International Journal of Cloud Computing and Services Science (IJ-CLOSER)
Vol.3, No.3, June 2014, pp. 145~162
ISSN: 2089-3337

Energy Consumption in Cloud Computing Data Centers

Awada Uchechukwu, Keqiu Li, Yanming Shen
School of Computer Science and Technology
Dalian University of Technology, Dalian, 116024, China

Article Info

Article history:

Received Apr 18th, 2014
Revised May 25th, 2014
Accepted Jun 10th, 2014

Keyword:

Cloud computing
Green cloud environment
Energy efficiency
Resource management
Energy saving

ABSTRACT

The implementation of cloud computing has attracted computing as a utility and enables penetrative applications from scientific, consumer and business domains. However, this implementation faces tremendous energy consumption, carbon dioxide emission and associated costs concerns. With energy consumption becoming key issue for the operation and maintenance of cloud datacenters, cloud computing providers are becoming profoundly concerned. In this paper, we present formulations and solutions for Green Cloud Environments (GCE) to minimize its environmental impact and energy consumption under new models by considering static and dynamic portions of cloud components. Our proposed methodology captures cloud computing data centers and presents a generic model for them. To implement this objective, an in-depth knowledge of energy consumption patterns in cloud environment is necessary. We investigate energy consumption patterns and show that by applying suitable optimization policies directed through our energy consumption models, it is possible to save 20% of energy consumption in cloud data centers. Our research results can be integrated into cloud computing systems to monitor energy consumption and support static and dynamic system level-optimization.

Corresponding Author:

Keqiu Li,
Departement of Computer Science and Technology,
Dalian University of Technology,
No.2 Linggong Road, Ganjingzi District, Dalian City, Liaoning Province, P. R. C., 116024.
Email: awada@mail.dlut.edu.cn

1. INTRODUCTION

Recently, the emerging cloud computing offers new computing models where resources such as online applications, computing power, storage and network infrastructure can be shared as services through the internet [1]. The popular utility computing model adopted by most cloud computing providers (e.g., Amazon EC2, Rackspace) is inspiring features for customers whose demand on virtual resources vary with time. The wide scale potential of online banking, social networking, e-commerce, e-government, information processing and others, result in workloads of great range and vast scale. Meanwhile, computing and information processing capacity of several private corporation and public organizations ranging from transportation to banking and manufacturing to housing have been increasing speedily. Such a vast and vivid increase

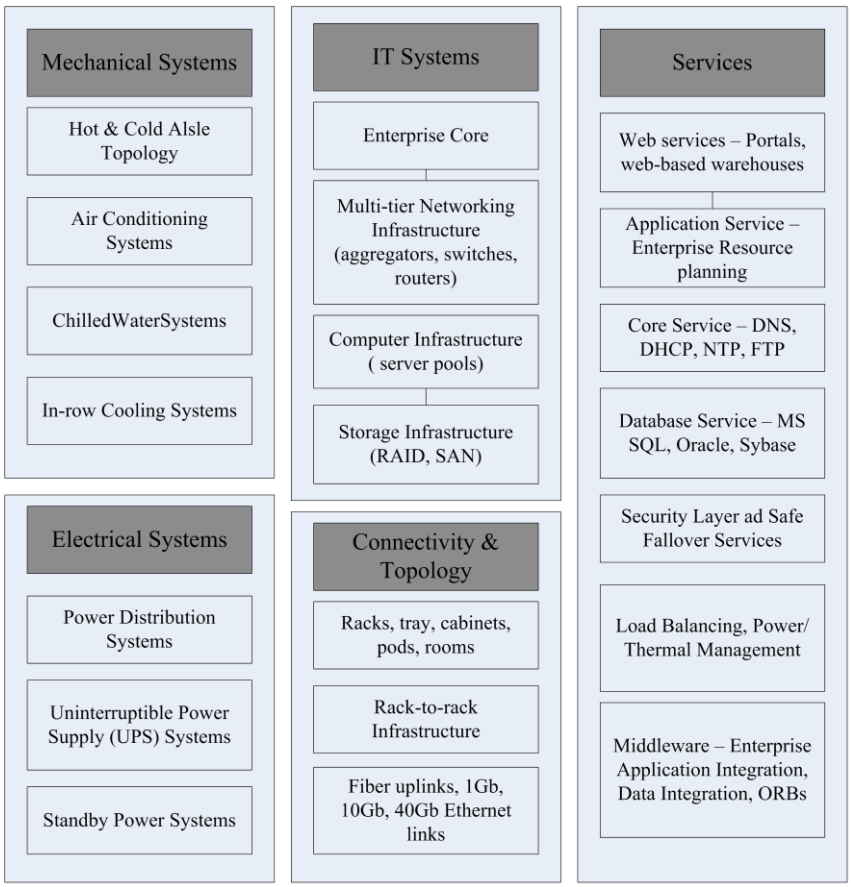


Fig. 1: High-level components comprising cloud computing environment

in the computing resources requires a scalable and efficient information technology (IT) infrastructure including servers, electrical grids, physical infrastructure, storage, network bandwidth, personnel and huge capital expenditure and operational cost. Cloud datacenters are the strength of today's demanding IT infrastructure, there is crucial need to improve its efficiency.

1.1 Energy-efficient cloud environment

As shown in Fig. 1, cloud computing environment is a large cyber-physical system consisting of electrical, mechanical and IT systems running a variety of tasks on a multitude of server pools and storage devices connected with multitier hierarchy of aggregators, routers and switches.

Energy consumption is the key concern in content distribution system and most distributed systems (Cloud systems). These demand an accumulation of networked computing resources from one or multiple providers on datacenters extending over the world. This consumption is censorious design parameter in modern datacenter and cloud computing systems. The power and energy consumed by the compute equipment and the connected cooling system is a major constituent of these energy cost and high carbon emission.

The energy consumption of date centers worldwide is estimated at 26GW corresponding to about 1.4% of worldwide electrical energy consumption with a growth rate of 12% per year [2], [3]. The Barcelona medium-size Supercomputing Center (a data center) pays an annual bill of about £1 million only for its energy consumption of 1.2 MV [4], which is equivalent to the power of 1, 200 houses [5]. Considering a U.S. Environmental Protection Agency (EPA) report to Congress [6], in which it is reported that U.S. datacenters consumed 61 billion kilowatt-hours of power in 2006, which constitutes 1.5% of all power consumed in the U.S. and represents a cost of $4.5 billions.

Table 1. Cloud Variables Definitions

Variable	Definition
E	Energy
U	Cloud utilization
Srv	Server
f	Fraction flow rate
SC	System configuration
I	Electric current
Q	CRAH heat
V	Voltage
T	Transmission rate
COP	Coefficient of performance
κ	Containment index
η	Loss coefficient
μ	Component utilization
ϖ	Flow rate
χ	Load distribution coefficient
β	Reduction factor
ρ	Heat

Electrical consumption of datacenters in the U.S., which hosts precisely 40% of the world's cloud datacenters servers, increased by approximately 40% during the financial breakdown [7], while energy consumed by servers, cooling, communication, storage, and power distribution equipment (PDU) accounts for between 1.7% and 2.2% [3]. This increased from 0.8% of U.S. energy consumption in 2000 and 1.5% in 2005 [8].

The environmental impact of cloud datacenters was estimated to be 116.2 million metric tons of CO_2 in 2006 [6]. Google datacenter used about 2.26 million MW hours of power to operate in 2010, resulting to carbon footprint of 1.46 million metric tons of carbon dioxide [9].The inter-government Panel on Climate change has called for the total reduction of 60%-80% by 2050 to avoid huge environmental damage.
Energy costs are the fastest-rising cost element in the data center portfolio, and yet data center managers are still not paying sufficient attention to the process of measuring, monitoring and modeling energy use in data centers.

1.2 Energy-inefficient cloud environment

Cloud computing environment comprises thousands to tens of thousands of server machines, working to render services to the clients [10], [11]. Present servers are far from energy uniformity. Servers consume 80% of the peak power even at 20% utilization [12]. The energy non-uniformity server is a key source to energy inefficiency in cloud computing environment. Servers are often utilized with between 10% to 50% of their peak load and servers experience frequent idle times [13]. This means that servers are not working at their optimal power-performance tradeoff points mostly, and idle mode of servers consumes big portion of overall power.

Another key contributor to power inefficiencies in cloud computing environment is the energy cost of Cooling and Air Conditioning Units (CACU), accounting to about 30% of the overall energy cost of cloud environment [14]. These values are reduced by introducing new cooling methods and new server and rack configurations for cloud computing environments. However, these values can also be reduced drastically for cloud datacenters located in good geographical locations so that they can benefit from ambient cooling. Yet, cooling energy consumption in cloud datacenters is still a major contributor to energy inefficiencies in cloud computing environments.

Yet another reason for energy inefficiency in cloud datacenters is the need for multiple power conversions in the power distribution system. Precisely, the main ac supply from the grid is first connected to dc so that it can be used to charge the battery backup system. The output of this electrical energy backup system then goes through an inverter to produce as power, which is then distributed throughout the cloud environment. These conversions are necessary due to the oversized and highly redundant uninterrupted power supply (UPS) modules, which are deployed for voltage regulation and power backup in cloud computing environment. However, most UPS modules in cloud datacenters operate at 10%-40% of their full capacity [15]. Unfortunately, the UPS conversion efficiency is quite low.

The power usage effectiveness (PUE), which explains how much power is lost in power distribution and conversion as well as in cooling and air conditioning in cloud computing environments, is calculated as the ratio of the total energy consumption in a cloud datacenter to the overall IT equipment power consumption [16]. The PUE metric has been steadily reducing over the last decade. In 2003, the PUE metric for a typical datacenter was estimated to be about 2.6 [17]. In 2010, Koomey estimated that the average PUE was between 1.83 and 1.92 [3]. Most recent cloud datacenters built by Google, Microsoft and Facebook have pushed PUEs under 1.2 or 1.1 [18, 19]. The cloud datacenter energy efficiency (CDCEE) metric may thus de defined as follows:

$$CDCEE = ITU \times ITE / PUE \quad (1)$$

where the IT utilization (ITU) denotes the ration of average IT use over the peak IT capacity in he cloud datacenter, and the IT efficiency (ITE) is the amount of useful IT work done per joule of energy.

1.3 Improving energy-efficiency in cloud computing environment

It is appropriate to attain energy proportionality at the server pool or cloud datacenter levels by dynamically shifting tasks among server and doing server consolidations so that the specific shape of the power dissipation versus utilization curve at the server level becomes less important, while the shape of the power-utilization curve at the cloud datacenter level becomes a line that goes through the origin [19]. Also, it has shown that energy-proportional operation can be attained for lightly utilized servers with full-system coordinated idle low-power modes [33]. Effects of using energy-proportional servers in datacenters are studied in [16]. The authors reported 50% energy consumption reduction by using energy-proportional servers with idle power of 10% of peak power instead if typical servers with 50% idle power consumption. The authors showed that increasing the energy efficiency of the disk, memory, network cards, and CPU helps in creating energy-proportional servers. Furthermore, dynamic power management (DPM) techniques, such as dynamic voltage scaling (DVS) and sleep mode for disk and CPU components, improve the energy proportionality of the servers.

High energy efficiency in cloud computing environments may be achieved by replacing traditional cloud datacenter equipment with more-powerful and energy-efficient state-of-the-art servers. These servers use more advanced internal cooling systems with less energy consumed by their fans. This is important because internal server energy consumption reductions are amplified by savings in the rack and cloud datacenter power distribution and cooling systems.

System-wide power management is an important key technique for improving energy efficiency in cloud computing environments. First, there is the total cost of ownership (TCO) for cloud computing environments, which includes the energy cost of operating a cloud datacenter. To minimize this cost, the cloud datacenter's overall power dissipation must be decreased. Secondly, there is the peak capacity of the power sources for cloud datacenters and electrical current limitations of the power delivery network in the cloud datacenter, which set a limit on the peak power draw at the server and datacenter levels.

Maximizing cooling efficiency is another way to lower the energy cost of cooling a cloud datacenter by deploying computer room air conditioning (CRAC) units and air handling units with demand-driven, variable frequency drive (VFD) fans within heat exchanges so as to match variable heat loads with variable airflow rates.

Finally, minimizing this energy consumption can result to conceal cost reduction. Apart from the enormous energy cost, heat released increases with higher power consumption, thus increases the probability of hardware system failures [20]. Minimizing the energy consumption has a momentous outcome on the total productivity, reliability and availability of the system. Therefore, minimizing this energy consumption does not only reduce the huge cost and improves system reliability, but also helps in protecting our natural environment. Thus, reducing the energy consumption of cloud computing system and data center is a challenge because data and computing application are growing in a rapid state that increasingly disks and larger servers are required to process them fast within the required period of time.

1.4 Paper overview and outline

To deal with this problem and certifying the future growth of cloud computing and data centers is maintainable in an energy-efficient manner, particularly with cloud resources to satisfy Quality of Service (QoS) requirement specified by users via Service Level Agreements (SLAs), reducing energy consumption is necessary. The main objective of this work is to present a new energy consumption models that gives detailed description on energy consumption in virtualized data centers so that cloud computing can be more environmental friendly and sustainable technology to drive scientific, commercial and technological advancements for the future.

The rest of the paper is organized as follows. Section II presents the related work on energy efficiency in cloud datacenter environments followed by the energy consumption pattern and formulas in Section III. Section IV formulates the energy consumption models for energy efficiency in cloud computing environments. The analysis of our energy consumption architecture is defined in Section V. Section VI presents the evaluation and implementation of our models. Finally, Section VII concludes the paper with discussion on the various issues and future research directions.

2. RELATED WORKS

Several issues about green ICT and energy reduction in modern cloud computing systems are receiving huge attention in the research community. Several other efforts have been made to build energy consumption models, develop energy-aware cost, manage workload fluctuation and try to achieve an efficient trade-off between system performance and energy cost. In [21] the authors proposed a cost model for calculating the total cost of utilization and ownership cost in cloud computing environments. They developed measurable metrics for this calculation. However, their calculation granularity is based on a single hardware component.

Energy management techniques in cloud environments have also been investigated in the past few years. In [22] described how servers can be turned ON/OFF using Dynamic Voltage/Frequency Scaling (DVFS) approach to adjust servers' power statues. DVFS adjust CPU energy consumption according to the workload. However, the scope is limited to the CPUs. Therefore, there is a need to look into the behavior of individual VMs. These can be possible by monitoring the energy profile of individual system components such as CPU, memory (at run time), disk and cache. Anne at al. [28] observed that nodes still consumed energy even when they are turned off, due to the card controllers embedded in the nodes which are used to wake up the remote nodes.

Sarji et al. [29] proposed two energy models for switching between the server's operational modes. They analyze the actual power measurement taken at the server's AC input, to determine the energy consumed in the idle state, the sleep state and the off state, to effect switching between this states. However, switching between power modes takes time and can translate to degraded performance if load goes up unexpectedly. Moreover, the set of servers that serve the load can also vary continually (a result of load balancing), leading to short idle times for most servers.

The power modeling techniques has been proposed by several authors. The power consumption model proposed by Buyya et al. [23] observed a correlation between the CPU energy utilization and the workload with time. Bohra et al. [24] also proposed a power consumption model that observed a correlation between the total system's power consumption and component utilization. The authors created a four-dimensional linear weighted power model for the total power consumption.

The work done by Chen et al. [25] treats a single task running in a cloud environment as the fundamental unit for energy profiling. With this technique, Chen and her colleagues observed that the total energy consumption of two tasks is not equivalent to the sum of individual consumed energy due to scheduling overhead. They created a power model for total energy consumption which focuses on storage, computation and communication resources.

On the other hand, several research effort has been also been made to minimize energy consumption in cloud environments mostly on virtualization. This technology permits one to overcome power clumsy by accommodating multiple VMs on a single physical host and by performing live migrations to optimize the utilization of the available resources. Yamini et al. [26] proposed a cloud virtualization as a potential way to reduce global warming and energy consumption. Their approach utilizes less number of servers instead of using multiple servers to offer service for multiple devices.

The power modeling techniques for the physical infrastructure (power and cooling systems) in data centers, proposed by Pelley et al. [34] is most relevant for us. They worked out first models which try to capture a data center at large.

In this paper, we provide tools to the Cloud computing environments to asses and reason about total energy consumption. Our approach target both Cloud data center simulation and analytic models. We show

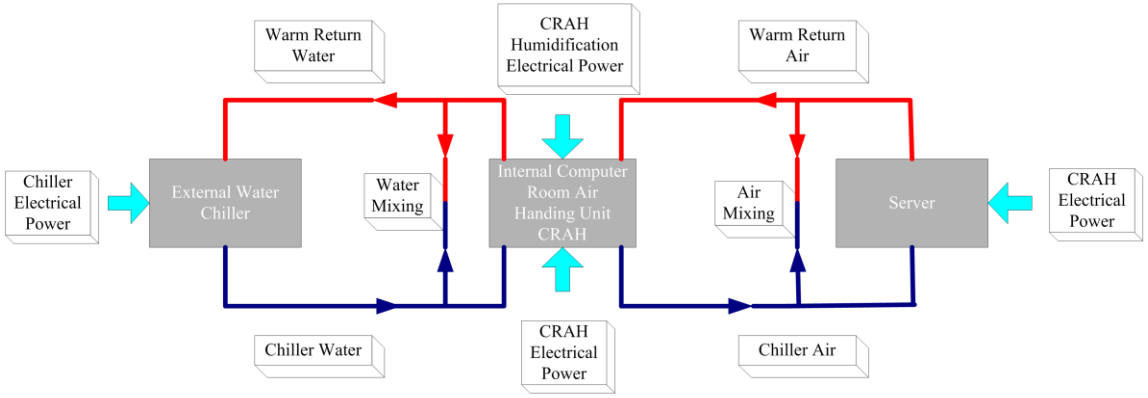


Fig. 2: Power and cooling flow of cloud environment

that by applying energy optimization policies through energy consumption models, it is possible to save huge amount of energy in cloud environments and data centers.

3. ENERGY CONSUMPTION PATTERN

The understanding of energy consumption pattern is necessary for its improvement. Servers consume a larger fraction of energy in cloud environments and their energy consumption varies with utilization. This consumption also vary with the type of computation going on in the server e.g., data retrieval and data processing.

Networking equipment, lightning and pumps also contribute to total energy consumption. However, the contribution of each totals a few percent of the overall consumption. Since these systems energy consumption does not vary significantly with data center load, we account for these systems by as a fixed energy overheads (around 6%) of the baseline power.

The power conditioning system supplies power to the uninterrupted power supplies (UPSs). The UPSs charge continuously and supply power until generators can start during a utility failure. UPS distribute electricity at high voltage (480V-1KV) to power distribution units (PDUs), which regulate voltage to match IT equipment requirements. The PDUs and UPSs consume significant amount of energy, and their consumption increases with workload.

The cooling system maintains humidity and air quality while it evacuates heat from the facility. This heat is the result of power dissipation. Removing this heat while maintaining humidity and air quality requires an extensive cooling system. Cooling starts with the computer room air handler (CRAH), which transfer heat servers' hot exhaust to a chilled water cooling loop while supplying cold air all over the facility. Extracting heat in this manner requires huge amount of energy. CRAH unit energy consumption dominates the total cooling system energy consumption and its requirement increases with both cloud environment thermal load and outside temperature. These overheads can be approximately modeled as a fixed figure of total energy consumed [30]. The power and cooling flow is presented in Fig. 2 and marketed by different industry segments which account for most of data center's energy consumption: (1) servers and storage systems, (2) power conditioning equipments, (3) cooling and humidification systems, (4) networking equipments, and (5) lighting /physical security. Thus, the first three sub-systems dominate power draw in cloud environments and data centers when in active mode.

However, the distribution of load in each sub-system can affect energy consumption, because of the non-linear interaction between sub-systems. Several research studies have proposed server energy consumption [8][9][10]. Whereas, the actual amount of utilization-energy varies, servers generally consume roughly half of their peak load energy when in idle mode, and energy consumption increases with resource utilization. Energy consumed during idle mode is a fixed part of the overall consumption. The dynamic part of energy consumption is the additional energy consumed by running tasks in the Cloud. Therefore, we divide energy consumption into two parts:

- Fixed energy consumption (energy consumed during server idle state)
- Dynamic energy consumption (energy consumed by Cloud tasks and cooling system)

3.1 Energy consumption formulas

The total energy consumption of an active server for a given time frame is the sum of energy consumed when the system is fixed and dynamic defined as ΔE_{Total} is formulated as follows:

$$\Delta E_{Total} = \Delta E_{Fix} + \Delta E_{Dyn} \quad (2)$$

There is additional energy is generated by scheduling overhead, denoted by E_{Sched} . This makes the energy consumption of two tasks, not equivalent to the sum of individual consumed energy. In this paper, we focus on the energy consumed by, server idle state, cooling systems, computation, storage and communication resource utilizations. These are defined as follows:

1. Energy consumption of server idle mode is denoted by E_{Idle}
2. Energy consumption of cooling system is denoted by E_{Cool}
3. Energy consumption of computation resources is denoted by E_{Compu}
4. Energy consumption of storage resources is denoted by E_{Store}
5. Energy consumption of communication resources is denoted by E_{Commu}

Therefore, the above formula (1) can be transformed into:

$$E_{Total} = (E_{Idle} + E_{Cool} + E_{Commu} + E_{Store} + E_{Compu}) + E_{Sched} \quad (3)$$

4. MODELING ENERGY CONSUMPTION

As discussed in the requirements, time variations in renewable energy availability and cloud computing environments efficiencies provide both opportunities and challenges for managing IT workload and cloud computing datacenters. In this section, we present novel models for energy efficiency aware management to improve the sustainability of cloud computing datacenters. In particular, we formulate measurable metrics based on runtime tasks to compare rationally the relation existing between energy consumption and cloud workload and computational tasks, as well as system performance.

Our models relate to overall cloud computing environments energy consumption to total utilization, represented by U . We represent the per-server utilization with μ_{Srv} , where the subscript denotes the server in question as follows:

$$U = \frac{1}{N_{Srv}} \sum_{Servers} \mu_{Srv[i]} \quad (4)$$

Table 2: Value of β_i for Intel Processors

Processor type	β_i
Intel Xeon dual-core E5502	0.942

Intel Xeon quad-core E5540	0.728
Intel Xeon hexa-core X5650	0.316

We characterize the individual server utilization, μ_{Srv}, as a function of U and a measure of task consolidation, χ, to abstract the effect of consolidation. χ is used to capture the degree of which load is distributed across cloud datacenter's servers. We define individual server utilization as:

$$\mu_{Srv} = \frac{U}{U + (1-U)\chi} \tag{5}$$

μ_{Srv} Only holds meaning for the $N_{Srv}(U + (1-U)\chi)$ servers that are non-idle. Fig. 3 depicts the relationship among μ_{Srv}, U, and χ, χ = 0 corresponds to perfect consolidation. The cloud's workload is packed onto the minimum number of servers, and the utilization of any active server is 1. χ = 1 represents the opposite extreme of perfect load balancing. All servers are active with $\mu_{Srv} = U$.

4.1 Modeling server idle state energy consumption

The idle energy consumption of a server can be determined by applying the following well known equation [38] from Joule's Ohm's law:

$$E = I \times V \tag{6}$$

where E denotes the energy or power (Watt), I represents the electric current (Ampere) and V indicates the voltage. The above equation can be adopted in order to determine the idle energy consumption at core level by assuming that each core contributes equally to the overall idle energy consumption of a processor:

$$E_i = I_i \times V_i \tag{7}$$

where E_i, I_i, V_i represents respectively the power, current and voltage of the corresponding core i. Furthermore, by analyzing the current I_i and voltage V_i relationship, we derive the following second order polynomial to model the current leakage as follows:

$$I_i = \alpha V_i^2 - \beta V_i + \gamma \tag{8}$$

where α = 0.114312 (($(V\Omega)^{-1}$), β = 0.22835 $(\Omega)^{-1}$ and γ = 0.139204 (V / Ω) are the coefficients [39].

With the implementation of energy saving mechanisms (e.g. AMD's PowerNow! and Cool 'n' Quite, Intel's SpeedStep), the idle energy consumption of a core (processor) decreases significantly. This is actualized by decreasing the Dynamic Voltage and Frequency Scaling (DVFS) of a core.

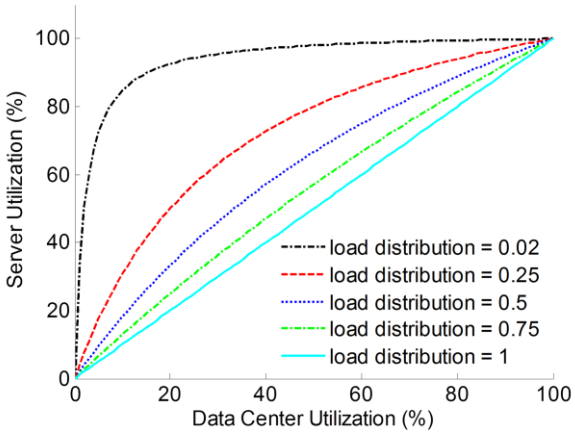


Fig. 3: Individual server utilization

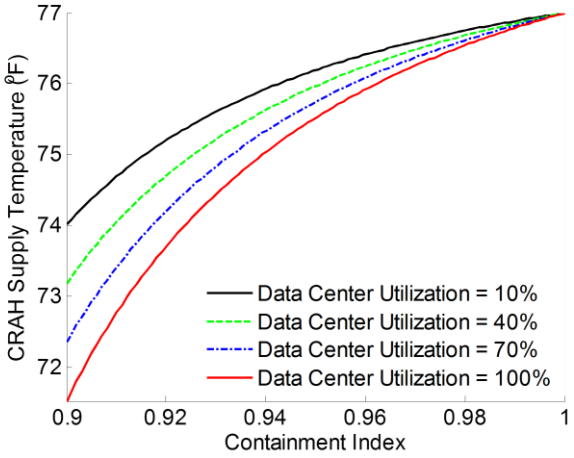


Fig. 4: CRAH supply temperature

$$E_{r_i} = \beta_i E_i \tag{9}$$

Where β_i is the factor for reduction in the power consumption E_i of core i, whereas E_{r_i} is the reduced power consumption of core i. However, β_i vary depending on the energy saving mechanism in use.

Given a processor of n numbers of cores with a specific energy saving mechanism, then its idle energy consumption is given by:

$$E_{Idle} = \sum_{i=1}^{n} E_{r_i} \tag{10}$$

The values of the reduction factor β_i for different types of Intel processors

4.2 Modeling cooling systems energy consumption

The CRAH unit energy consumption dominates the total cooling system energy consumption, it transfer heat servers' hot exhaust to a chilled water cooling loop while supplying cold air all over the facility. Its requirement increases with both cloud environment thermal load and outside temperature. The efficiency of cooling process varies on the speed of the air exiting the CRAH unit, the substance used in the chiller, etc.

In general, heat is transferred between two bodies according to the thermodynamic principle as follows:

$$\rho = \varpi C_{hc}(T_{ha} - T_{ca}) \tag{11}$$

Here ρ is the power transferred between a device and fluid ϖ represents the fluid mass flow, and C_{hc} is the specific heat capacity of the fluid. T_{ha} and T_{ca} represent the hot and cold temperatures respectively. The value of ϖ, T_{ha} and T_{ca} depend on the physical air flow throughout the data center and air recirculation.

Cloud datacenter designers use computational fluid dynamics (CFD) to model the complex flow and CRAHs to minimized recirculation. We replace CFD with simple parametric model that capture its effect on cloud computing energy consumption. Based on previous metrics for recirculation [35], [36], we introduce a containment index (κ). Containment index is defined as the fraction of air ingested by a server that is

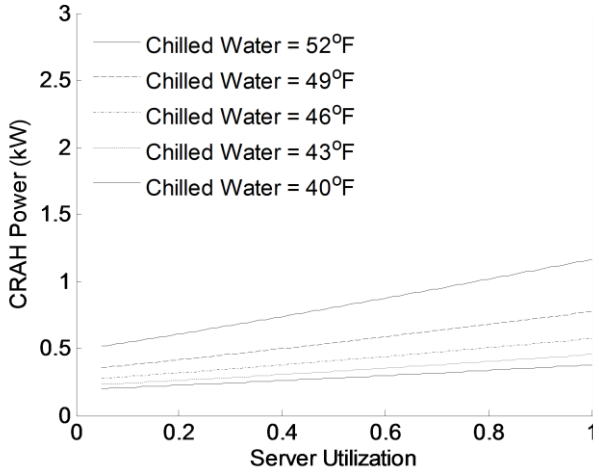


Fig. 5: CRAH power

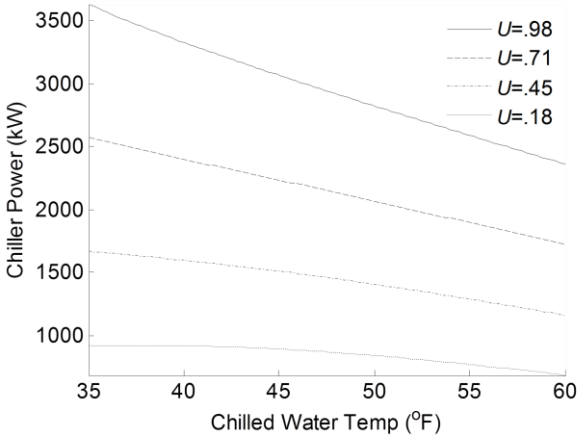


Fig. 6: Chilled Water Temperature

supplied by a CRAH. Thus, a κ of 1 implies no recirculation from the device. Our model uses a single, global containment index to represent average behavior, resulting as follows:

$$\rho = \kappa \varpi_{air} C_{hc_{air}} (Ta_{ha} - Ta_{ca}) \tag{12}$$

Here ρ is the heat transferred by the server or CRAH, ϖ_{air} represents the total air flowing through the device, Ta_{ha} the temperature of the air exhausted by the server, and Ta_{ca} is the temperature of the cold air supplied by the CRAH. CRAHs transfers heat out of the server room to the chilled water loop. Thus, we model equation (12) above using a modified effectiveness-NTU method [37]:

$$\rho_{CRAH} = E \kappa \varpi_{air} C_{hc_{air}} f^{0.7} (\kappa Ta_{ha} + (1-\kappa) Ta_{ca} - Twt) \tag{13}$$

ρ_{CRAH} is the heat removed by the CRAH. E is the transfer efficiency at the maximum mass flow rate (0 to 1), f represents the volume flow rate, and Twt the chilled water temperature.

The efficiency of a CRAH unit is measured using the Coefficient of Performance (COP), which is defined as the ratio of the amount of heat that is removed by the CRAH unit (Q) to the total amount of energy that is consumed in the CRAH unit to chill the air (E) [31]:

$$COP = Q / E \tag{14}$$

The COP of a CRAH unit varies by the temperature (T_s) of the cold air that it supplies to the cloud facility. The summation of energy consumed by the CRAH (E_{CRAH}) and IT (E_{IT}) equipments in cloud environment equal the total power dissipation [32]. The energy consumed by the CRAH unit may be specified as:

$$E_{CRAH} = \frac{E_{IT}}{COP(T_S)} \quad (15)$$

Energy consumed by the CRAH unit is dominated by fan power, which increases dynamically with the cube of mass flow rate (f) to some maximum amount. Additionally, some fixed energy is consumed by sensors and control system. Thus, the energy consumed by the CRAH unit totals it's fixed and dynamic activity:

$$E_{Cool} = E_{CRAH_{Idle}} + E_{CRAH_{Dyn}} f^3 \quad (16)$$

The efficient of heat exchange and the mass available to transfer heat increases as the volume flow rate through the CRAH increases. We use a single value of κ to simplify our model by allowing the conservation of air flow between the CRAH and servers. Fig. 4 demonstrates air recirculation places on the cooling system. The figure shows the CRAH supply temperature for typical maximum safe server inlet temperature of 77^0F. As κ decreases, the required CRAH supply temperature quickly drops. Thus, lowering supply temperature results in superlinear increases in CRAH and chiller plant power and preventing are recirculation can drastically improve cooling efficiency.

The effects of containment index and chilled water supply temperature on CRAH power are shown in Fig. 5. Here the CRAH model has a peak heat transfer efficiency of 0.5, a maximum airflow of 6900 CFM, peak fan power of 3KW, and idle energy consumption cost of 0.1KW. When the chilled water supply temperature is low, CRAH units are relatively insensitive to changes in containment index. For this reason, cloud computing operators often chosse low chilled water supply temperature, leading to overprovisioned cooling in the common case.

Chillers at a constant outside temperature and chilled water supply temperature will require energy that grows quadratically with the quantity of heat removed with utilization. The HVAC community has developed several modeling approaches to assess chiller performance. Although physics-based models do exist, we chose the Department of Energy's DOE2 chiller model [41]. Fitting the DOE2 model to a particular chiller requires numerous physical measurements. We use a benchmark set of regression curves provided by the California Energy Commission [42].

A chiller intended to remove 8MW of heat at peak load using 3,200 KW at a steady outside air temperature of 85^0F, a steady chiller water supply temperature of 45^0F, and a cloud load balancing coefficient of 0 will consume the following power as a function of total cloud utilization (KW):

$$E_{Chiller} = 742.8U^2 + 1{,}844.6U + 538.7 \quad (17)$$

Fig. 6 demonstrates the energy required to supply successively lower chilled water temperature at various U for an 8MW peak thermal load. However, as thernal load increases, the energy required to lower the chilled water temperature becomes substantial. The difference in chilled power for a 45^0F and 55^0F chilled water supply at peak load is nearly 500KW. Fig. 7 displays the rapidly growing energy reqiurement as cooling load increases for a 45^0F chilled water supply. The grah also shows the strong sensitivity of chiller energy consumption to outside air temperature.

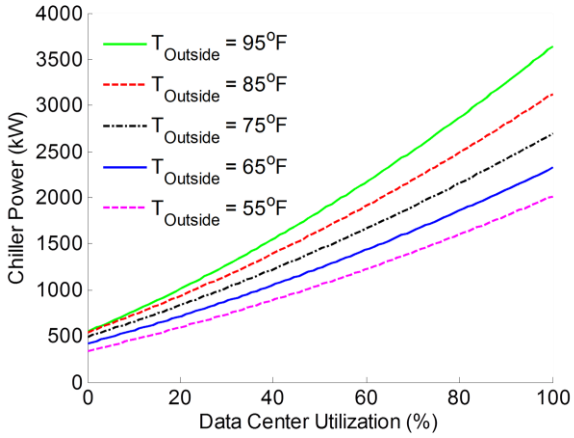


Fig. 7: Effects of U and $T_{Outside}$ on P Chiller

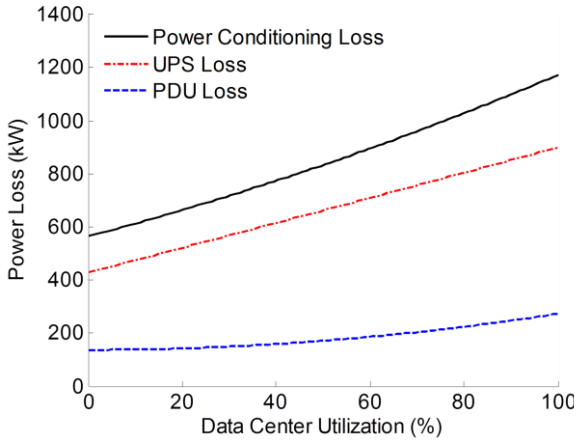


Fig. 8: Power conditioning losses

4.3 Modeling power conditioning systems energy Consumption

Cloud computing environments need considerable infrastructure simply to supply uninterrupted, stable electric power. Power distribution units transform the high voltage power supplied throughout the cloud environment to voltage levels appropriate for servers. They incur a constant energy loss as well as a energy loss proportional to the square of the load [40]:

$$E_{PDU} = E_{PDU_{Idle}} + \eta_{PDU}\left(\sum_{Servers} \mu_{Srv}\right)^2 \quad (18)$$

where E_{PDU} represents the energy consumed by the PDU, η_{PDU} denotes the PDU energy loss coefficient, and $E_{PDU_{Idle}}$ the PDU's idle energy consumption. UPSs provide temporary energy supply during utility failure. UPS systems are placed in series between the utility supply and PDUs and impose some energy consumption overhead even when operation on utility energy. UPS energy overheads follow the relation [40]:

$$E_{UPS} = E_{UPS_{Idle}} + \eta_{UPS} \sum_{PUDs} E_{PDU} \quad (19)$$

where η_{UPS} denotes the UPS loss coefficient. UPS losses about 9% of their input energy at full load.

Fig. 8 shows the power losses for a 10MW of cloud environment. At peak load, power conditioning loss is 12% of total server energy consumption. These losses result in additional heat that must be evacuated by the cooling system.

4.4 Modeling dynamic derver state energy consumption

The energy consumption of a task (communication, storage, computation) is determined by the number of processes, size of data to be processed, size of data to be transmitted and the system configuration (SC_i). The energy consumption profiling metrics are presented in Table 3. Thus, the energy consumed by each task can be formulated as:

Table 3: PROFILING METRICS

Task	ID	Energy Consumed	Process Number	Size of Data Processed	Size of Data Transmitted
Communication	m_i	Em_i	N_{m_i}	Dm_i	Tm_i
Storage	s_i	Es_i	Ns_i	Ds_i	Ts_i
Computation	c_i	Ec_i	Nc_i	Dc_i	Tc_i

$$Em_i = fm_i(Nm_i, Dm_i, Tm_i, SC_i) \quad (20)$$

$$Es_i = fs_i(Ns_i, Ds_i, Ts_i, SCi) \quad (21)$$

$$Ec_i = fc_i(Nc_i, Dc_i, Tc_i, SC_i) \quad (22)$$

The energy consumed by cloud tasks E_{Dyn} is formulated as follows:

$$E_{Dyn} = \sum_{i=1}^{n} Em_i + \sum_{i=1}^{i} Es_i + \sum_{i=1}^{j} Ec_i \quad (23)$$

Adding the energy generated by schedule overhead and interference, equation (8) above can be transformed as follows:

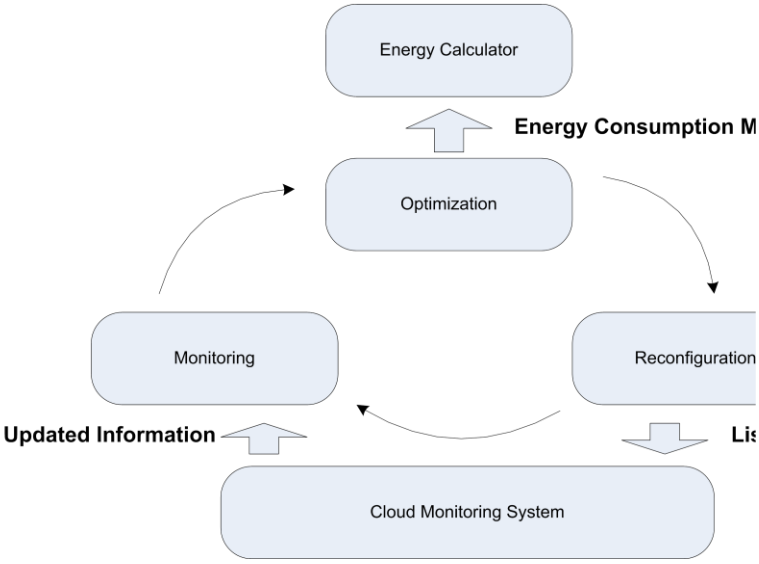
$$E_{Dyn} = \sum_{i=1}^{n} Em_i + \sum_{i=1}^{i} Es_i + \sum_{i=1}^{j} Ec_i + E_{Sched} \quad (24)$$

5. ENERGY CONSUMPTION ANALYSIS TOOL

The architecture of our energy-saving mechanism, presented in Fig. 9, is based on Optimization, Reconfiguration and Monitoring. The entire state of Cloud environment is automatic monitored. Another major contribution of this paper is committed to detailed analysis of the state of Cloud environments and data centers resources with relevant energy consumption attributes and interconnections.

This state is recurrently analyzed by the Optimization module in order to find a surrogate software application and service allocated configurations that enables energy minimization. Once an appropriate energy-saving configuration is detected, the loop is closed by issuing a set of action on Cloud environment to reconfigure the allocation of this energy-saving setup.

Monitoring and Reconfiguration modules communicate with the Cloud environment monitoring framework to perform their tasks. The Optimization module ranks the target configurations, this is established by applying energy-saving policies without violating existing SLAs, with respect to their energy consumption that are predicted by the Energy Calculator module. The accuracy predictions of this module is essential to take the most appropriate energy minimization decisions, it has the ability to forecast the energy consumption of Cloud environment after a possible reconfiguration option.



6. EVALUATION

In this section, we demonstrate the utility of our models. We provide a comparison of power requirement between several presumptive Cloud data centers using the energy consumption models [34, 43]. Each scenario based on the previous, present new power-saving attribute. We decompose each data center on 25% utilization. Next, we present the configured Cloud data centers as well as induced workload.

6.1 Scenario 1 and 2

These represent conventional Cloud data center with legacy physical infrastructure typical of facilities commissioned in the last three to five years. We use yearly average for outside air temperature (^{o}F). We assume limited server consolidation and a relatively poor containment index of 0.9. Furthermore, we assume typical (inefficient) servers with idle power at 60% of peak power, and static chilled water and CRAH air supply temperature set to 45^{o}F and 65^{o}F, respectively. We scale the Cloud date center such that the *Scenario 1* facility consumes precisely 10MV at peak utilization.

6.2 Scenario 3

This represents a data center where virtual machine consolidation or other mechanisms reduce χ from 0.26 to 0.57 and reducing the number of active servers from 81% to 44%. Improved consolidation drastically decreases aggregate power draw, but, paradoxically, it increases PUE. These results illustrate the shortcoming of PUE as a metric for energy efficiency; it fails to account for the inefficiency of IT equipment.

6.3 Scenario 4

This allows servers to idle at 5% of peak power by transitioning rapidly to a low power sleep state, reducing overall data center power by another 22% [33]. This approach and virtual machine consolidation take alternative approaches to target the same source of energy inefficiency: server idle power waste.

6.4 Scenario 5

This posits integrated, dynamic control of the cooling infrastructure. We assume an optimizer with global knowledge of data center load/environmental conditions that seek to minimize chiller power. The optimizer chooses the highest T_{wt} that still allows CRAHs to meet the maximum allowable server inlet temperature. This scenario demonstrates the potential for intelligent cooling management.

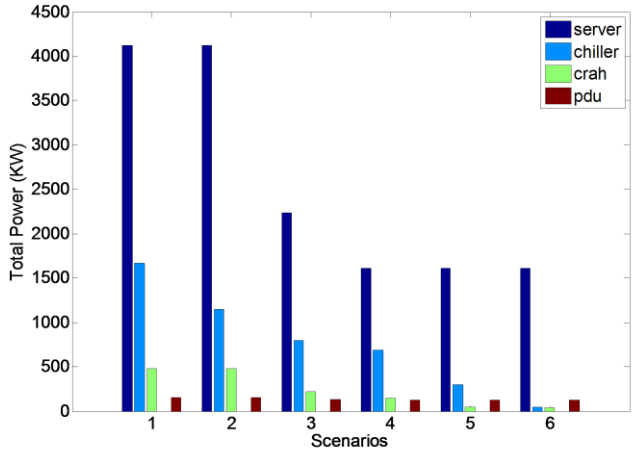


Fig. 10: Power saving features

Table 1: Presumptive cloud data centers

Cloud scenario	χ	κ	E_{Idle}/E_{Peak}	$T(F)$	Optimal cooling
1	0.95	0.9	0.6	70	No
2	0.93	0.9	0.6	50	No
3	0.25	0.9	0.6	50	No
4	0.25	0.9	0.05	50	No
5	0.25	0.9	0.05	50	Yes
6	0.25	0.99	0.5	50	Yes

6.5 Scenario 6

Finally, this represents a data center with a containment system (e.g., servers enclosed in shipping containers), where containment index is increased to 0.99. Under this scenario, the cooling system power draw is drastically reduced and power conditioning infrastructure becomes the limiting factor on power efficiency. We have presented holistic models of Cloud data center fundamentals reasonable to use in a detailed Cloud environment simulation infrastructure, abstract estimation and green energy prediction as an effective solution. Figure 10 displays the respective energy draws of our scenarios, actualizing 30% better energy efficiency and the list of scenarios is shown on Table 4.

7. CONCLUSION

Cloud computing is becoming more and more crucial in IT sector due to abundant advantages it renders to its end users. With the high user demands, Cloud environment possess very large ICT resources. To this, power and energy consumption of Cloud environment have become an issue due to ecological and economical reasons. In this paper, we have presented energy consumption formulas for calculating the total energy consumption in Cloud environments and show that there are incentives to save energy. To this respect, we described an energy consumption tools and empirical analysis approaches. Furthermore, we provide generic energy consumption models for server idle and server active states. This research result is crucial for developing potential energy legislation and management mechanisms to minimize energy consumption while system performance is achieved for Cloud environments.

As future work, we will investigate several Cloud environments and propose new optimization policies which will minimize the CO_2 emissions of Cloud environment, we will integrate energy cost rate into our new models in differing environmental impact and to minimize the total energy cost.

ACKNOWLEDGEMENTS

This work is supported by the National Science Foundation for Distinguished Young Scholars of China (Grant No. 61225010), NSFC under grant nos.61173160, 61173161, 61173162, 61173165 and New Century Excellent Talents in University (NCET-10-0095) of Ministry of Education of China.

REFERENCES

[1] M. Armbrust, "Above the clouds: A Berkeley view of cloud computing", Technical Report UCB/EECS-2009-28, 2009.

[2] BONE project, "WP 21 Tropical Project Green Optical Networks: Report on year 1 and unpdate plan for activities", No. FP7-ICT-2007-1216863 BONE project, Dec. 2009.

[3] J. Koomey, "Estimating Total Power Consumption by Server in the U.S and the World", February 2007, http://enterprise.amd.com/Downloads/svrpwrusecompletefinal.pdf

[4] Jordi Toress, "Green Computing: The next wave in computing", In Ed. UPC Technical University of Catalonia, February 2010.

[5] Peter Kogge,"The Tops in Flops", pp. 49-54, IEEE Spectrum, February 2011.

[6] U.S Environmental Protection Agency. (2006). *Report to Congress on Server and Datacenter Energy Efficiency Public Law* [Online]. Available: http://hightech.lbl.gov/documents/data_centers/epa-datacenters.pdf

[7] Greenpeace. (2011). *Greenpeace "Likes" Facebook's New Datacenter, But Wants a Greener Friendship* [Online]. Available: http://www. greenpeace.org/international/en/press/releases/Greenpeace-likes-Face- books-new-datacentre-but-wants-a-greener-friendship

[8] J. Kovar. (2011, Aug. 10). *Data Center Power Consumption Grows Less Than Expected: Report* [Online]. Available: http://www. crn.com/news/data-center/231400014/data-center-power-consumption-grows-less-than-expected-report.htm?pgno=2

[9] R. Miller. (2011, Sep.). *Google's Energy Story: High Efficiency, Huge Scale* [Online]. Available: http://www.datacenterknowledge.com/ archives/2011/09/08/googles-energy-story-high-efficiency-huge-scale

[10] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A view of cloud computing," *Commun. ACM*, vol. 53, no. 4, pp. 50–58, 2010.

[11] R. Buyya, "Market-oriented cloud computing: Vision, hype, and reality of delivering computing as the 5th utility," in *Proc. Int. Symp. Cluster Comput. Grid*, May 2009, p. 1.

[12] L. A. Barroso and U. Hölzle, "The case for energy-proportional computing," *IEEE Comput.*, vol. 40, no. 12, pp. 33–37, Dec. 2007.

[13] L. A. Barroso and U. Hölzle, "The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines". San Rafael, CA: Morgan and Claypool, 2009.

[14] N. Rasmussen, "Calculating total cooling requirements for datacenters," *Amer. Power Convers.*, white paper 25, 2007.

[15] Data Center Energy Efficiency Training, U.S. Department of Energy. (2011). *Electrical Systems* [Online]. Available: http://hightech.lbl.gov/training/modules/10-electrical-systems.pdf

[16] C. Belady, A. Rawson, J. Pfleuger, and T. Cader. *Green Grid Datacenter Power Efficiency Metrics: PUE and DCIE* [Online]. Available: http://www.thegreengrid.org/gg–content/TGG-Data–Center–Power–Efficiency–Metrics–PUE– and–DCiE.pdf

[17] S. Ghemawat, H. Gobioff, and S.-T. Leung, "The Google file system," in *Proc. ACM Symp. Operating Syst. Principles*, 2003, pp. 29–43.

[18] S. Higginbotham. (2012, Mar.). *Whose Data Centers Are More Efficient? Facebook's or Google's?* [Online]. Available: http:// gigaom.com/cloud/whose-data-centers-are-more-efficient-facebooks- or-googles/

[19] T. Seubert. (2012, Feb. 27). *Microsoft Builds New Data Center in Dublin* [Online]. Available: http://facilitygateway.com/news/?p=1937

[20] W.C. Feng, X. Feng and C. Rong, "Green Supercomputing Comes of Age," *IT Professional*, vol.10, no.1, pp.17-23, Jan.-Feb. 2008.

[21] X. Li, Y. Li, T. Liu, J. Qiu, and F. Wang, "The method and tool of cost analysis for cloud computing," in the IEEE International Conference on Cloud Computing (CLOUD 2009), Bangalore, India, 2009, pp. 93-100.

[22] L. Shang, L.S. Peh, and N. K. Jha, "Dynamic voltage scaling with links for power optimization of interconnection networks," In the 9th International Symposium on High-Performance Computer Architecture (HPCA 2003), Anaheim, California, USA, 2003, pp. 91-102.

[23] B. Rajkumar, B. Anton and A. Jemal, "Energy Efficient Management of Data Center Resources for Cloud Computing: A Vision Architectural Elements and Open Challenges", In Proc. International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 2010), Las Vegas, USA, July 12-15, 2010.

[24] A. Bohra and V. Chaudhary, "VMeter: Power modelling for virtualized clouds," *International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW), 2010 IEEE*, vol., no., pp.1-8, 19-23 April 2010.

[25] F. Chen, J. Schneider, Y. Yang, J. Grundy and Q. He, "An energy consumption model and analysis tool for Cloud computing environments," *Green and Sustainable Software (GREENS), 2012 First International Workshop on*, vol., no., pp.45-50, 3-3 June 2012.

[26] B. Yamini and D.V. Selvi, "Cloud virtualization: A potential way to reduce global warming," *Recent Advances in Space Technology Services and Climate Change (RSTSCC), 2010*, vol., no., pp.55-57, 13-15 Nov. 2010.

[27] Z. Zhang and S. Fu, "Characterizing Power and Energy Usage in Cloud Computing Systems," *Cloud Computing Technology and Science (CloudCom), 2011 IEEE Third International Conference on*, vol., no., pp.146-153, Nov. 29 2011-Dec. 1 2011.

[28] A.C. Orgerie, L. Lefevre, and J.P. Gelas, "Demystifying energy consumption in Grids and Clouds," *Green Computing Conference, 2010 International*, vol., no., pp.335-342, 15-18 Aug. 2010.

[29] I. Sarji, C. Ghali, A. Chehab, A. Kayssi, "CloudESE: Energy efficiency model for cloud computing environments," *Energy Aware Computing (ICEAC), 2011 International Conference on*, vol., no., pp.1-6, Nov. 30 2011-Dec. 2 2011.

[30] X. Fan, W.D. Weber, and L. A. Barroso, "Power provisioning for a warehouse-sized computer," in the *34th International Symposium on Computer Architecture (ISCA 2007)*, San Diego, California, USA, 2007, pp. 13-23.

[31] J. Moore, J. Chase, P. Ranganathan, R. Sharma, "Making Scheduling "Cool": Temperature-Aware Workload Placement in Data Centers" In Proc. *of the 2005 USENIX Annual Technical Conference, Anaheim, CA, USA, April 10–15, 2005.*

[32] P. Ehsan and P. Massoud, "Minimizing data center cooling and server power costs", *In Proc. Of the 4th ACM/IEEE International Symposium on Low Power Electronic and Design (ISLPED)*, 2009, Pages 145-150

[33] D. Meisner, B. T. Gold, T. F. Wenisch, "PowerNap: Eliminating Server Idle Power", *In Proc. of the 14th international conference on Architectural support for programming languages and operating systems (ASPLOS)*, USA, 2009.

[34] S. Pelley, D. Meisner, T. F. Wenisch, and J. W. VanGilder, "Understanding and absracting total data center power," in *WEED: Workshop on Energy Efficienct Design,* 2009.

[35] R. Tozer, C. Kurkjian, and M. Salim, "Air management metrics in data centers," in *ASHRAE* 2009, January 2009.

[36] J. W. VanGilder and S. K. Shrivastava, "Capture index: An airflow-based rack cooling performance metric," *ASHRAE Transactions*, vol. 113, no. 1, 2007.

[37] Y. A. Çengel, Heat transfer: a practical approach, 2nd ed. McGraw Hill Professional, 2003.

[38] Meade RL, Diffenderfer R (2003) Foundations of Electronics: Circuits & Devices. Clifton Park, New York. ISBN: 0-7668-4026-3

[39] ZES ZIMMER PA, Brodersen RW (1995) Minimizing Power Consumption in CMOS Circuits. Tech. rep., University of California at Berkely. http://bwrc.eecs.berkely.edu/publications/1995/\\Min_pwr_consum_CMOS_crct/paper.fm.pdf

[40] N. Rasmussen, "Electrical efficiency modeling for data centers," APC by Schneider Electric, Tech. Rep. #113, 2007.

[41] DOE (Department of Energy), "Doe 2 reference manual, part 1, version 2.1," 1980.

[42] CEC (California Energy Commission), "The nonresidential alternative calculation method (acm) approval manual for the compliance with california's 2001 energy efficiency standards," April 2001.

[43] A. Uchechukwu, K. Li and Y. Shen,"Improving Cloud Computing Energy Efficiency*." In Proc. Asia Pacific cloud Computing Congress,* Nov., 2012.

BIOGRAPHY OF AUTHORS

Awada Uchechukwu received the B.Sc. degree in computer science from Ebonyi State University, Abakaliki, Nigeria and the M.Eng degree in computer applied technology from Harbin Engineering University, Harbin, China, in 2011, he is currently working towards the Ph.D. degree in computer science and engineering with the Network and Cloud Computing Laboratory at Dalian University of Technology. His research interests include energy consumption, cloud computing, big data and distributed computing. He is a student member of the IEEE Computer Society.

Keqiu Li received B.S. and M.S degree in applied mathematics from Dalain University of Technology, Dalian, China in 1994 and 1997 respectively and Ph.D. degree in information technology from Japan Institute of Science and Technology in 2005. Currently, he is a Professor at Dalian University of Technology, where he is the Dean of the school of computer science and technology and Director of the network and cloud computing laboratory. His research interests include web technology, grid/cloud computing, mobile computing, network and security. He is a senior member of the IEEE and the IEEE Computer Society.

Yanming Shen received the B.S. degree in automation from Tsinghua University, Beijing, China, in 2000 and the Ph.D. degree from the Department of Electrical and Computer Engineering at the Polytechnic University (now Polytechnic Institute of New York University), Brooklyn. He is an Associate Professor in the Computer Science and Engineering Department at Dalian University of Technology, Dalian, China. He was a summer intern with Avaya Labs in 2006, conducting research on IP telephony. His general research interests include cloud computing, distributed systems, packet switch design, peer-to-peer video streaming, and algorithm design, analysis, and optimization.

`