active}), which is the power consumption due to the workload, depends on the workload of the machine and the way it utilizes CPU, memory and I/O components. Hence, the power model can be expressed as the sum of baseline power and active power [43–45]. The formula is as follows:

$$P_{BA} = P_{base} + P_{active} + Delta \quad (2)$$

where, *Delta* is the correction term, which can be either a fixed value or an expression. In addition, the formula can be transformed. P_{base} can be regarded as a constant term, and P_{active} can be expressed as a function (e.g., linear-function, power function, high-degree polynomial). Hence, the BA model can be classified as linear regression models, power function models, non-linear model and polynomial model. In particular, from the perspective of heat generation and heat dissipation, the power consumption of the server includes the IT components and cooling components. Thus, the server power model can be transformed as their sum [46,47].

3.2.2.1. Linear regression model. Linear regression models include

Table 4
The power models and their factors and benchmark/testing tool.

Category	Authors	Power model	Factors	Benchmark/Testing tool	Error
Additive model	Basmadjian et al. [36]	$P_{Server} = \sum_{i=1}^{l} P_{Mainboard}$; (*for server type*) $P_{Server} = \sum_{i=1}^{l} P_{Mainboard} + \sum_{j=1}^{m} P_{Fan} + \sum_{k=1}^{n} P_{PSU}$; (*for tower or rack type*)	CPU voltage and frequency, memory frequency, hard disk read and write rate, and fan rotate speed	Lookbusy benchmark	2% for best case; < 10% for worst case
	Perumal and Subbiah [37]	$E_{node} = E_{CPU} + E_{memory} + E_{disk} + E_{NIC} + E_{mainboard}$	CPU frequency, voltages and percentage of active gates	Eucalyptus	
	Song et al. [38]	$E_{total} = E_{CPU} + E_{memory} + E_{disk} + E_{NIC} + E_{network-devices}$	Average component power	MPI shift benchmark	< 5%, max; < 3.5%, average
	Roy et al. [39]	$E = E_{CPU} + E_{Memory}$			
	Tudor and Teo [40]	$E = E_{CPU} + E_{Memory} + E_{I/O}$	Clock frequency, service time of one memory request, I/O service time	NPB 3.3, memslap 0.44, and PARSEC 2.1 benchmarks	~9%
	Ge et al. [41]	$P_{node} = P_{cpu} + P_{memory} + P_{other}$	L1 cache access, L2 cache access, main memory access, retired operations, and retired floating-point operations	NAS EP and FT benchmarks	
	Arroba et al. [42]	$P_{tot}(m, k) = P_{cpu}(m, k) + P_{mem}(m, k) + P_{others}(m, k)$		Lookbusy, modified RandMem, Web Search application from the CloudSuite benchmark suite, SPEC CPU2006 mcf and SPEC CPU2006 perlbench	
BA model	Dhiman et al. [43]	$P = P_{base} + P_{active}$	CPU utilization	Four benchmarks from the SPEC2000 suite, namely: mcf, gcc, mesa and gap	< 10%
	Xiao et al. [44]	$P_{server} = P_{static} + \sum_{i=1}^{M} P_i^{vm} = P_{static} + \sum_{j \in J} (k_j \cdot U_j)$	The utilization of each kind of physical components	401.bzip2, 429.mcf, TPC-W, Cachebench, and IOZone benchmark	< 5.2%
	Chen et al. [45]	$P = P_{fix} + P_{var}$	Throughput	IOzone benchmark, Fibonacci sequences, JPetstore, and JMeter	
	Ham et al. [46]	$P_{server} = P_{IT} + P_{sfan}$; $P_{IT} = 1.566 \times 10^{-5} + 42.29u_{cpu} + 0.379T_{die} + 0.03002T_{die}^2$; $P_{sfan} = 0.0012\text{RPM} - 12 \times 10^{-8}RPM^2 + 28 \times 10^{-2}RPM^3$	CPU utilization and die temperature, and fan revolution speed		
	Garraghan et al. [47]	$P_{server} = P_{IT} + \sum_{1}^{m} mP_{pump} + \sum_{1}^{n} nP_{fan}$	CPU utilization, die temperature and fan speed	SPECPower ssj2008 benchmark	3%
Simple regression model	Fan et al. [49]	$P(u) = P_{idle} + (P_{busy} - P_{idle}) \cdot u$	CPU utilization		
	Kavanagh and Djemame [50]	$EU\_P_x = Host\_Idle + (Host\_P - Host\_Idle) \times \frac{EU\_Util_x}{\sum_{1}^{EU\_Count} EU\_Util_y}$	CPU utilization	Stress	
	Islam and Pierson [51]	$P_{S[t1,t2]} = P_{idles} + \frac{Conn_{S[t1,t2]}}{ConnMax_S}(P_{Maxs} - P_{idles})$	Server utilization		
	Jin et al. [53]	$P_n = P_n^0 + (P_{max} - P_n^0) \times r_n/r_n^*$; $P_v = P_v^0 + (P_{max} - P_v^0) \times r_v/r_v^*$	Goodput	π-based CPU/RAM intensive application, HTTP request benchmark	
	Gupta et al. [54]	$P(\lambda) = \frac{\lambda}{\mu} \times (P_{CPU} + P_{Other}) + \left(1 - \frac{\lambda}{\mu}\right) \times P_{Idle}$	Throughput and speedup	Standard M/M/1 queuing model	
	Lefurgy et al. [55]	$P(k) = \text{At}(k) + B$; $P(k+1) = P(k) + Ad(k)$	The performance state of the processors	SPEC CPU2000 and the P4MAX workload	

(*continued on next page*)

Table 4 (*continued*)

Category	Authors	Power model	Factors	Benchmark/Testing tool	Error
	Beloglazov et al. [56]	$P(u) = k \bullet P_{max} + (1-k) \bullet P_{max} \bullet u$	CPU utilization	Non Power-Aware and Single Threshold policy	
	Rezaei-Mayahi et al. [57]	$P_{srv} = \frac{\beta}{\alpha}(P_{peak} - P_{idle})(T_{srv} - T_{in}) + P_{idle}$	Server temperature and inlet air temperature		
	Wang et al. [58]	$P_{estimated} = P_{idle} + U_{CPU\%}(P_{100\%} - P_{idle}) + Delta(T)$; $\mathrm{Delta}(T) = a_0 + a_1 T + a_2 T^2$	CPU utilization, inlet air temperature	Server Efficiency Rating Tool	3.2%–4.6% (Mean absolute percentage error)
Multiple regression model	Economou et al. [59]	$P_{blade} = 14.45 + 0.236 \times u_{cpu} - (4.47E-8) \times u_{mem} + 0.00281 \times u_{disk} + (3.1E-8) \times u_{net}$; $P_{itanium} = 635.62 + 0.1108 \times u_{cpu} + (4.05E-7) \times u_{mem} + 0.00405 \times u_{disk} + 0 \times u_{net}$	Utilization of CPU, memory, disk, and network	SPECcpu2000 integer and floating-point benchmarks, SPECjbb2000, SPECweb2005, the *streams* benchmark, and matrix multiplication.	< 5%
	Alan et al. [60]	$P_t = C_{cpu,n} \times u_{cpu,t} + C_{memory} \times u_{memory,t} + C_{disk} \times u_{disk,t} + C_{nic} \times u_{nic,t}$	CPU utilization, memory access rate, hard disk I/O request rate, and network I/O request rate	Scp, rsync, ftp, bbcp, and gridftp	< 6% for fine-grained model
	Li et al. [61]	$P_{server} = \alpha \bullet \sum_{k=1}^{n} U_{CPU}(k) + \beta \bullet \sum_{k=1}^{n} U_{MEM}(k) + \gamma \bullet \sum_{k=1}^{n} U_{IO}(k) + n \bullet e + E_{baseline}$	CPU utilization, memory usage, and IO throughput	Hadoop including PI, Sudoku, Random, Writer, Sort, Word, and Count	< 4%
	Lent [62]	$\mathrm{P} = \mathrm{I} + \sum_{i=0}^{N-1} \alpha_N \rho_N(i) + \sum_{j=0}^{C-1} \alpha_C \rho_C(j) + \sum_{k=0}^{D-1} \alpha_D \rho_D(k) + \Psi_m\left(\sum_{j=0}^{C-1} \rho_C(j)\right) + \Psi_M \left(\sum_{j=0}^{C-1} \rho_C(j)\right)$	Utilization of core, disk or port subsystem	HTTPerf	< 2.5 W
	Kansal et al. [63]	$E_{sys} = \alpha_{cpu} u_{cpu} + \alpha_{mem} u_{men} + \alpha_{io} u_{disk} + \gamma$	CPU utilization, the number of LLC misses, and the number of bytes read and written	SPEC CPU 2006 suite and IOmeter benchmarks	< 5%
	Costa and Hlavacs [64]	$\mathrm{P} = P_0 + \sum_{i=1}^{I} \alpha_i Y_i + \sum_{j=1}^{J} \beta_j S_j$	Performance variables	SPEC Power	
	Witkowski et al. [65]	$\mathrm{y} = \beta_0 + \beta_1 f_1(x_1) + \beta_2 f_2(x_2) + \cdots + \beta_n f_n(x_n)$	Performance and CPU temperature		1%–4%
	Bohra and Chaudhary [66]	$P_{\{CPU,cache\}} = a_1 + a_2 p_{CPU} + a_3 p_{cache}$; $P_{\{DRAM,disk\}} = a_4 + a_5 p_{DRAM} + a_6 p_{disk}$; $P_{(total)} = \alpha P_{\{CPU,cache\}} + \beta P_{\{DRAM,disk\}}$	Performance counter values for CPU, cache and DRAM	NAS-NPB 2.4, Iozone, Bonnie++, BYTEmark, Cachebench, Dense matrix multiplication, Gcc	7% of average mean error;6% of median error
	Lewis et al. [67]	$E_{system} = \alpha_0(E_{proc} + E_{mem}) + \alpha_1 E_{em} + \alpha_2 E_{board} + \alpha_3 E_{hdd}$	Ambient temperature, CPU die temperature, system bus, and misses in the L2 cache	Eight benchmarks from the SPEC CPU2006 benchmarksuite: bzip2, cactusadm, gromac, lbm, leslie3d, mcf, omnetpp, and perlbench	< 4%
Power function model	Chen et al. [68]	$\mathrm{P} = P_{fixed} + P_f \times f^3$	CPU frequency	Real HTTP traces	
	Elnozahy et al. [69]	$\mathrm{P} = c_0 + c_1 \times f^3$	CPU frequency	Olympics98 and Finance	
	Tian et al. [70]	$P_i = \rho_i k_i \mu_i^{\alpha i} + P_i^*$	Server utilization and service rate		
	Yao et al. [71]	$\mathrm{P} = \frac{b_i(t)^{\alpha}}{A} + P_{idle}$	Server utilization	MapReduce jobs	
	Wu et al. [72]	$P_{vm}(u_p) = \alpha \bullet u_p^{\beta}$; $P_{vm}(u_v, n) = \alpha \bullet \left(\frac{n}{N}\right)^{\beta} \bullet u_v^{\beta}$	CPU utilization	SysBench, Nbench, Fhourstone + Dhrystone, Ubench	4.26% for VMs;0.88% for host server
	Lien and Bai [73]	$\mathrm{P} = P_{base} + (P_{max} - P_{base}) \alpha U^{\beta}$	CPU utilization	Windows Media Load Simulator	< 3%, hardware-revised method; < 6%, software-revised method; ~11%, filled-manually method

(*continued on next page*)

Table 4 (*continued*)

Category	Authors	Power model	Factors	Benchmark/Testing tool	Error
Non-linear model	Fan et al. [49]	$P(u) = P_{idle} + (P_{busy} - P_{idle}) \cdot (2u - u^r)$	CPU utilization		
Polynomial model	Zhang et al. [74]	$P_{total} = a + b \times R_{cpu}$; $P_{total} = a + b \times R_{cpu} + c \times R_{cpu}^2$; $P_{total} = a + b \times R_{cpu} + c \times R_{cpu}^2 + d \times R_{cpu}^3$	CPU utilization	SPECPower_ssj2008 benchmark	< 4%;
	Lin et al. [75]	$E = N \times \frac{P}{s} = \frac{N}{\eta_{EE}}$; $\eta_{EE}(u_{ser}) = \frac{s}{P} = \frac{c_0 + c_1 \cdot u_{ser} + c_2 \cdot u_{ser}^2}{d_0 + d_1 \cdot u_{ser} + d_2 \cdot u_{ser}^2}$	Server utilization	SPECpower_ssj2008	2.794% for power function model; 2.974% for quadratic model; 4.506% for linear model; 1.615% for polynomial model; 5.918% for linear model of fixed coefficient; 8.222% for power function model of fixed coefficient
Other model	Horvath and Skadron [76]	$P_i(f_i, U_i) = a_{i3} f_i U_i + a_{i2} f_i + a_{i1} U_i + a_{i0}$	Clock frequency and CPU utilization		1%- average error; < 4%-worst case
	Xu et al. [77]	$power_{total} = \sum_{i=1}^{n} (power_{cpu,i}) + \sum_{i=1}^{n} (power_{disk,i}) + power_{idle}$	CPU utilization, clock frequency, and disk utilization	SPEC2006 and PARSEC (*mcf*, *bzip2* and *canneal* are chosen to represent user tasks, and *freqmine* are chosen to represent the tasks run by cloud computing providers.)	6.7%
	Bai et al. [78]	$P_i = (a_1 + u_i a_2)(b_1 + b_2 T_{chip}^i)$	Server utilization and chip temperature		
	Enokido and Takizawa [79]	$E_t(\tau) = \begin{cases} NFE_t \text{ if } NC_t(\tau) \geq 1 \\ minE_t \text{ otherwise} \end{cases}$; $E_t(\tau) = \begin{cases} maxE_t \text{ if } NC_t(\tau) \geq M_t \\ \rho_t \cdot NC_t(\tau) + NFE_t \text{ if } 1 \leq NC_t(\tau) \leq M_t \\ minE_t \text{ otherwise} \end{cases}$				
	Li et al. [80]	$P_{norm} = \frac{P_{sys} - P_{idle}}{P_{busy} - P_{idle}}$; $h(U) = c_1 U^{c2} + c_3 U^{c4} + c_5$; $P_{norm}(U) = 1 - h(U)^{-1}$	CPU utilization	SAP ERP performance model	
	Lim et al. [81]	$E = E_{dynamic} + E_{idle} + E_{sleep} = P(s, 1.0)mr/s$	The average number of instructions, the predicted number of requests, CPU utilization, CPU speed	RUBiS and TPC-W benchmarks	< 10%
	Kataoka et al. [82]	$E_t(\tau) = minE_t + \gamma_t(\tau) \cdot [bC_t + \sum_{i=0}^{nc_t - 1} \alpha_{ti}(\tau) \cdot \{(cE_t + \beta_{ti}(\tau) \cdot tE_t)\}]$		Nanosleep, gettimeodday, and taskset	
	Enokido and Takizawa [83]	$PC_t(tn_t(\tau)) = \delta_t \cdot tn_t(\tau) + minE_t + \{(ac_t(\tau) \cdot cE_t) + (nv_t(\tau) \cdot vE_t) + (mv_t(\tau) \cdot mE_t)\}$			

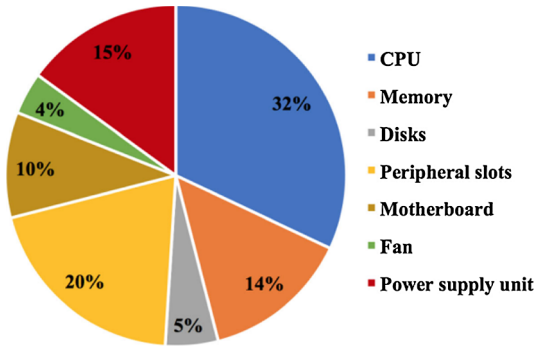

Fig. 6. The power consumption composition of a server [7]

simple regression models and multiple regression models. Bellosa [48] was the first to propose power models based on the correlation between power consumption and performance counters that captured activity across various functional units of the CPU. Fan et al. [49] proposed simple linear regression power models based on just CPU utilization to account for the system power as a whole. The model was validated by the experimental results. A similar model for VMs was also provided [50], and its formula has different independent variables for different application scenarios. For instance, regarding content distribution networks, the utilization ratio of the server during the time intervals is the variable of the model [51]. Similarly, throughput has a strong correlation with power [52]. For server virtualization, the ratio of the goodput to the maximum supportable goodput is regarded as the independent variable [53]. Gupta et al. [54]assumed that the server power consumption scales linearly with the ratio of the throughput to the processing speed. For feedback control loop, the model is a function of the performance state of the processors in the control period [55]. Furthermore, if the idle server consumes a constant ratio of the power consumed by the server running at the full CPU speed, then the idle power can be replaced by the maximum power multiplied by the ratio [56]. Rezaei-Mayahi et al. [57] adopted the simple linear regression model by considering the difference of CPU temperature and inlet air temperature to investigate the negative impact of ambient temperature on power consumption. The authors assumed that the CPU temperature is linearly related to server utilization, and the relationship between inlet temperature and power increment, which is reflected in the delta in function (2), was analyzed [58].

Although simple regression models based on CPU utilization are able to provide reasonable prediction accuracy for CPU-intensive workloads, they tend to be considerably inaccurate for predicting power consumption caused by I/O- and memory-intensive applications [43]. Hence, to make the model more universal, multivariate linear regression is used to build models between power consumption and the metrics of multiple components. Economou et al. [59] and Alan et al. [60] provided a power model as a function of the utilization of CPU, memory, disk, and network, and Li et al. [61] proposed a classified-piecewise ternary linear regression model to achieve more accurate predictive power. Lent [62] assumed that subsystems such as core, disk and port show linear power consumption with respect to their individual utilization. In contrast, Kansal et al. regarded CPU utilization, the number of LLC misses, and the number of bytes read and written as the training model parameters [63]. Costa and Hlavacs [64] described server power consumption based on performance counters (i.e., perf_cycles, perf_cache.references, perf_cache.misses). Furthermore, Witkowski et al. [65] added the CPU temperature in the model. VMeter is a power modeling technique based on the online monitoring of system resources and can predict the instantaneous power with an average mean and median accuracy of 93% and 94%, respectively [66]. Unlike these models, Lewis et al. [67] combined the ambient temperature, CPU die temperature, system bus, misses in the L2 cache together to regress the power model, and the accuracy is within 4%. These models tend to predict power consumption well as long as the characteristics of workload do not change.

3.2.2.2. Power function model. For some models, CPU is the main component of the energy consumption model. It operates at a frequency f and its dynamic power consumption is proportional to $V^2 \times f$, where V is the operating voltage; the frequency will be reduced when the voltage is lowered. The power consumption of all other components (except the CPU) is independent of the frequency, then, the power consumption of the server can be expressed as a cubic relationship with the operating frequency [68,69]. Furthermore, in cloud computing, the server utilization and service rate were regarded as the main components of the model by considering the tradeoff between energy cost and performance [70]. However, some studies used server utilization as the variable for characterizing the power model. The exponential value is equal to 3 [71]. Additionally, there is a power function relationship between the power consumption of servers and CPU utilization [72,73].

3.2.2.3. Non-linear model and polynomial model. In addition to simple linear regression, Fan et al. [49] proposed a non-linear model to fit the server power consumption. This model is more accurate than the linear model when estimating the total power usage of an individual machine. However, as mentioned above, the simple linear regression models and power function models are a poor fit for all but CPU-intensive servers and resort to multiple-input quadratics to provide an average absolute error of 9%. Considering that, Zhang et al. [74] used high-degree polynomial models to fit the server power consumption and found that the cubic polynomial can be the best choice. Similarly, Lin et al. [75] provided the relationship between power consumption and the second-order polynomial of server utilization.

3.2.3. Other models

Additive and BA models are the widely used power consumption models of servers, and other models have also been developed. Horvath and Skadron [76] used the CPU utilization, frequency and product to describe the power consumption of individual active machines. Similarly, Xu et al. [77] built a model with CPU utilization, the product of CPU utilization and frequency, and disk utilization. Bai et al. [78] used the product of the server utilization and chip temperature to represent the power consumption. Enokido and Takizawa [79] created simple power consumption for a server, where the electric power consumption

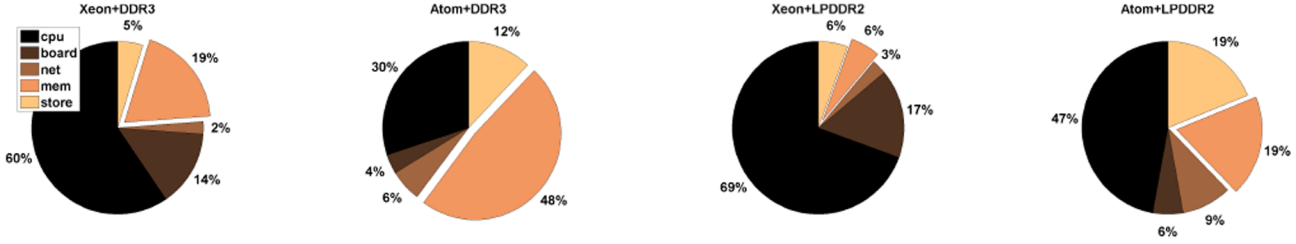

Fig. 7. Power breakdown across server components for different rack server configurations [35]

is maximally consumed if at least one application process is performed, and otherwise, the electric power consumption is minimal. In addition to the CPU and other devices, cooling devices such as CPU fans consume the electronic power. Ham et al. [46] considered the power consumption of servers to consist of IT and cooling power consumption. IT power consumption is a function of CPU utilization and die temperature, and the fan power is related to the fan revolution speed. Unlike the previously described approaches, Li et al. [80] developed a power consumption model through the normalization of system power instead of through direct modeling. Specifically, power capping behavior will occur when CPU utilization is higher than 80%. This behavior can be fitted well by a power function by using CPU power over the measured time [81]. The relationships among the energy consumed for processing requests, the idle period, and the sleep period were established. Moreover, power consumption models considering multi-core CPU [82] and the number of active VMs [83] were proposed, and the artificial neural network method was popularly used in power consumption models for cloud data centers [84].

3.3. Comparison of power consumption models of servers

Fig. 8 depicts the statistical result of the main variables in the existing power models. It can be seen that CPU utilization is the most widely used variable in the power consumption models, followed by CPU frequency, performance counters, and temperature. Different model components achieve different prediction results, and the accuracy will affect the prediction of the energy conservation and the choice of control strategy. Some studies have validated the accuracy of the models, and the errors are presented in Table 4. According to the statistics, the errors of additive models and BA models are < 9% and 11%, respectively. For additive models, the errors of most studies were < 5%, and the worst error rate was 10% in the worst case. For BA models, the errors of simple linear regression models, multivariate linear regression models, power function models, and polynomial models were generally < 4.6%, 7%, 6%, and 4%, respectively. Specifically, the accuracy of a linear regression model will be improved 13.6% by considering the inlet air temperature at 45 °C [58]. To analyze the accuracy of the models under the same conditions, Zhang et al. [74] compared the linear, quadratic, and cubic models and found that the cubic model can fit the measured data well. Lin et al. [75] analyzed the errors among six power models and concluded that the polynomial model has the lowest error (1.615%), followed by the power function model (2.794%) and the quadratic model (2.974%). However, if the coefficient is fixed, then the errors of the linear model and power function model can increase to 5.918% and 8.222%, respectively. Hence, it can be concluded that the polynomial model fits the power consumption of the servers best.

4. Model application

With the development of information technology, the power density and energy consumption of IT equipment are increasing. Meanwhile, the demands for cooling and the energy consumption of the cooling system are also increasing. Therefore, thermal management and energy management in data centers face enormous challenges. By using the server as the basic unit of energy flow and heat flow, the power model can be applied for energy prediction and management to reduce the amount of energy required by the server and cooling system.

4.1. Energy management of servers and server clusters

According to [85], saving 1 Watt of power at the CPU can easily turn into 1.5 W of savings due to power delivery efficiency losses inside the server and up to 3 W in the data center. Hence, considering the data center energy costs, reducing the energy consumption of servers is a major challenge. According to the review of the literature, from the last decade, the overall energy consumption of the IT equipment can be reduced by reducing the energy consumption of a single device or distributing the workload to the server clusters. To reduce the energy used by servers, a higher inlet air temperature will result in higher leakage power and fan power, and it is necessary to comprehensively evaluate this relationship. Some researchers have provided control policies based on the server power model by considering the optimal thermal environment and obtained the energy savings. Wang et al. [86] proposed a model-based approach that uses the power and temperature models, to create a multiple-input multiple-output fan controller and then handled the power optimization problem of blade servers. Compared with the feedback controller, this method can reduce fan power usage by approximately 20% and guarantee most temperatures between 64 °C and 66 °C. However, it cannot consider the influence of the temperature on the leakage power. Shin et al. [87] comprehensively considered the fan and leakage power. The results showed that tracking the energy-optimal steady-state temperature can reduce the overall energy by up to 17.6%. Moreover, Huang et al. [88] adjusted the server thermal setpoint and allowed the system to heat up when doing so saves more fan power than it costs in terms of leakage power. This thermal-aware power optimization will reduce the total power by 5.2% by increasing the inlet air temperature by 3 °C, and the thermal setpoint converged to 78 °C. Additionally, Zapater et al. [89,90] analyzed the relationships among computational power, temperature, leakage, and cooling power and proposed a cooling management policy that

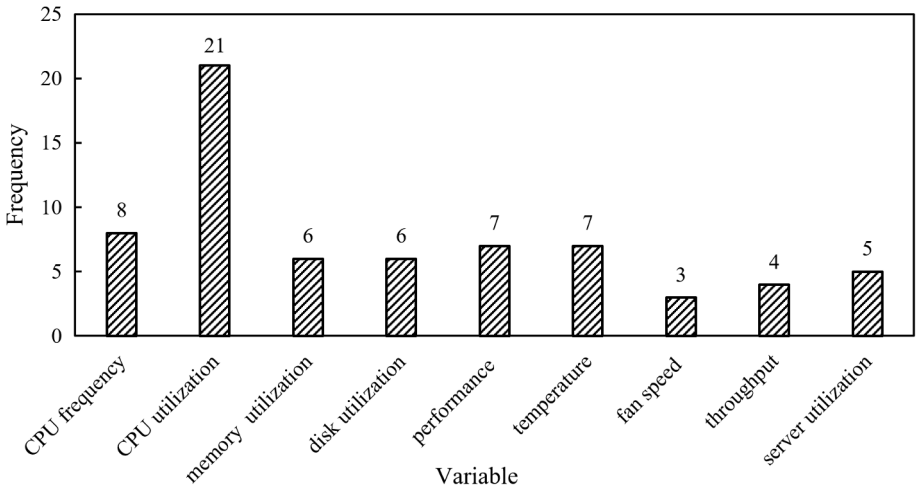

Fig. 8. The statistical result of the main variables in power models.

minimizes the server energy consumption by setting the optimal fan speed during operation. The energy savings of the whole cluster in terms of leakage plus fan power will increase with the increase in ambient temperature. The maximum savings are 10.3% at 32 ˚C.

Additionally, the server power model can be applied during server operation to guide server power mode (i.e., on/off), workload allocation and virtualization. Servers generally consume more than 50% of the peak power when they are idle [43]. However, the word “idle” can mean several things in practice: (1) the server must almost instantly react to a request for new work; (2) the server could pause for a brief period; (3) the server has plenty of time to wake up [91]. Although much energy can be saved by turning idle servers off, turn a server on requires some energy. The compromise is to put the idle server in sleep mode, which can save approximately 140 W of energy [92]. Virtualization technology allows servers, storage and networking devices to be shared, which improves resource utilization. In virtualization, applications can be easily migrated from one physical server to another [37]. Then, the managers can distribute the workload to the specified servers, and the others can be powered off or put in sleep mode by estimating the task needs. In this process, a suitable VM power model can help data center operators save power and price the VM energy consumption of cloud computing platforms [93]. Noguchi and Nishi compared the capacity of the servers and the summation capacity of the incoming requests to determine whether to start the servers [94]. Chen et al. [95] proposed server provisioning and load dispatching algorithms, which can reduce energy consumption by 30% without sacrificing user experiences. Based on a unified energy model and mechanisms for energy-aware resource accounting and allocation, a novel framework has been proposed for energy optimization [96]. Additionally, a joulemeter [63] is used to monitor actual power usage and to help determine how the migration of each VM will affect power level. This method can be used to realize significant power provisioning cost savings in virtualized data centers. Reference [45] further identified the relationship between energy consumption and runtime tasks under different cloud system configures and the correlation with cloud system performance. Elnozahy et al. [69] discussed the relationship between power consumption and CPU frequency and compared the energy efficiencies of five policies: independent voltage scaling, coordinated voltage scaling, vary-on vary-off, combined policy, and coordinated policy for server clusters. It can be found that the policy that saves the most energy is the coordinated policy, which achieves 33% and 50% of the cluster energy. Furthermore, Dhiman et al. presented a system for online power prediction using the Gaussian mixture model, and the average prediction error can be less than 10% [43]. Moreover, Lefurgy et al. proposed a feedback controller that uses precise, system-level power measurement to periodically select the state with the highest performance while keeping the system within a fixed power constraint [55]. Although there have been some attempts for the power/energy management of the server or server clusters, they are restricted to certain conditions. With the continuous improvement of energy-saving requirements, refined control/management technology based on the power consumption model has become an important direction for future research.

4.2. Cooling load and energy efficiency of cooling system

The matching of cooling supply and demand is key to saving energy of cooling system. According to the power consumption models of servers, the cooling capacity can be calculated accurately to ensure that IT equipment works in the required environment and avoid wasting the cooling capacity during design and operation. In the design of the cooling system, the cooling capacity depends on the cooling requirements of data center rooms, and the total cooling load can be calculated by considering various parameters, such as solar radiation through transparent surfaces, heat conduction through the building envelope, heat generated in the space by lights, occupants, and IT equipment, and heat transfer through air infiltration [97]. Among them, the heat dissipated by the IT equipment occupies the main part of the cooling load. The designers often take the nameplate power of IT equipment as its heat dissipation [98]. However, this estimation is seldom true in practice. As depicted in Fig. 3 and Fig. 4, the mean full-utilization power accounts for only approximately 50% of the nameplate power that has hardly changed recent years. Additionally, the servers do not operate at peak power all the time, and the average rack was only one-third filled. Hence, the thermal load will be significantly overestimated in the data center, and the cooling equipment are designed to be oversized according to the above design method [99]. At the same time, the power system has too much redundancy. In [100], the authors used the simple linear regression model [49] to model the power consumption of the data center, and they estimated the server’s maximum and idle power consumption by linear regression using 491 data points from various manufactures submitted to the SPECpower2008 database between 2007 and 2017. Giang et al. [101] and Pelley et al. [102] employed the linear model to investigate the total energy consumption of data centers. However, the power draw is not precisely linear during operation, and if possible, a more accurate representation of server utilization should be employed.

In addition, the models can be applied for cooling control or airflow management to realize energy savings. Zapater et al. [103] proposed the cooling control policy that set per-rack inlet temperature and airflow to minimize data center power based on the BA model and combined cooling and workload management. The results showed that this policy saves energy during winter (1.2%) and more energy in summer (14.4%). Noguchi and Nishi [94] presented an active controller shutter to prevent cold air from leaking through idle or shut-down servers when

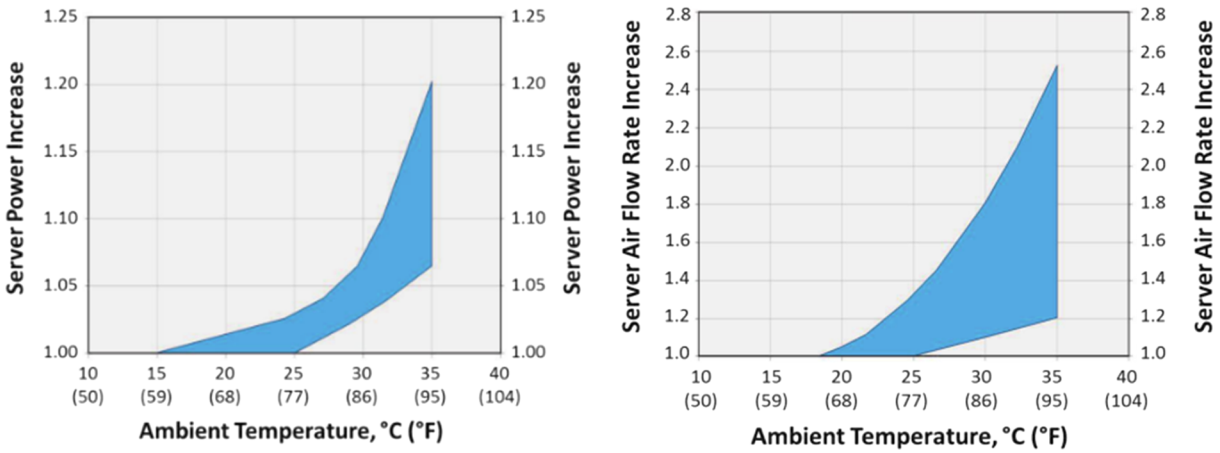

Fig. 9. Server power and airflow rate increase versus ambient temperature for Class A2 [104]

the server temperature is less than the threshold value (40 °C). Ham et al. [46] applied the server model to simulate the hourly cooling energy consumption by considering thermal characteristics. It can be found that cooling energy consumption will be increased when the supply air temperature is higher than 19 °C, owning to the increase in fan energy consumption.

5. Discussion and further work

5.1. Model

The physical architecture and configuration of the server differ very much among manufactures, and become more complicated with the innovation of technologies. The difference between peak power and idle power of the servers is not constant from generation to generation, as discussed in Section 3.1.2. At the same time, different workload types presented in Table 2 make the CPU, memory, I/O, and storage at different usage levels, which result in the power models less portable. Therefore, whether existing power consumption models accommodate these changes must be determined. Besides, the power increase of a server is a result of fan power, component power, and the power conversion of each. The component power increase is a result of an increase in the leakage current for some silicon devices. For Class A2, if the server inlet air temperature is raised from 15 °C to 35 °C, the server power would increase in the range from 7% to 20%, as shown in Fig. 9 [104]. At the same time, the required air volume is 1.2–2.5 times that at 15 °C, which can increase the power consumption of the server fan. This power increase may be ignored for server power consumption but not for the data centers that house tens of thousands of servers. Few studies [57,58] consider the inlet air temperature variation due to the fluctuating environment in the power consumption model of the server. In other words, most existing power consumption models are ineffective when the inlet air temperature of the server is changing. Hence, future models should reflect the coupling relationship between the power consumption and the thermal environment. Additionally, the power management of a server is a factor that influences model establishment. For instance, demand-based power management of Dell servers performs best in terms of energy-savings at mid-to-low processor utilization levels, which will cause the energy consumption of servers to be segmented. Then, the future models should be piecewise functions.

Table 5
Base idle state level power allowance [105]

Category	Maximum possible number of installed processors (#P)	Managed server	Base idle state power allowance, P_{BASE} (Watts)
A	1	No	47.0
B	1	Yes	57.0
C	2	No	92.0
D	2	Yes	142.0
Resilient	2	Yes	205.0

Note: A managed server must meet all of the following criteria: 1) is designed to be configured with redundant power supplies; and 2) contains an installed dedicated management controller. The Resilient category applies only to two socket systems that meet the definition of Resilient Servers as set forth in Appendix B of [105].

According to the statistics of the power consumption models of the servers, the simple linear regression is the most popular model, and the peak power and idle power are important. However, a method for easily acquiring these data is needed. For the peak power model, the power consumption is hypothesized to be strongly correlated with thermal design power (TDP), as the variation trend of TDP is the same as that of the peak power of 1U and 2U servers, as shown in Fig. 10. Energy Star divided the idle state power into two parts: base idle power and additional power, and the idle power limits are applicable to one and two socket systems only, which not include the blade and multi-node systems. The base idle power can be divided into five categories (A-D and Resilient) by considering the number of processors, redundant power supplies, and dedicated management controller. The additional idle power mainly considers additional components, such as power supplies, hard drives, memory, buffered DDR channel and I/O devices [105]. The relative values are shown in Table 5 and Table 6.

5.2. Application

A. Current problems for the applications of power consumption models

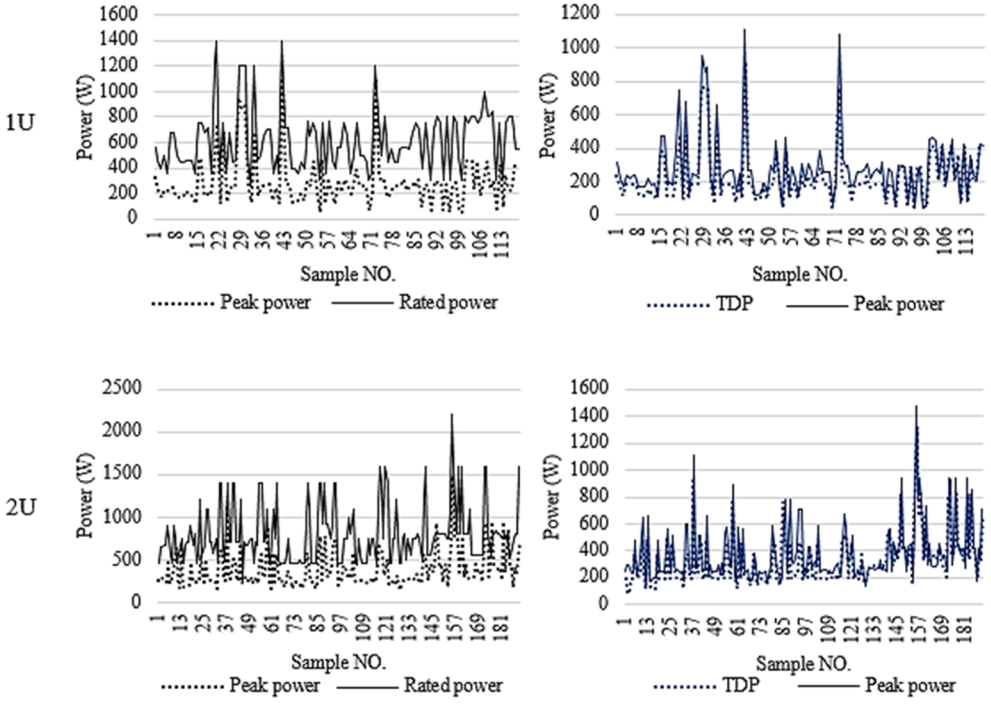

Fig. 10. The comparison among peak power, rated power and TDP for 1U and 2U servers.

Table 6
Additional idle power allowance [105]

System characteristic	Applies to:	Additional idle power allowance
Additional power supplies	Power supplies installed explicitly for power redundancy	20 W per power supply
Hard drives (including solid state drives)	Per installed hard drive	8.0 W per hard drive
Additional memory	Installed memory greater than 4 GB	0.75 W per GB
Additional buffered DDR channel	Installed buffered DDR channels greater than 8 channels (Resilient Servers only)	4.0 W per buffered DDR channel
Additional I/O devices	Installed devices greater than two ports of ≥1 Gbit, onboard Ethernet	< 1 Gbit: No Allowance = 1 Gbit:2.0 W/Active Port > 1 Gbit and < 10 Gbit: 4.0 W/Active Port ≥10 Gbit: 8.0 W/Active Port

Regarding the application of power consumption models, accuracy and simplicity are the two main requirements, but they are contradictory and restricted. If the model needs to be accurate, then the power consumption of each component of the server needs to be considered. In contrast, if the formula is simple and only considers the power consumption of the main components of the server, then it may not be so accurate. For example, additive models can explain the components of power consumption according to the configuration of the servers. However, this kind of power model needs to calculate the power consumption of each component of the servers, which is complicate, time-consuming, and hardly possible for heterogeneous systems. Furthermore, the simple linear regression is a popular model for predicting energy consumption. However, it is usually suitable for the CPU-dominated servers, especially under moderate utilization and constant power consumption. Once the systems become less CPU-dominated due to the demand of faster and larger memory and storage systems, the simple regression model introduces large prediction errors to these systems [106]. The multi-variable model facilitates monitoring and the collection of parameters at the server level, but it may not suitable for the system or room level depending on server type and workload. The regression coefficient should be adjusted when the number of servers changes. Otherwise, the accuracy of the model will be poor. Therefore, when using existing models for predicting or evaluating the power/energy consumption of servers or data centers, the applicability of them needs to be investigated, and the accuracy needs to be verified. The application scenarios, advantages and disadvantages of these models have presented in Table 7.

B. Power consumption model applications

As mentioned in Section 4.2, the existing cooling load calculation in the data center still follows the traditional calculation method. According to the working characteristics and server configurations, the servers hardly consume the rated power; hence, if the value of nameplate power is regarded as that of heat dissipation, then the equipment will be designed to be oversized, resulting in low equipment efficiency and significant power and cooling system redundancies, leading to the increase in the initial equipment investment. Due to this redundancy, the energy consumed by the cooling system can be offset by the energy savings of the servers. Therefore, the power consumption models of servers used to estimate the cold load is of great significance to the cooling design and energy saving of the data centers.

Furthermore, considering the safety of servers and data, many experiments have been carried out in the laboratory instead of in actual data centers. Hence, the server simulators should determine the heat dissipation characteristics based on the power consumption model of the servers. Wang et al. [107] and Nelson [108] have used heating elements instead of actual servers to investigate the reliability of the power system and the thermal environment of data centers, which are similar to therml manikins that an in-line electric heat source replaces the real person, and the thermal conductivity of the material is basically the same as that of human skin [109]. However, the heat dissipated by these heating elements is always equal to the nameplate power. This simplification is attributed to the lack of comprehensive and accurate power consumption models of servers, which still limits the diversity of the experimental conditions. Therefore, the model-based server simulator needs to be developed.

In addition, with the continuous improvement of equipment efficiency, the energy savings of equipment will be maximized, and the energy conservation potential will be reflected in operational control. Regarding thermal management and energy conservation, computer room air conditioner (CRAC) power leads to cooler server inlet ambient air temperature, which makes the server cooling fans work less hard and thus reduces server fan power. Similarly, lower CRAC power leads to warmer ambient air temperature, which causes server fans to work harder and consume more IT power [88]. Hence, coupling the server fan and CRAC fan is key to the energy efficiency. The power prediction models provide a powerful criterion for controlling the operation of the fans.

5.3. Further work

Based on the aforementioned references, the power consumption models of the servers play a crucial role in the power/energy management and thermal management of the data center. However, some models are proposed and applied to estimate the potential of power management to reduce energy consumption, and further research on the models and their applications is necessary. The adaptability of the existing models requires further verification, and the piecewise power consumption model of servers needs to be investigated according to the server power trends, server configuration, energy-saving technologies, and inlet air temperature. Model-based server simulators that present the operational characteristics of servers have potential for the experimental study of the data center, which is important for the development of highly efficient equipment and the implementation of energy-saving strategies for data centers. For data centers, thermal safety, which is the premise of energy savings, is extremely important. Hence, regarding the energy savings, energy-aware and thermal-aware management based on the power consumption models of servers should be considered at the same time. Thus, the combination of thermal management and energy management and the relationship between CRAC fans and server fans should be studied in the future.

6. Conclusion

The power consumption model of servers plays a critical role in the thermal management and energy management of data centers. This paper reviewed server classification and power consumption trends, then summarized and classified the existing power consumption models of servers. A comparison of these models has been carried out. Additionally, the application scenarios of these models were investigated, and the outlook for model development to increase energy savings was discussed. Based on the presented overview, the following conclusions can be drawn:

(1) The rated power of the servers hardly increased over time, and the

Table 7
The application scenarios, advantages and disadvantages of the power models.

Power consumption model		Application scenarios	Advantages	Disadvantages
Additive models		Server power consumption and optimization	Precisely predict the power consumption of a server.	Various monitoring parameters and not suit for heterogeneous systems or whole data center.
BA models	Simple linear regression model, power function model, non-linear model and polynomial model	Power consumption estimation of CPU-dominated servers, medium utilization systems, cooling load calculation, and cloud computing management	Only monitor one parameter and they are suit for prediction for cooling load or energy consumption of data centers.	Perform large prediction error for less CPU-dominated systems, and limit to partial utilization regions and server types.
	Multiple linear regression model	Server power consumption and server power management	Few parameters need to be monitor, and suit for predicting energy consumption of a server or small data centers, and the prediction error is between simple linear regression and additive model.	Once the characteristic of components changed, the regression coefficient need to be adjusted.
Other models		Artificial neural network method was popularly used in power consumption models for cloud data centers		

peak power of servers accounts for approximately 40% to 50% of the nameplate power. The idle power decreased from 55% to 15% of the rated power due to energy-saving technologies. Furthermore, the processing performance of the server per Watt has been significantly increased. These changes and the workloads of servers need to be considered for the application of existing models and the development of new models.

(2) The existing power consumption models can be classified as additive models, BA models, and other models according to their calculation formula and other factors. Specifically, the BA models can be further divided into linear regression models, power function models, non-linear models and polynomial models. CPU utilization is the most common variable in the models, followed by CPU frequency and performance counters. However, these models are for a particular server under a fixed environment, and most of them are suitable for single-core systems. Considering simplicity, convenience and accuracy, whether at the server, system or room level, linear regression and polynomial models can be the best choice, and their errors are $< 7\%$ and 4%, respectively. Moreover, to establish a model, energy-efficient technologies and the inlet air temperature must be considered, and the piecewise function should be used.

(3) The current power consumption models are established to estimate the potential of energy management, to predict the power consumption of servers or data centers, and to balance energy consumption and reliability. These models have been applied in some schemes of the IT field and the refrigeration field, but more efforts are needed. While the accurate power consumption models of servers result in accurate cooling load calculation and avoid excessive redundancy, energy- and thermal-aware managements based on the model results in the best overall energy-savings. Additionally, the model-based server simulators improve the experimental study, as they closely represent the actual situation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The study is supported by the National Natural Science Foundation of China (51778080).

References

[1] CAICT. China Internet Industry Development Trend and Prosperity Index Report. 2018.
[2] CAICT. White Paper on AI Development - Industrial Application. 2018.
[3] www.199it.com/archives/670459.html.
[4] Cooling DC. China data center annual research report on cooling technology development 2018. China Architecture & Building Press; 2019.
[5] Arman Shehabi SS, Dale Sartor, Richard Brown, Magnus Herrlin United Stated Data Center Energy Usage Report. Ernest Orlando Lawrence Berkeley National Laboratory; 2016.
[6] Dai J, Ohadi MM, Das D, Pecht MG. Optimum cooling of data centers. Springer; 2014.
[7] Vasques TL, Moura P, Almeida A. A review on energy efficiency and demand response with focus on small and medium data centers. Energ Effic. 2019;12:1399–428.
[8] Lu HJ, Zhang ZB, Yang L. A review on airflow distribution and management in data center. Energ Build 2018;179:264–77.
[9] Alkharabsheh S, Fernandes J, Gebrehiwot B, Agonafer D, Ghose K, Ortega A, et al. A brief overview of recent developments in thermal management in data centers. J Electron Packag 2015;137.
[10] Chu WX, Wang CC. A review on airflow management in data centers. Appl Energ 2019;240:84–119.
[11] Rambo J, Joshi Y. Modeling of data center airflow and heat transfer: State of the art and future trends. Distributed Parallel Databases 2007;21:193–225.
[12] Wang LZ, Khan SU. Review of performance metrics for green data centers: a taxonomy study. J Supercomput. 2013;63:639–56.

[13] Jin C, Bai X, Yang C. Effects of airflow on the thermal environment and energy efficiency in raised-floor data centers: A review. Sci Total Environ 2019;695:133801.
[14] Ge C, Sun ZL, Wang N. A survey of power-saving techniques on data centers and content delivery networks. IEEE Commun Surv Tut 2013;15:1334–54.
[15] Mittal S. A survey of techniques for improving energy efficiency in embedded computing systems. Int J Comput Aid Eng Technol 2014;6:440.
[16] Power Mittal S. Management techniques for data centers: a survey. Eprint Arxiv 2014.
[17] Orgerie AC, De Assuncao MD, Lefevre L. A survey on techniques for improving the energy efficiency of large-scale distributed systems. Acm Comput Surv. 2014;46.
[18] Beloglazov A, Buyya R, Lee YC, Zomaya A. A taxonomy and survey of energy-efficient data centers and cloud computing systems. Adv Comput. 2011;82:47–111.
[19] Bostoen T, Mullender S, Berbers Y. Power-reduction techniques for data-center storage systems. Acm Comput Surv. 2013;45.
[20] Wang J, Feng L, Xue W. A review of energy efficiency technology in computer servers and cluster systems. 3rd International Conference on Computer Research and Development (ICCRD). 2011.
[21] Hammadi A, Mhamdi L. A survey on architectures and energy efficiency in data center networks. Comput Commun. 2014;40:1–21.
[22] Daraghmeh HM, Wang CC. A review of current status of free cooling in data-centers. Appl Therm Eng 2017;114:1224–39.
[23] Ni JC, Bai XL. A review of air conditioning energy performance in data centers. Renew Sust Energ Rev 2017;67:625–40.
[24] Shuja J, Bilal K, Madani SA, Othman M, Ranjan R, Balaji P, et al. Survey of techniques and architectures for designing energy-efficient data centers. Ieee Syst J 2016;10:507–19.
[25] Bhattacherjee S, Khatua S, Roy S. A review on energy efficient resource management strategies for cloud. Adv Intell Syst. 2017;568:3–15.
[26] Atiewi S, Yussof S, Ezanee M, Almiani M. A Review Energy-Efficient Task Scheduling Algorithms in Cloud Computing. 2016 Ieee Long Island Systems, Applications and Technology Conference (Lisat). 2016.
[27] Reda S, Nowroz AN. Power modeling and characterization of computing devices: a survey. Found Trends Electronic Des Automat 2012;6:121–216.
[28] Zhang WW, Wen YG, Wong YW, Toh KC, Chen CH. Towards joint optimization over ICT and cooling systems in data centre: a survey. IEEE Commun Surv Tut. 2016;18:1596–616.
[29] Mobius C, Dargie W, Schill A. Power consumption estimation models for processors, virtual machines, and servers. IEEE T Parall Distr. 2014;25:1600–14.
[30] Dayarathna M, Wen YG, Fan R. Data center energy consumption modeling: a survey. Ieee Commun Surv Tut. 2016;18:732–94.
[31] ASHRAE. IT Equipment Design Impact on Data Center Solutions: W. Stephen Comstock; 2016.
[32] https://en.wiktionary.org/wiki/server.
[33] ASHRAE. IT Equipment Power Trends: W. Stephen Comstock; 2018.
[34] https://www.spec.org/.
[35] Malladi KT, Nothaft FA, Periyathambi K, Lee BC, Kozyrakis C, Horowitz M. Towards Energy-Proportional Datacenter Memory with Mobile DRAM. 2012 39th Annual International Symposium on Computer Architecture (Isca). 2012:37-48.
[36] Basmadjian R, Ali N, Niedermeier F, Meer Hd, Giuliani G. A methodology to predict the power consumption of servers in data centres. ACM Sigcomm International Conference on Energy-efficient Computing & Networking. 2011.
[37] Perumal V, Subbiah S. Power-conservative server consolidation based resource management in cloud. Int J Netw Manag. 2014;24:415–32.
[38] Song SL, Barker K, Kerbyson D. Unified performance and power modeling of scientific workloads. Proceedings of the 1st International Workshop on Energy Efficient Supercomputing. 2013.
[39] Roy S, Rudra A, Verma A. An energy complexity model for algorithms. Proceedings of the 4th conference on Innovations in Theoretical Computer Science. 2013.
[40] Tudor BM, Teo YM. On understanding the energy consumption of ARM-based multicore servers. ACM Sigmetrics/international Conference on Measurement and Modeling of Computer Systems. 2013.
[41] Ge R, Feng XZ, Cameron KW. Modeling and evaluating energy-performance efficiency of parallel processing on multicore based power aware systems. Int Parall Distrib, P. 2009. p. 1960.
[42] Arroba Patricia, Risco-Martín José L, Zapater Marina, Moya José M, Ayala José L, Olcoz Katzalin. Server power modeling for run-time energy optimization of cloud computing facilities. Energy Procedia 2014;62:401–10. https://doi.org/10.1016/j.egypro.2014.12.402.
[43] Dhiman G, Mihic K, Rosing T. A system for online power prediction in virtualized environments using gaussian mixture models. Des Aut Con 2010:807–12.
[44] Xiao P, Hu ZG, Liu DB, Yan GF, Qu XL. Virtual machine power measuring technique with bounded error in cloud environments. J Netw Comput Appl 2013;36:818–28.
[45] Chen F, Grundy J, Yang Y, Schneider J-G, He Q. Experimental analysis of task-based energy consumption in cloud computing systems. Proceedings of the ACM/SPEC international conference on International conference on performance engineering. 2013.
[46] Ham SW, Kim MH, Choi BN, Jeong JW. Simplified server model to simulate data center cooling energy consumption. Energ Build 2015;86:328–39.
[47] Garraghan P, Al-Anii Y, Summers J, Thompson H, Kapur N, Djemame K. A Unified model for holistic power usage in cloud datacenter servers. Int Conf Util Cloud 2016:11–9.
[48] Bellosa F. The Benefits of Event–Driven Energy Accounting in Power-Sensitive Systems. 2000.
[49] Fan XB, Weber WD, Barroso LA. Power Provisioning for a Warehouse-sized Computer. Isca'07: 34th Annual International Symposium on Computer Architecture, Conference Proceedings. 2007. p. 13–23.
[50] Kavanagh R, Djemame K. Rapid and accurate energy models through calibration with IPMI and RAPL. Concurrency and Computation: Practice and Experience. 2019. p. 1–21.
[51] Islam Su, Pierson J-M. Evaluating energy consumption in CDN servers. ICT as Key Technology against Global Warming. 2012.
[52] Ghosh S, Chandrasekaran S, Chapman B. Statistical modeling of power/energy of scientific kernels on a multi-GPU system. 2013 International Green Computing Conference (Igcc). 2013.
[53] Jin YC, Wen YG, Chen QH, Zhu ZQ. An empirical investigation of the impact of server virtualization on energy efficiency for green data center. Comput J 2013;56:977–90.
[54] Gupta V, Nathuji R, Schwan K. An analysis of power reduction in datacenters using heterogeneous chip multiprocessors. ACM SIGMETRICS Performance Evaluation Review. 2011. p. 87.
[55] Lefurgy C, Wang X, Ware M. Server-level power control. international conference on autonomic computing. IEEE; 2007.
[56] Beloglazov A, Abawajy J, Buyya R. Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing. Future Gener Comp Sy. 2012;28:755–68.
[57] Rezaei-Mayahi M, Rezazad M, Sarbazi-Azad H. Temperature-aware power consumption modeling in Hyperscale cloud data centers. Future Generat Comput Syst Int J Esci 2019;94:130–9.
[58] Wang Y, Nörtershäuser D, Masson SL, Menaud J-M. An Empirical Study of Power Characterization Approaches for Servers. ENERGY 2019 - The Ninth International Conference on Smart Grids, Green Communications and IT Energy-aware Technologies. Athens, Greece2019. p. 1-6.
[59] Economou D, Rivoire S, Kozyrakis C, Ranganathan P. Full-System Power Analysis and Modeling for Server Environments. Workshop on Modeling Benchmarking and Simulation. 2006.
[60] Alan I, Arslan E, Kosar T. Energy-aware data transfer tuning. Ieee Acm Int Symp. 2014:626–34.
[61] Li YF, Wang Y, Yin B, Guan L. An online power metering model for cloud environment. 11th IEEE International Symposium on Network Computing and Applications (Nca). 2012. 2012. p. 175–80.
[62] Lent R. A model for network server performance and power consumption. Sustain Comput Informat Syst 2013;3:80–93.
[63] Kansal A, Zhao F, Liu J, Kothari N, Bhattacharya AA. Virtual machine power metering and provisioning. Proc of the 1st ACM symposium on Cloud Computing (SOCC'10). 2010.
[64] Costa GD, Hlavacs H. Methodology of measurement for energy consumption of applications. IEEE/ACM International Conference on Grid Computing: IEEE. 2010.
[65] Witkowski M, Oleksiak A, Piontek T, Weglarz J. Practical power consumption estimation for real life HPC applications. Future Generat Comput Syst Int J Esci 2013;29:208–17.
[66] Bohra AEH. Chaudhary V. VMeter: Power modelling for virtualized clouds. IEEE International Symposium on Parallel & Distributed Processing: IEEE. 2010. p. 1–8.
[67] Lewis A, Ghosh S, Tzeng N-F. Run-time energy consumption estimation based on workload in server systems. Workshop on Power Aware Computing & Systems. 2008.
[68] Chen Y, Das A, Qin W, Sivasubramaniam A, Wang Q, Gautam N. Managing server energy and operational costs in hosting centers. ACM SIGMETRICS Performance Evaluation Review. 2005;33:303.
[69] Elnozahy EN, Kistler M, Rajamony R. Energy-efficient server clusters. Power-Aware Computer Syst 2003;2325:179–96.
[70] Tian Y, Lin C, Li KQ. Managing performance and power consumption tradeoff for multiple heterogeneous servers in cloud computing. Cluster Comput 2014;17:943–55.
[71] Yao Y, Huang LB, Sharma A, Golubchik L, Data Neely M. Centers power reduction: a two time scale approach for delay tolerant workloads. IEEE Infocom Ser. 2012:1431–9.
[72] Wu W, Lin W, Peng Z. An intelligent power consumption model for virtual machines under CPU-intensive workload in cloud environment. Soft Comput 2016.
[73] Lien CH, Bai YW, Lin MB. Estimation by software for the power consumption of streaming-media servers. IEEE T Instrum Meas 2007;56:1859–70.
[74] Zhang X, Lu JJ, Qin X, Zhao XN. A high-level energy consumption model for heterogeneous data centers. Simul Model Pract Th 2013;39:41–55.
[75] Lin WW, Wang WQ, Wu WT, Pang XW, Liu B, Zhang Y. A heuristic task scheduling algorithm based on server power efficiency model in cloud environments. Sustain Comput Informat Syst 2018;20:56–65.
[76] Horvath T, Skadron K. Multi-mode Energy Management for Multi-tier Server Clusters. Pact' 08: Proceedings of the Seventeenth International Conference on Parallel Architectures and Compilation Techniques. 2008. p. 270–9.
[77] Xu X, Teramoto K, Morales A, Huang HH. DUAL: reliability-aware power management in data centers. proceedings of the 2013. 13th Ieee/Acm International Symposium on Cluster, Cloud and Grid Computing (Ccgrid 2013). 2013. p. 530–45.
[78] Bai Y, Gu LJ, Qi X. Comparative study of energy performance between chip and inlet temperature-aware workload allocation in air-cooled data center. Energies 2018;11.
[79] Enokido T, Takizawa M. An extended power consumption model for distributed applications. Int Con Adv Info Net. 2012:912–9.
[80] Li H, Casale G, Ellahi T. SLA-driven planning and optimization of enterprise applications. Proceedings of the first joint WOSP/SIPEW International Conference on

Performance Engineering. San Jose, California, USA2010.
[81] Lim SH, Sharma B, Tak BC, Das CR. A dynamic energy management scheme for multi-tier data centers. Int Sym Perform Anal. 2011:257–66.
[82] Kataoka H, Duolikun D, Enokido T, Power Takizawa M. Consumption and computation models of a server with a multi-core CPU and experiments. IEEE International Conference on Advanced Information Networking and Applications Workshops. 2015.
[83] Enokido T, Takizawa M. Power consumption model of a server to perform communication type application processes on virtual machines. 2015 10th International Conference on Broadband and Wireless Computing, Communication and Applications. 2015.
[84] Lin W, Wu G, Wang X, Li K. An artificial neural network approach to power consumption model construction for servers in cloud data centers. IEEE Trans Sustain Comput 2019.
[85] Gough G, Steiner I, Saunders WA. Energy efficient servers-blueprints for data center optimization. Apress; 2015.
[86] Wang ZK, Bash C, Tolia N, Marwah M, Zhu XY, Ranganathan P. Optimal Fan Speed Control for Thermal Management of Servers. Ipack 2009: Proceedings of the Asme Interpack Conference 2009, Vol 2. 2010:709-19.
[87] Shin D, Kim J, Chang N, Choi J, Chung SW, Chung E-Y. Energy-Optimal Dynamic Thermal Management for Green Computing. 2009 International Conference on Computer-Aided Design (ICCAD'09). San Jose, CA, USA: ACM; 2009.
[88] Huang W, Allen-Ware M, Carter JB, Elnozahy E, Hamann H, Keller T, et al. TAPO: thermal-aware power optimization techniques for servers and data centers. Green Computing Conference and Workshops (IGCC): IEEE. 2011.
[89] Zapater M, Ayala JL, Moya JM, Vaidyanathan K, Gross K, Coskun AK. Leakage and temperature aware server control for improving energy efficiency in data centers. Des Aut Test Europe. 2013:266–9.
[90] Zapater M, Tuncer O, Ayala JL, Moya JM, Vaidyanathan K, Gross K, et al. Leakage-aware cooling management for improving server energy efficiency. Ieee T Parall Distr. 2015;26:2764–77.
[91] ASHRAE. Server Efficiency-metrics for computer server and storage: W. Stephen Comstock; 2015.
[92] Gandhi A, Gupta V, Harchol-Balter M, Kozuch MA. Optimality analysis of energy-performance trade-off for server farm management. Perform Evaluation. 2010;67:1155–71.
[93] Wen CJ, Xiang L, Yang Y, Ni F, Mu YF. System power model and virtual machine power metering for cloud computing pricing. Third International Conference on Intelligent System Design and Engineering Applications (Isdea). 2013. p. 1379–82.
[94] Noguchi T, Nishi H. Active controlled shutter for effective cooling of servers in data center. IEEE Ind Elec. 2015:1668–73.
[95] Chen G, He W, Liu J, Nath S, Rigas L, Xiao L, et al. Energy-aware server provisioning and load dispatching for connectionintensive internet services. 5th USENIX Symposium on Networked Systems Design & Implementation. San Francisco, CA. 2008.
[96] Stoess J, Lang C, Bellosa R. Energy management for hypervisor-based virtual machines. Usenix Association Proceedings of the 2007 Usenix Annual Technical Conference. 2007. p. 1–14.
[97] Hatamipour MS, Mahiyar H, Taheri M. Evaluation of existing cooling systems for reducing cooling power consumption. Energ Build 2007;39:105–12.
[98] Rasmussen N. Calculating Total Cooling Requirements for Data Centers. 2017.
[99] Mitchell-Jackson J, Koomey JG, Nordman B, Blazek A. Data center power requirements: measurements from Silicon Valley. Energy 2003;28:837–50.
[100] Cheung H, Wang SW, Zhuang CQ, Gu JF. A simplified power consumption model of information technology (IT) equipment in data centers for energy system real-time dynamic simulation. Appl Energ. 2018;222:329–42.
[101] Tran VG, Debusschere V, Bacha S. Data center energy consumption simulator from the servers to their cooling system. 2013 Ieee Grenoble Powertech (Powertech). 2013.
[102] Pelley S, Meisner D, Wenisch TF, VanGilder JW. Understanding and Abstracting Total Data Center Power. Workshop on Energy-Efficient Design. 2009.
[103] Zapater M, Turk A, Moya JM, Ayala JL, Coskun AK. Dynamic Workload and Cooling Management in High-Efficiency Data Centers. 2015 Sixth International Green Computing Conference and Sustainable Computing Conference (Igsc). 2015.
[104] ASHRAE. Thermal guidelines for data processing environment. fourth ed2015.
[105] Star E. ENERGY STAR® Program Requirements for Computer Servers-Partner Commitments V2.1. 2016.
[106] Chatzipapas A, Pediaditakis D, Rotsos C, Mancuso V, Crowcroft J, Moore AW. Challenge: resolving data center power bill disputes the energy-performance trade-offs of consolidation. 2015.
[107] Wang CH, Tsui YY, Wang CC. On cold-aisle containment of a container datacenter. Appl Therm Eng 2017;112:133–42.
[108] Nelson G. Development of an experimentally-validated compact model of a server rack. Georgia Institute of Technology; 2007.
[109] Tanabe S, Arens EA, Bauman FS, Zhang H, Nladsen TL. Evaluating thermal environments by using a thermal manikin with controlled skin surface temperature. Ashrae Trans 1994;100:39–48.