Energy efficient optimization method for green data center based on cloud computing

Runze WU[1, a], Wenwei CHEN[1, b], Li LI[2,c] and Kuangyi ZHAO[2,d]

[1]School of electrical & Electronic Engineering (North China Electric Power University), Beijing

[2] State Grid Jibei Electric Power Company Economic Research Institute, Beijing

[a]wrz_y@126.com, [b]572451718@qq.com, [c] jbjyyll@126.com , [d]1476340851@qq.com

Keywords: Data center; Maximize efficiency; Clustering analysis; Scheduling mechanism

Abstract. With the rapid development of cloud computing data centers, the problems of tremendous power consumption has become increasingly prominent, so it is of great practical significance to study the energy efficiency of cloud computing green data center. This paper proposes a mathematical model based on the constraint of QoS which combines benefits and costs, the model takes account some factors including electricity price, renewable power generation, service rate and Qos. Finally clustering algorithm is applied to the data center task scheduling mechanism to reduce data center energy consumption and increase efficiency.

Introduction

The growing demand for big data technology and cloud computing has greatly increased the energy consumption with big data centers. For example, Microsoft's data center has 965 kilometers of electric wire, and 1.5 metric tons of backup batteries, which is equal to the power consumption of 40000 homes [4]. Therefore, how to deal with the problem of energy efficiency is the key issue of data center. Due to establish a lager data center, there has been a growing interest towards developing algorithms to minimize data center's energy consumption and maximize energy efficiency. In order to improve the energy efficiency of green energy grid, cloud computing data center must be based on principles of safety, economy, green and efficient. To address this issue, we propose a system model about optimization-based energy profit maximization in this paper，which combines new energy power generation and power generation in the traditional way, realizes the self-sufficiency of data center, specifically explains the calculation process of data center efficiency, provides a method for reducing energy consumption and measuring energy efficiency. Finally, the scheduling model based on cluster analysis is built from the perspective of the resource providers of cloud computing data center. The model improves the efficiency of the scheduling task to complete and contributes to high efficiency and energy saving.

Cloud Computing Data Center

The data center is called "server farm" which refers to the large core computer resources in a certain environment, and it has a unified management and provides the required business for users and enterprises every moment. A powerful and efficient data center needs large servers, switches, routers, power supply equipment and related facilities. "Resource pool" is the sum of the virtual resources provided by cloud computing data center. The energy consumption of data center mainly results from the power consumption, so the use of new energy power generation will play a significant role in energy saving. High energy consumption will lead to high energy costs and bring carbon emissions which may go against sustainable development. So the establishment of a data center must also take the economic benefits and environmental impact into account.

The energy consumption of data centers is mainly from two aspects, namely, the power consumption of all servers and network equipment in the data center and consumption of data control

center to ensure the normal operation of the whole system. In general, let η denote an indicator to measure the use of energy efficiency, η is defined as follows:

$$\eta = \frac{E}{E_{IT}} \tag{1}$$

Where E denotes the total energy consumption of data center, E_{IT} is the total energy consumption of IT equipment. η has been 1.08 in green data center, this result indicates that the energy consumption of entire data center mainly comes from power supply to devices.

Mathematical Model Based on Energy Efficiency Optimization

In order to minimize the energy cost of data center and maximize the user's business needs, the mathematical model of energy efficiency optimization for the data center is established. Owning to the power consumption of the computer servers, the total energy consumption of the business needs received every day includes data center power utilization, electricity price change, renewable energy utilization, etc. In this system model, the actual service level agreements and parameters for the actual service level of the user's task to complete the time limit, the user's payment and the service provider to provide the compensation in the time limit. According to the above parameters, we propose the mathematical model of the optimization about the cost function and income based on QoS constraints, then the model is shown in Figure 2.

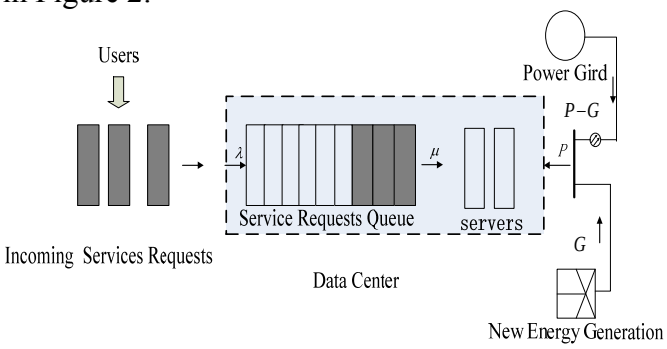

Fig.2 Mathematical model of energy efficiency

The system model is explained in terms of power consumption, price of electricity, incoming workload and quality of service.

Power Consumption. The total power consumption of a data center is the power consumption of the computer servers and its infrastructure, such as lighting, cooling. η is defined as power usage effectiveness, which is an important indicator of measuring the energy efficiency of a data center. Total power consumption can be obtained by using the following formula:

$$P = M[P_{idle} + (\eta - 1)P_{peak} + \left(P_{peak} - P_{idle}\right) * U] \tag{2}$$

We can calculate η according to (1), let M denote the number of servers that are the normal operation of the state at data center. P_{idle} is the average idle power draw of a single server and P_{peak} is the average peak power when a server is dealing with a service request. The ratio P_{peak} and P_{idle} denotes the power elasticity of servers. The higher the ratio is, the minimum power required for the free server and the minimum power consumption. U is the CPU utilization of servers. This model shows that the total power consumption of the server and the CPU utilization should have a certain linear relationship, and it is positive growth.

Electricity Price. The electricity price model of each region is based on the region's electricity market is regulated. More specifically, if the electricity market is regulated, the region's electricity prices won't change in a day and is in a steady state, on the other hand, if it is not regulated, the price of electricity is a significant change in a day, the fluctuation of the price reflects the fluctuations of electricity market. The cost of the data center can be obtained by pricing information, and the

corresponding pricing model can be helpful to the establishment of data center based on energy efficiency.

Renewable Power Generation. In order to reduce cost of electricity, a data center may use renewable power generators, as shown in Figure 2, a wind turbine will be connected to the grid and realize wind power generation. Let G be the renewable power generated by renewable generators, $P-G$ is the amount of power change with the power grid. Local renewable power generation is lower than local power consumption when $P>G$, namely, $P-G$ is positive and current direction is from power grid to data center, if $P=G$, the power generated by the new energy is just used for the entire data center. If $P<G$, $P-G$ is negative and current direction is from data center to power grid. In this case, assume that the data center does not receive the compensation of the input power when data center input additional new energy to the power grid. Power grid is used to compensate the power consumption of the data center according to the model of electricity market.

QoS. Data center has a limited computing capacity, and the actual working load is random, data center can't process the incoming service requests in time, so all service requests are placed in a certain queue in order to meet the requirements of service quality. In order to meet the quality of service, the waiting time of service request and the queuing delay are limited in a certain range according to the service level agreement (SLA). The exact SLA depends on the type of service based on cloud computing, such as video streaming and Web HTML services. Each SLA is measured by three non-negative parameters D, δ and γ. D represents the maximum waiting time for a service request, namely, the time of the user's task completed. δ indicates data center obtain user fees for completing the task of the user before D. Data center gives the user's compensation γ because it can not complete the task before D.

Service Rate. Let μ denote the rate at which service requests are removed from the queue and handled by a server, and it depends on the number of servers that is "on". Let S denote the time it takes for a server to finish dealing with a service request. $f=1/S$ is defined as the number of service requests per second, so the total service rate is obtained as:

$$\mu = fM \Rightarrow M = \mu / f \tag{3}$$

The formula (3) shows that increasing the number of normal operating servers can improve the service rate, and it also increases the energy consumption of data centers, so it is important for the system model to choose a reasonable service rate that can reduce the energy consumption of data centers.

Establishment of energy profit maximization function. The rate at which the service request to the data center is changed over time, in order to improve the efficiency of data center, the number M of servers in normal operation should be determined according to the rate of the received service requests. More servers should be turned on when services requests are received at higher rates. We can adjust the number of servers in operation state by monitoring service request rate. So the servers will provide service to users according to a certain proportion, by this way, we could reduce the power consumption of data centers. Because server varies from the state on and off, it will cause the state delay of the computer server. M is generally not quick to change, and it takes a few minutes to update, so the operating time of data center is divided into the same time slot $T_1, T_2, \ldots T_N$, the length of slot is $t=15\min$. Servers are updated at the beginning of each time slot. The profit of the entire data center is that the total revenue minuses the cost of energy consumption, energy saving means the minimum cost and the maximum benefit. Let β denote the maximum waiting time of the service request over SLA, the model and the calculation method are analyzed in reference [8]. Assume that λ is the average rate of receiving service requests within rime slot T_N. The total revenue of data center at the time slot of interest can be calculated as

$$reward = (1-\beta)\delta\lambda t - \beta\gamma\lambda t \tag{4}$$

Where $(1\text{-}\beta)\delta\lambda t$ denotes the total payment received by data center within time slot T_N, where $\beta\gamma\lambda t$ denotes the total penalty paid by data center within time slot T_N for the services requests that are not

handled before the SLA. So the total revenue is limited by QoS service quality constraints.Each server turned on can handles n service requests within time slot T_N.

$$n = t(1-\beta)\lambda / M \tag{5}$$

Assume that the busy time of CPU is t , the CPU utilization rate for each server is U , then we can get

$$U = (1-\beta)\lambda / f\, M \tag{6}$$

Replacing (3) and (6) in (2), we can get the power consumption P of the data center at the time slot :

$$P = a\mu + b\lambda(1-\beta)/f \tag{7}$$

where

$$a \triangleq P_{idle} + (\eta - 1)P_{peak} \tag{8}$$

$$b \triangleq P_{peak} - P_{idle} \tag{9}$$

Multiplying (7) by the electricity price ω , so we can get the total energy cost of data center at the time slot of interest is obtained as:

$$\cos t = t\omega\left[a\mu + b\lambda(1-\beta)/f\right] \tag{10}$$

So the total profit of the data center at the time slot is obtained as:

$$profit = reward - cost \tag{11}$$

To achieve the maximum benefit of the data center, we can select the appropriate service rate of the data center, so we can transform the optimal solution into solving the following optimization problem:

$$\underset{\lambda \le \mu \le fM_{\max}}{Max} \left[(1-\beta)\delta\lambda t - \beta\gamma\lambda t\right] - t\omega\left[a\mu + b\lambda(1-\beta)/f\right] \tag{12}$$

$M_{\max}$ is the total number of servers available in the data center. In order to assure stabilizing the service request queue, we can get $\lambda \le \mu$. In this model, the data center is simultaneously supplied by the power grid and new energy sources, and the optimal solution of the following (13) is the optimal service rate for the data center benefit maximization:

$$\underset{\lambda \le \mu \le fM_{\max}}{Max} \left[(1-\beta)\delta\lambda t - \beta\gamma\lambda t\right] - t\omega\left[a\mu + b\lambda(1-\beta)/f - G\right]^{+} \tag{13}$$

Where $[x]^{+} = \max(x, 0)$, $a\mu + b\lambda(1-\beta)/f - G$ indicate the amount of power to be purchased from the grid. If $P - G$ is negative, the data center's electricity cost will be zero if the power grid does not provide compensation for the input power. The advantage of this model is that renewable energy generation can balance the power consumption of the data center, the number of normal operating servers depends on the rate of service request, a certain queue model can avoid the server wasting too much idle time. The optimal allocation of resource pool of cloud computing data center is achieved. Through a reasonable allocation of idle resources, the cloud controller gives the task scheduling to the idle server cluster. The main idea of clustering analysis is to put similar objects into the same cluster, and clustering analysis is applied to idle servers.

Scheduling Mechanism Based on Data Mining

Scheduling is a decision-making process in which the resources are allocated to the users in a specific time point or time period according to users' requirements. Scheduling mechanism is the key to the development of high efficiency data center, which is related to the rational allocation of resources and the efficient completion of service request. In view of this, all idle resources that meet the needs of users are divided into K clusters according to the data center network structure by clustering algorithm. If there is a requirement of users in the cluster when the user sends out the service request, we can complete the task with the cloud controller, there are multiple clusters to meet the needs of the user, then an optimal cluster is selected according to the other constraints to accomplish the task. The basic flow chart is shown in Figure 3.

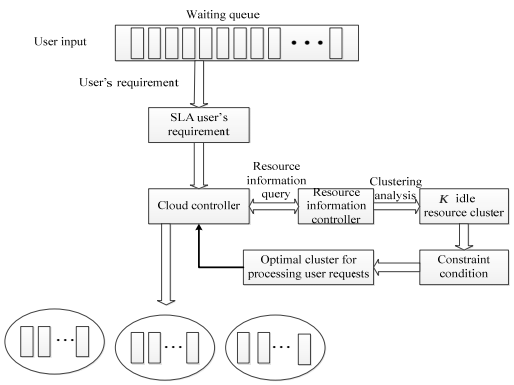

Fig.3 Scheduling model of clustering algorithm

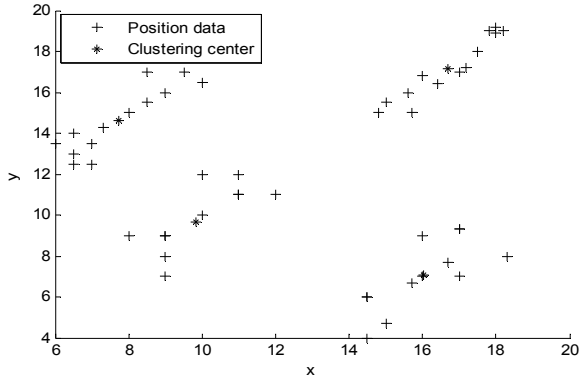

Fig.4 Results of clustering algorithm

Resource information controller mainly collects and records the current idle server classification, which is aimed at improving the efficiency of finishing scheduling tasks, and reducing the cost of the service provider. There are many methods for clustering analysis [6], the improved K-means algorithm is applied to finish clustering in this paper. There are 50 idle servers that are normal operation, two-dimensional coordinates represent the position of the idle servers, the input of the improved K-means is these 50 position data, then the data are divided into four clusters by clustering analysis. The simulation results are showed in Figure 4.

Summary

In the cloud era, energy reserves are dwindling and energy demand continues to grow. So energy efficiency has become a key problem to restrict the development of large enterprises, data centers are running a large number of computer servers, and always consume a huge power, data center based on energy efficiency will become an inevitable trend. The data center benefit maximization model and the establishment of data center structure is presented in this paper from the perspective of reducing energy consumption. large-scale data center will produce massive data, data mining is applied in the data center scheduling mechanism, the user needs to optimize the allocation of idle resources to the corresponding user tasks, users select the data center cloud resources according to their own needs, a win-win goal that resource supply and resource consumption will be achieved.

Acknowledgments

This work was supported by Nation Nature Science Foundation of China (No. 51507063).

References

[1] Liang LUO, Wenjun WU, Fei ZHANG. Energy modeling based on cloud data center[J]. Journal of Software, 2014,07:1371-1387.

[2] Wei DENG, Fangming LIU, Hai JIN. Leveraging renewable energy in cloud computing Data centers: State of the art and future research[J]. Chinese journal of computers, 2013,03:582-598.

[3] Xiong N, Han W, Vandenberg A. Green cloud computing schemes based on networks: a survey[J]. IET Communications, 2012, 6(18): 3294-3300.

[4] Ghamkhari M, Mohsenian-Rad H. Energy and performance management of green data centers: a profit maximization approach[J]. Smart Grid, IEEE Transactions on, 2013, 4(2): 1017-1025.

[5] Bera S, Misra S, Rodrigues J J P C. Cloud computing applications for smart grid: a survey[J]. 2014.

[6] Lijuan ZHOU, Hui WANG. Parallel KMeans algorithm for massive data[J]. J.Huazhong Univ. of Sci. & Tech. 2012,S1:150-152.