Bulletin of Networking, Computing, Systems, and Software – www.bncss.org, ISSN 2186–5140
Volume 1, Number 1, pages 16–20, January 2012

Evaluating Energy Efficiency of Data Centers with Generating Cost and Service Demand

Cheng-Jen Tang, Miau-Ru Dai,
and Hui-Chin He
The Graduate Institute of Communication Engineering
Tatung University,
Taipei, Taiwan 104
Email: ctang@ttu.edu.tw, d9610002@ms2.ttu.edu.tw,
hche@ttu.edu.tw

Chi-Cheng Chuang
Smart Network System Institute,
Institute for Information Industry,
Taipei, Taiwan 105
Email: polon@nmi.iii.org.tw

***Abstract*—Cloud computing regards applications (or software) as services. A cloud computing data center must conduct huge amount of computations in addition to traditional server tasks. The energy cost of cloud computing datacenters are therefore staggering high. Cloud service data center has a different power consumption pattern from the traditional storage oriented Internet services. The computation oriented implementation of cloud service broadens the gap between the peak power demand and base power demand of a data center. Improving the energy efficiency is a crucial strategy for overcoming the issue. This paper presents a new model for the energy efficiency of a cloud service data center. This model considers the energy efficiency of both the power demand side and the power supply side.**

I. Introduction

Workload of a data center is increasing due to the emerging Cloud Computing. Cloud computing services are computation intensive services. A cloud service data center clusters distributed computers that provide applications as services and on-demand resources, including computing, data accessing etc., over Internet to satisfy user demands. Along with the increasing workload, power demand, processing time, and operation cost are also lifting accordingly.

From the perspective of power demand, the computation oriented nature of cloud computing differs the power consumption patterns of data centers from the traditional storage oriented model. This computation oriented model is likely to widen the difference between the amount of energy required for peak periods and off-peak periods. Some issues can be identified for the increasing separation on peak and off-peak demands, such as:

1) Increased computation needs enlarge CPU utilization rate. Power consumption of modern server CPUs is proportional to their utilization.
2) Increased computation needs also bring up the required memory sizes. DRAM has been identified as one of the main contributors to energy consumption in a computer.
3) Frequent computation requests raise the frequency of random disk accesses. Random disk accesses consume much more energy than sequential disk accesses.

In short, the coming-out of cloud computing services enlarges the power demand of a data center. In addition to the increasing power demand, the followed large fluctuation of power demand broadens the gap between the peak and the off-peak power demand. This is clearly not a welcome trend for the electricity suppliers.

II. Background

Many studies point out that the ever-growing energy demand of data centers has become a critical challenge. In 2007, Koomey [1] pointed out that the energy usage growth of either the total consumption or the data centers doubled from 2000 to 2005 worldwide. A great portion of this growth comes from the increasing of the number of servers. Rapidly developing information and communication technologies (ICTs) and Internet elaborates establishment of datacenters. Modern datacenters usually equip with fully populated rack of blades to better use their limited spaces. For example, Google has over 450,000 servers. [2] [3] In a report from Tschudi et al. [4], server computers are responsible for 51 percent of energy cost, in which the investigated items include the server load, computer room air condition (CRAC) units, cooling tower plant, electrical room cooling, office space conditioning, lighting and others. According to an estimation from Amazons [5], expenses related to the cost and operation of their servers is about 53 percent in their monthly cost. Patterson et al. [6] has measured and estimated the power consumption density (PCD, $watts/feet^2$) growth of IT equipments in a data center. Although the network equipments scores the highest PCD, servers still consume most of the supplied power since there are many more server computers than the switches or the routers in a data center.

In a data center, the main consumers of energy are the computation devices and the cooling facilities. [7] [8] With the emergence of cloud computing, data centers now have to respond to the mounting computation requests. Data centers hunger for even more power than ever as the increasing computation needs raising the electricity demand. This increasing power demand surely brings up the energy cost. Thus, improving energy efficiency becomes a key issue for the cost-effectiveness of the data center management.

Computing resource consolidation aims at improving the effectiveness, including energy efficiency, of data centers. Virtualization, consolidating physical machines, and cloud computing, consolidating computation capabilities, are two distinct techniques among others towards such direction. In order to support the increasing computation needs of a cloud service data center, the service provider needs to upgrade the infrastructure to satisfy the new requirements, which include high performance server computers, cooling systems, network and electricity facilities, etc. All these additional devices require electricity to function. Thus, cloud computing raises the energy requirement of data centers. For either the power supplier or the service provider, additional effort is required for either side due to the increased power demand. The power supplier needs to guarantee the capability of supplying, especially during the peak periods. The service provider needs to pay higher energy bills in addition to the equipment investments. In order to reduce the amount of required physical machines, server virtualization has become a popular technique. Data centers with virtualization are allowed to consolidate resource of their physical servers into several high-level servers to achieve better hardware usage. In general, this mechanism is capable of reducing energy consumption. [9] [10] However, due to the limited memory size, virtual machines are more often swapping data out and in from their swap spaces, which eventually consume much more energy.

III. Energy Efficiency from Different Perspectives

Many studies investigate the energy usage effectiveness of data centers from the demand side. These definitions of the energy efficiency include:

1 Green Grid [11] proposes Power Usage Effectiveness (PUE) and Datacenter Efficiency (DCE) metrics. PUE and DEC evaluate the energy efficiency of data centers to determine whether a data center requires improving energy efficiency or not. The entities in a data center are categorized as: IT devices and other equipments. IT devices consist of servers, storage units, networking equipments. As shown in Eq. 1, ideal PUE is 1.

$$PUE = \frac{TotalFacilityPower}{ITequipmentpower}, DCE = \frac{1}{PUE} \quad (1)$$

2 (2) Tsirogiannis et al. [12] define the energy efficiency as the ratio between the number of finished jobs and the consumed energy, as shown in Eq. 2.

$$EE = \frac{Workdone}{Energy} = \frac{Workdone}{Power \times Time} = \frac{Performance}{Power} \quad (2)$$

This paper presents a different viewpoint on energy efficiency, which considers this issue as an interaction result between the demand side (data centers) and the supply side (utilities). For utilities, they use fuels to generate electricity. For data centers, they consume electricity to make jobs done. Therefore, this becomes a fuel-to-job system.

TABLE I
Symbols about data centers

Symbols	Descriptions
$P_{ds}(t)$	The supplied power to a data center d at time t
n_d	The number of server machines in a data center d
K_d	A constant represents the line loss and power consumed by other active electrically-driven devices in a data center d
$P_{di}(t)$	The power consumed by a server machine i in a data center d, where $i = 1...n_d$

For a utility, the main challenge is to generate enough electricity to meet the demand while avoiding too much wasted power. Because of the large fluctuations of power demand, electricity generators are categorized as the base load, intermediate load, and peak load generators. Base load generators provide the most essential power that has to be stable. Intermediate and peak load generators are capable of quickly on and off to deal with the variation in power demands. The power demand, which is time varied, determines what kind of generators to supply during a particular time period. For a utility, improving energy efficiency means reducing fuel cost but still satisfying demand.

The number and the frequency of the network requests are the important factors affecting energy consumption of a data center. According to a statistic presented by the InternetWorldStat [13], there were already 2,095,006,005 Internet users as of March 31, 2011. If each Internet user spends one joule per second for a minute on computation tasks daily, the total consumed energy for one year is about 18 TWh. Obviously, how the computation tasks are handled is one the important issues regarding the energy efficiency of data centers.

This study evaluates the overall energy efficiency by considering both the supply side and the demand side. The energy efficiency is defined as the ratio between the number of finished jobs and the fuel cost for generating electricity.

Suppose the amount of supplied power to the server is P_s (the apparent power); the real power is P_u, which is the power consumed by working machines in a data center. The power factor, PF, is defined as the Eq. 3.

$$PF = \frac{P_u}{P_s} \quad (3)$$

Suppose $PF_d(t)$ is the power factor of a datacenter d at time t that is the Eq. 4.

$$PF_d(t) = \frac{K_d + \sum_{i=1}^{n_d} P_{di}(t)}{P_{ds}(t)} \quad (4)$$

In the Eq. 4, the number of requests for serving affects this part $P_{di}(t)$; $P_{ds}(t)$ is limited on the costs of generation. The followings investigate the detail of the $P_{di}(t)$, and $P_{ds}(t)$.

1) The power consumed by a server machine i in a data center d, $P_{di}(t)$: Suppose the power usage of a server i at the idle state is $P_{idle_{di}}$. Suppose the server i performs one request at a time. Other requests are queued until the working request

TABLE II
SYMBOLS ABOUT THE COST OF POWER GENERATION

Symbols	Descriptions
$P_{Bd}(t)$	The power generated for a data center d by base load generators at time t
$P_{Id}(t)$	The power generated for a data center d by intermediate load generators at time t
$P_{Pd}(t)$	The power generated for a data center d by peak load generators at time t
C_B	The unit cost of the power generated by base load generators
C_I	The unit cost of the power generated by intermediate load generators
C_P	The unit cost of the power generated by peak load generators
P_{MAX_B}	The maximum output of base load generators
P_{MAX_I}	The maximum output of intermediate load generators
P_{MAX_P}	The maximum output of peak load generators

has finished. Thus, the $P_{di}(t)$ is:

$$P_{di}(t) = \begin{cases} P_{idle_{di}}, & \text{if there is no jobs at time t} \\ P_{idle_{di}} + P_{kdi}, & \text{if there is request k at time t} \end{cases} \tag{5}$$

This P_{kdi} showed in the Eq. 5 represents the power requirement of the request k performing on server i of a data center d. Suppose the energy required for finishing a client request k on a server i of a data center d is E_{kdi}. Suppose a server i needs a period of time L_{kdi} to finish the request k. The power requirement of the request k performing on a server i of a data center d is:

$$P_{kdi} = \frac{E_{kdi}}{L_{kdi}} \tag{6}$$

The power consumption of all server machines in a data center d at time t is able to be found. In the tested data center, all server machines have the same hardware configuration. This paper also assumes all requests consume the same amount of energy, and P_j is the power requirement for performing a request. Therefore, for a period t_0 to t_1:

$$\sum_{i=1}^{n_d} P_{di}(t) = n \times P_{idle_{di}}(t) + \frac{J_d}{t_1 - t_0} \times P_j \tag{7}$$

2) The supplied power to a data center d at time t, $P_{ds}(t)$: As mentioned earlier, different types of power generators are required to meet this fluctuating demand. Generators are usually divided into three different types according to their missions:

- Base load generators,
- Intermediate load generators, and
- Peak load generators.

In general, the utility takes the minimum cost as the basic to plan their generation. Peak load generators usually cost most, followed by intermediate load generators, and then base load generators. The power generation cost for the data center d at time t defines:

$$C_d(t) = C_B P_{Bd}(t) + C_I P_{Id}(t) + C_P P_{Pd}(t) \tag{8}$$

Therefore, the supplied power to a data center d at time t is:

$$P_{ds}(t) = \begin{cases} P_{Bd}(t), \\ \text{for } P_{ds}(t) < P_{MAX_B} \\ P_{MAX_B} + P_{Id}(t), \\ \text{for } P_{ds}(t) < P_{MAX_B} + P_{MAX_I} \\ P_{MAX_B} + P_{MAX_I} + P_{Pd}(t), \\ \text{for } P_{ds}(t) < P_{MAX_B} + P_{MAX_I} + P_{MAX_P} \end{cases} \tag{9}$$

This paper does not consider the condition of $P_s(t)$ exceeding P_{MAX_P} , which causes the circuit breaker of the datacenter to be tripped.

This paper defines the overall energy efficiency as this ratio of the number of requests and the cost of generation. Therefore, for a period t_0 to t_1:

$$Eff_d = \frac{J_d}{\int_{t_0}^{t_1} C_d(t)dt} \tag{10}$$

The $C_d(t)$ is able to determine through the Eq. 4, 9 and 7.

$$P_{ds}(t) = \frac{K_d + n \times P_{idle_{di}}(t) + \frac{J_d}{t_1 - t_0} \times P_j}{PF_d(t)} \tag{11}$$

From Eq. 11, the P_{MAX_B} of a carefully designed power generation system of a data center d is:

$$P_{MAX_B} = \frac{C + n \times P_{idle_{di}}}{PF_d} \tag{12}$$

Peak load generators are operated under some critical conditions. For most of time, the $P_s(t)$ is:

$$P_s(t) = P_{MAX_B} + P_{Id}(t) \tag{13}$$

The $C_d(t)$ is therefore:

$$C_d(t) = C_B P_{MAX_B} + C_I P_{Id}(t) \tag{14}$$

From Eq. 11, 12, and 13, the $P_{Id}(t)$ is able to found.

$$P_{Id}(t) = \frac{\frac{J_d}{t_1 - t_0} \times P_j}{PF_d} \tag{15}$$

Therefore, the C_d is:

$$C_d(t) = C_B P_{MAX_B} + C_I \frac{\frac{J_d}{t_1 - t_0} \times P_j}{PF_d} \tag{16}$$

The overall energy efficiency is able to obtain from Eq.s 10, and 16.

$$Eff_d = \frac{J_d}{\int_{t_0}^{t_1} C_B P_{MAX_B} + C_I \frac{\frac{J_d}{t_1 - t_0} \times P_j}{PF_d} dt} \tag{17}$$

$$Eff_d = \frac{J_d}{(C_B P_{MAX_B} + C_I \frac{\frac{J_d}{t_1 - t_0} \times P_j}{PF_d})(t_1 - t_0)} \tag{18}$$

Let a constant $C_0 = C_B P_{MAX_B}(t_1 - t_0)$, a constant $C_1 = \frac{C_I P_j}{PF_d}$, the Eq. 18 is written as:

$$Eff_d = \frac{J_d}{C_0 + C_1 J_d} \tag{19}$$

$$Eff_d = \frac{1}{C_1} - \frac{C_0}{C_1(C_0 + C_1 J_d)} \tag{20}$$

From Eq. 20, to maximize Eff_d needs to maximize J_d. The upper bound of J_d in a period t_0 to t_1 is $\frac{t_1 - t_0}{L_j}$. Therefore, an easy way to gain a better energy efficiency of a data center is to reduce the L_j , which is the period of time required for finishing a request.

IV. Analysis and Discussion

This study uses the logged user record from a server at http://www.gottu.tw to examine the proposed idea. The server http://www.gottu.tw provides rich information for visitors to help them choosing an appropriate department and university in Taiwan after the end of the college entrance examination. This site has attracted more than 660,000 unique visiting IP addresses in one month. Each IP address has issued 7.2 requests in average. The Fig. 1 shows the average number of requests in every 15 minutes of one day from July 12, 2011 to August 11, 2011. There are 3 peaks at about 10:00 AM, 15:00 PM, and 21:00 PM of a day. The bottom is located at about 4:00 AM. It is clearly similar to the daily work-rest pattern.

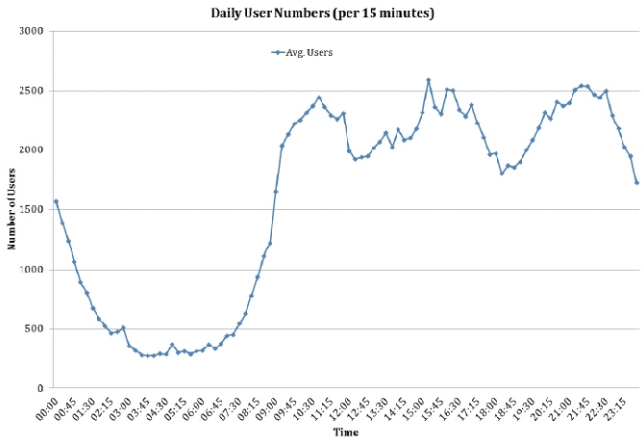

Fig. 1. Daily user numbers

Different types of generators have different time to ramp up or down. Peak load generators ramp up or down very quick, followed by intermediate load generators, and finally are base load generators. Suppose the time to ramp up intermediate load generators is 4 hours; 2 hours for peak load generators. Base load generators are always up. For the load curve shown in Fig. 2, the planning of three type generators is:

- Base load generators support the green block.
- Intermediate load generators are responsible for the blue blocks.
- Peak load generators are for the red blocks.

The deployment is taking the lowest cost as the basic condition to determine the generation. In fact, an optimal deployment is obtained by a series of simulations. The amount of generation show in Fig. 3.

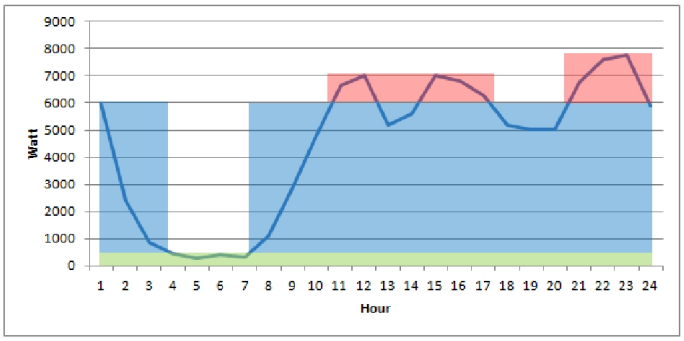

Fig. 2. Daily load curve

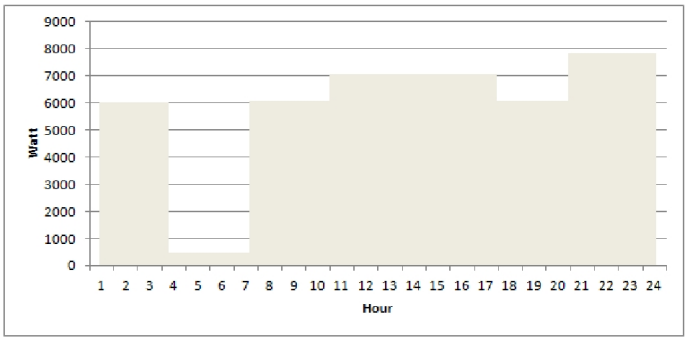

Fig. 3. Daily load curve

Previous studies aim at lowering the energy costs. However, lower energy costs do not always imply better energy efficiency. For a utility, the minimum gap between the peak demand and off-peak demand is preferred and is more energy efficient. The computation oriented nature of cloud computing differs the power consumption patterns of data centers from the traditional storage oriented behavior. Such cloud service broadens the gap between the peak power demand and base power demand of a data center. This study creates two test sites that one generates web pages without sorting their contents, and the other sorts the contents whenever a request is received. Fig. 4 shows the differences between power usages of two sites.

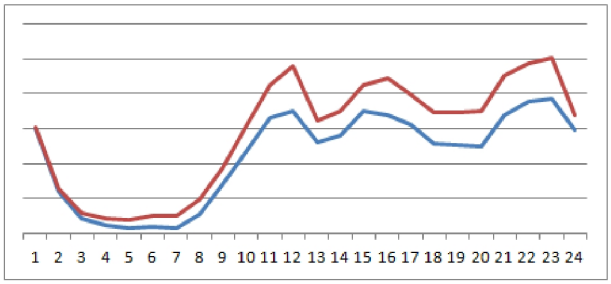

Fig. 4. Daily load curve

Modern cloud service data center constructs their micro power system to reduce the energy cost. Such cloud service data center needs to pay more attention to choose the power source. The energy efficiency of such data center needs to

consider the relationship between the workload and the cost of generation. This study models the energy efficiency of a data center as:

$$Eff_d = \frac{J_d}{\int_{t_0}^{t_1} C_d(t)dt} \tag{21}$$

J_d represents the number of finished jobs, and $C_d(t)$ denotes the cost of the generation in a period t_0 to t_1.

V. Conclusion

Cloud computing is a computation intensive service that clusters distributed computers providing applications as services and on-demand resources over Internet. Theoretically, such consolidated resource enhances the energy efficiency of both clients and servers. In reality, cloud computing is not a panacea for enhancing energy efficiency under some certain conditions. Most of the existing proposals focus on reducing the use of energy. This paper presents a new model for the energy efficiency of a cloud service data center.

This model considers the energy efficiency from both the demand side and the supply side. The supply side, such as utilities, dispatches different type generators based on the load curve and their characteristics. The demand side, such as cloud service data centers, needs to integrate the workload, the cost of their own power system, and the electricity contract to plan an optimal energy planning.

Acknowledgment

This study is conducted under the "Advanced Metering Infrastructure(AMI) Enhancement Project" of the Institute for Information Industry which is subsidized by the Ministry of Economy Affairs of the Republic of China.

References

[1] J. Koomey, "Estimating total power consumption by servers in the us and the world," 2007.

[2] D. Carr, "How google works," *Baseline Magazine*, vol. 6, no. 6, 2006.

[3] J. Grimmelmann, "The google dilemma," 2009.

[4] W. Tschudi, T. Xu, D. Sartor, and J. Stein, "High-performance data centers: A research roadmap," 2004.

[5] J. Hamilton, "Cooperative expendable micro-slice servers (CEMS): low cost, low power servers for internet-scale services," in *Conference on Innovative Data Systems Research (CIDR?09)(January 2009)*. Citeseer.

[6] M. Patterson, D. Costello, P. Grimm, and M. Loeffler, "Data center tco; a comparison of high-density and low-density spaces," *Thermal Challenges in Next Generation Electronic Systems (THERMES 2007)*, 2007.

[7] X. Fan, W. Weber, and L. Barroso, "Power provisioning for a warehouse-sized computer," in *Proceedings of the 34th annual international symposium on Computer architecture*. ACM, 2007, pp. 13–23.

[8] J. Hamilton, "Cost of power in large-scale data centers," *Blog entry dated*, vol. 11, p. 28, 2008.

[9] Y. Wang, X. Wang, M. Chen, and X. Zhu, "Partic: Power-aware response time control for virtualized web servers," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 22, no. 2, pp. 323 –336, feb. 2011.

[10] S. Murugesan, "Harnessing green it: Principles and practices," *IT Professional*, vol. 10, no. 1, pp. 24 –33, jan.-feb. 2008.

[11] G. Grid, "Green grid metrics: Describing datacenter power efficiency," Tech. Rep., 2007.

[12] D. Tsirogiannis, S. Harizopoulos, and M. Shah, "Analyzing the energy efficiency of a database server," in *Proceedings of the 2010 international conference on Management of data*. ACM, 2010, pp. 231–242.

[13] "Internet users in the world, distribution by world region," 2011. [Online]. Available: http://www.internetworldstats.com/stats.htm