ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)
Vol. 2, Issue 4, April 2016

AN EFFICIENT COST OPTIMIZATION IN CLOUD COMPUTING BASED ON WEIGHTED ROUND ROBIN SCHEDULING

S.P.ABINAYA1, T.SHEIK YOUSUF 2

(Anna Univ Affiliated)M.E Department of Computer Science and Engineering 1
(Anna Univ Affiliated) Associate professor Department of Computer Science and Engineering2
Mohamed Sathak Engineering College,kilakarai, Ramanathapuram district, TN.
1abipalani0610@gmail.com
2sheikras@gmail.com

***Abstract*— Cloud computing is an emerging technology, which is a new paradigm for delivering remote computing resources through a network. Here there are many critical issues for the cloud providers such as achieving an energy-efficiency control and simultaneously satisfying a performance guarantee. This paper implement three power-saving policies in cloud systems in order to mitigate server idle power. This paper study the challenges of controlling service rates and applying the N-policy to optimize operational cost within a performance guarantee. Here a cost function is develop that includes the costs of power consumption, system congestion and server startups. Here the effect of energy-efficiency controls on response times, operating modes and incurred costs are demonstrate. The objectives are to find the optimal service rate and mode-switching restriction, so as to minimize cost within a response time guarantee under varying arrival rates. Here propose an algorithm called a Weighted Round Robin (WRR) Scheduling algorithm which is developed for solving constrained optimization problems and making costs/performances tradeoffs in systems with different power-saving policies, which reducing operational costs and improving response times which can be verified by applying the power-saving policies combined with the proposed algorithm as compared to a typical system under a same performance guarantee.**

***Keywords*— Cost optimization, energy-efficiency control, response time, power-saving policy**.

I.INTRODUCTION

The cloud computing model is comprised of a front end and a back end. These two elements are connected through a network. The front end is the vehicle by which the user interacts with the system and the back end is the cloud itself. The front end is composed of a client computer, or the computer network of an enterprise, and the applications used to access the cloud. The back end provides the applications, computers, servers, and data storage that creates the cloud of services. Cloud computing describes a type of outsourcing of computer services, similar to the way in which the supply of electricity is outsourced. Users can simply use it. They do not need to worry where the electricity is from, how it is produced, or transported. In cloud, services allowing users to easily access resources anywhere anytime.

Users can pay for a service and access the resources made available during their subscriptions until the subscribed periods expire. Users are then forced to demand such resources if they want to access them also after the subscribed periods. We mainly focused on the service provision issues on IaaS, which abstracts hardware resources into pool of computing resources and virtualization infrastructure. IaaS providers build flexible cloud solutions according to the hardware requirements of customers; furthermore it let customers run operating systems and software applications on virtual machine (VMs).Customers merely pay for the resources that are actually used. To host web application services, service operators would apply resource subscription plans to dynamically adjust service capacity to satisfy a time-varying demand. While subscribing IaaS resources, the web service operators aimed to provide a certain level Agreement (SLA) with their clients, e.g.,a guarantee on request response time. The resource provisioning of IaaS allows consumers to elastically increase or decrease the system capacity by changing configurations of computing resources. Moreover, cloud providers have multiple usage based pricing models based on different VM configurations, such as different CPU cores, memory size, and rental costs.

When you store your photos online instead of on your home computer, or use webmail or a social networking site, you are using a “cloud computing” service. If you are an organization, and you want to use, for example, an online invoicing service instead of updating the in-house one you have been using for many years, that online invoicing service is a “cloud computing” service. Cloud computing refers to the delivery of computing resources over the Internet. Instead of keeping data on your own hard drive or updating applications for your needs, you use a service over the Internet, at another location, to store your information or use its applications. Doing so may give rise to certain privacy implications.

Cloud Computing is the delivery of computing services over the Internet. Cloud services allow individuals and businesses to use software and hardware that are managed by third parties at remote locations. Examples of cloud services include online file storage, social networking sites, webmail, and online business applications. The cloud computing model allows access to information and computer resources from anywhere that a network connection is available. Cloud computing provides a shared pool of resources, including data storage space, networks, computer processing power, and specialized corporate and user applications. The following definition of cloud computing has been developed by the U.S. National Institute of Standards and Technology (NIST): Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability and is composed of five essential characteristics, three service models, and four deployment models.

The characteristics of cloud computing include on-demand self service, broad network access, resource pooling, rapid elasticity and measured service. On-demand self service means that customers (usually organizations) can request and manage their own computing resources. Broad network access allows services to be offered over the Internet or private networks. Pooled resources means that customers draw from a pool of computing resources, usually in remote data centers. Services can be scaled larger or smaller; and use of a service is measured and customers are billed accordingly.

The cloud computing service models are Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). In a Software as a Service model, a pre-made application, along with any required software, operating system, hardware, and network are provided. In PaaS, an operating system, hardware, and network are provided, and the customer installs or develops its own software and applications. The IaaS model provides just the hardware and network; the customer installs or develops its own operating systems, software and applications.

Cloud services are typically made available via a private cloud, community cloud, public cloud or hybrid cloud. Generally speaking, services provided by a public cloud are offered over the Internet and are owned and operated by a cloud provider. Some examples include services aimed at the general public, such as online photo storage services, e-mail services, or social networking sites. However, services for enterprises can also be offered in a public cloud. In a private cloud, the cloud infrastructure is operated solely for a specific organization, and is managed by the organization or a third party. In a community cloud, the service is shared by several organizations and made available only to those groups. The infrastructure may be owned and operated by the organizations or by a cloud service provider. A hybrid cloud is a combination of different methods of resource pooling for example, combining public and community clouds.

II.RELATED WORKS

[11]Even though the data centre network is lightly utilized, virtualization can still cause significant throughput instability and abnormal delay variations. [7]The feasibility of building next-generation Cloud provisioning systems based on peer-to-peer network management and information dissemination models. [6]Workload models indicate that the proposed provisioning technique detects changes in workload intensity that occur over time and allocates multiple virtualized IT resources accordingly to achieve application QoS targets.[3]Assign each task to the resource on which the energy consumption for executing the task is explicitly or implicitly minimized without the performance degradation of that task. [8]Operating characteristics of such systems-in particular, average queue length and queueing time are evaluated. A cost structure is superimposed on the system and optimization procedures are outlined. The close relationship with priority queueing and storage models is pointed out. [4]VM-aware power budgeting uses multiple distributed managers integrated into the Virtual Power Management framework whose actions are coordinated via a new abstraction,termed the virtual power management VPM tokens. [2]The disparate energy management systems and defines a model for resource allocation that can be used for these and other energy management systems.

III.SYSTEM REPRESENTATION

A distributed service system consists of lots of physical servers, virtual machines (VMs) and a job dispatcher. The job dispatcher in system is used to identify an arrival job request and forward it to a queue of a corresponding VM manager that can satisfy its QoS levels, meet its target web application or specific requirements. When there has no job in a queue or no job is being processed, a server becomes idle and it remains until a subsequent job has been sent to its processor node. Generally, a server operates alternately between a busy mode

and an idle mode for a system with random job arrivals in a cloud environment.

A. N-Policy

A busy mode indicates that jobs are processed by a server running in one or more of its VMs' and an idle mode indicates that a server remains active but no job is being processed at that time. To mitigate or eliminate idle power wasted, three power-saving policies with different energy-efficient controls, decision processes and operating mode configurations are presented. First, try to make an energy-efficient control in a system with three operating modes m= { Busy, Idle, Sleep}, where a sleep mode would be responsible for saving power consumption.

1) ISN Policy

The energy-efficient control with the N policy is denoted by ***N***. According to the switching process from **I**dle to **S**leep and the energy-efficient control (**N** policy), called such an approach the "ISN policy".

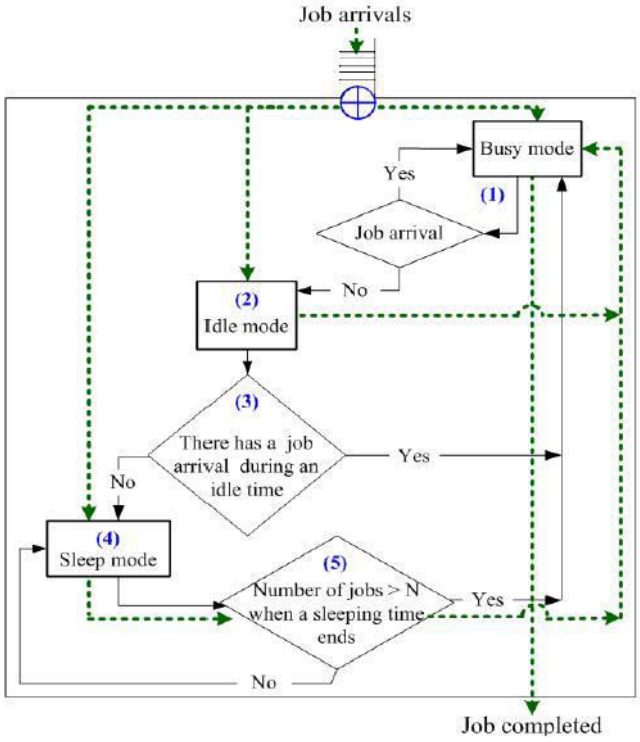

Fig:1 Decision processes of the ISN policy

2) SN Policy

A server switches into a busy mode depending on the number of jobs in the queue to avoid switching too often, the switching restriction is denoted by N . According to the switching process (directly to Sleep) and the energy-efficient control (N policy), called such an approach the "SN policy".

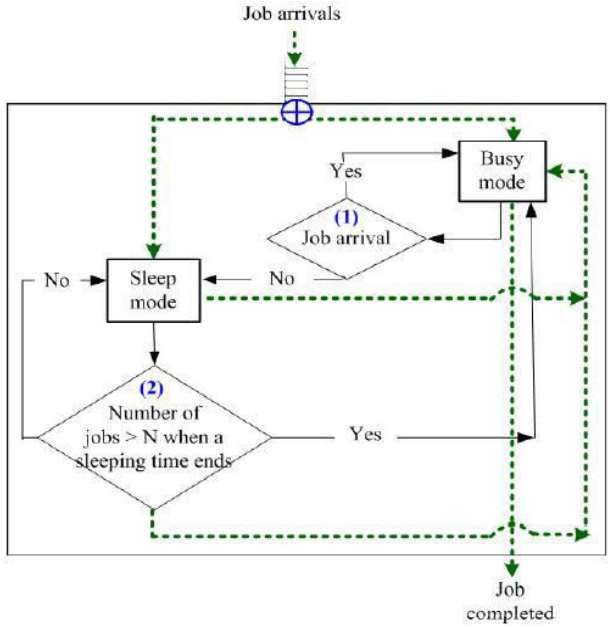

Fig:2 Decision Process of the SN Policy

3) SI Policy

A server switches into a sleep mode right away rather than an idle mode when there has no job in a system. This is similar to the SN policy but a server only stays in a sleep mode for a given time. When a sleeping time expires, it will enter into an idle mode or a busy mode depending upon whether a job has arrived in the queue or not. According to the switching process (from Sleep to Idle), called such an approach "SI policy".

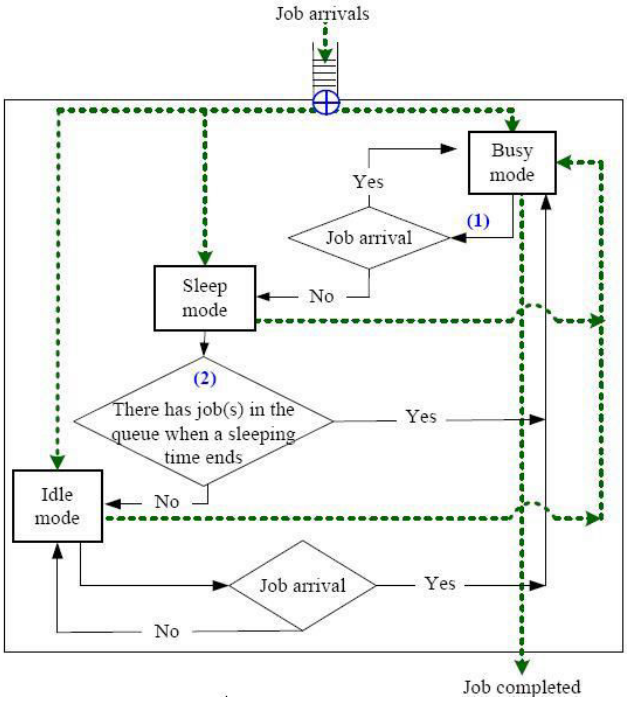

Fig:3 Decision Process of the SI Policy

IV.QUEUING MODELS

A. Scheduling Process

Scheduling is an important tool for manufacturing and engineering, where it can have a major impact on the productivity of a process. In manufacturing, the purpose of scheduling is to minimize the production time and costs, by telling a production facility when to make, with which staff, and on which equipment. Production scheduling aims to maximize the efficiency of the operation and reduce costs. Production scheduling tools greatly outperform older manual scheduling methods. These provide the production scheduler with powerful graphical interfaces which can be used to visually optimize real-time workloads in various stages of production, and pattern recognition allows the software to automatically create scheduling opportunities which might not be apparent without this view into the data. Forward scheduling is planning the tasks from the date resources become available to determine the shipping date or the due date. Backward scheduling is planning the tasks from the due date or required-by date to determine the start date and/or any changes in capacity required.

In computing, resource allocation is necessary for any application to be run on the system. When the user opens any program this will be counted as a process, and therefore requires the computer to allocate certain resources for it to be able to run. Such resources could have access to a section of the computer's memory, data in a device interface buffer, one or more files, or the required amount of processing power. A computer with a single processor can only perform one process at a time, regardless of the amount of programs loaded by the user (or initiated on start-up). Computers using single processors appear to be running multiple programs at once because the processor quickly alternates between programs, processing what is needed in very small amounts of time. This process is known as multitasking or time slicing. The time allocation is automatic, however higher or lower priority may be given to certain processes, essentially giving high priority programs more/bigger slices of the processor's time. On a computer with multiple processors different processes can be allocated to different processors so that the computer can truly multitask.Most conventional measures to save the power consumed by servers and the communication network have been discussed and implemented independently. For example, the processing of a server to reduce its power consumption can prolong not only its processing time but also the bandwidth holding time in the network, which in turn increases the power consumed by the network. Conversely, rising the processing sped of a server increases its power consumption but reduces the processing time, and consequently reduces the power consumed by the network. This may reduce the total power consumption. Therefore, it is important to take an integrated approach to saving the power consumed by both servers and the network.

B. Weighted Round Robin scheduling

Three power-saving policies follow the identical assumptions as follows. It is assumed that job request arrivals follow a Poisson process with parameter λ and they are served in order of the capacity of the server, that is, the queue discipline is based on the weighted round robin(WRR). All service times are independent and exponentially distributed with mean $1/\mu$ and the system utilization is $\rho=\lambda/\mu$, which is required to be less than one for a stable state. Idle times that those follow the exponential distribution with mean $1/\Theta i$ and follow a fixed (deterministic) time with mean $1/\Theta d$ are both considered in the ISN policy, denoted by ISN-1 and ISN-2, respectively. A sleep length is exponentially distributed with mean $1/\Theta s$ and both aforementioned variables are independent of each other. Here the state space is settled by $S = \{(n, m), 0 \leqq n < \infty, m=\{ 0, 1, 2\}\}$ where n denotes the number of jobs in the system, and m denotes that the system is in the mode m. The state-transition-rate diagram for a queuing system with the ISN-1 policy. State (0, 1) denotes that the system is in an idle mode when there has no job in the system; state (n, 0), $n \geq 1$ indicates that the system is in a regular busy mode when there have n jobs in the system; state (n, 2), $n \geq 0$ indicates that the system is in a sleep mode when there have n jobs in the system.

V.OPTIMIZATION COST FORMULATION

In general, a larger controlled N value can gain more power saving but result in excessive delay. Conversely, a smaller controlled N value can reduce delay times but lead to a shorter operational cycle. Therefore, the power consumption overhead due to server startup cannot be ignored. The operational costs and system congestion cost considered in our cost function include power consumption (service rates, operating modes and server startup) and performance degradation (congestion management cost and delay cost). The corresponding cost notations are defined and listed

.

Notation	Description
C_0	Power consumption cost when a server is in a busy mode per unit time;
C_1	Power consumption cost when a server is in an idle mode per unit time;
C_2	Power consumption cost per service rate per unit time;
C_3	Power consumption cost when a server is in a sleep mode per unit time;
C_4	Server startup cost incurred by activating a server;
C_5	Cost incurred by jobs waiting in a system per unit time;
C_6	Cost incurred by congestion management cost per unit time;

Table:1 Cost Function

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)
Vol. 2, Issue 4, April 2016

The waiting time cost C5 mainly indicates the performance penalty cost that is used to compensate for user delay experienced. On the other hand, the congestion management cost (also known as the holding cost in a queuing system) is spent to manage arrival jobs according to a service discipline and avoid a waiting queue growing without bound. Besides, the mean length of operational period, denoted by E[C], is also considered in our cost function to estimate startup cost. Since system performances and operational cost strongly depend on the service rate and mode-switching restriction, a cost objective function per unit time is developed in which both the service rate and the controlled N value are the main decision variables to address a tradeoff problem.

Furthermore, it is known that a response time guarantee is regarded as one of the most important performance concerns in designing a green cloud system since no customer wants to suffer from long delay caused by power conservation. Therefore, the SLA constraint is focused on the response time guarantee, which indicates the time from a job arrival in a buffer to the time that a job request has been processed and completed. That is, both the waiting time in the queue and the job execution time are considered.

Let *PB*, *PI* and *PS* denote the probabilities that a server is in a busy, idle and sleep mode, Let k be the number of phases in the service station. To represent a queuing model in which the offered service is a series of k identical phases, the Erlang-k service model is adopted and controlled by the SN policy. Let *L* denote the mean number of jobs in the system respectively.

The cost minimization problem can be stated mathematically as:

Minimize Fc

Where $Fc = Fc(\mu, N)$

$= C0PB + C1PI + C2\mu + C3PS + C4/E[C] + C5W + C6L$

Subject to

$0 \leq \rho \leq 1$

$W \leq x$

VI.CONCLUSION AND FUTURE WORKS

Thus this project optimize the cost and simultaneously provide the performance guarantee with improve response time.This paper plans to use the Fuzzy with Round Robin Scheduling algorithm allows cloud providers to optimize the problem in decision-making in service rate and mode-switching restriction, so as to minimize the operational cost without sacrificing a SLA constraint.

VII.REFERENCES

[1] Yi-Ju Chiang, Yen-Chieh Ouyang and Ching-Hsien Hsu " An Efficient Green Control Algorithm in Cloud Computing for Cost Optimization," in Proc. IEEE Int. Syst. Conf., 2014.

[2]G.P. Duggan and P. M. Young, "A resource allocation model for energy management systems," in Proc. IEEE Int. Syst. Conf., 2012, pp. 1–3.

[3] M. Mazzucco, D. Dyachuky, and R. Detersy, "Maximizing Cloud Providers Revenues via Energy Aware Allocation Policies," in Proc. IEEE 3rd Int. Conf. Cloud Comput., 2010, pp.131–138.

[4] Q. Zhang, M. Zhani, R. Boutaba, and J. Hellerstein, "Dynamic heterogeneity- aware resource provisioning in the cloud," IEEE Trans. Cloud Comput., vol. 2, no. 1, pp. 14–28, Jan.–Mar. 2014.

[5] M. Guazzone, C. Anglano and M. Canonico, "Energy-efficient resource management for cloud computing infrastructures," in Proc. IEEE Int. Conf. Cloud Comput. Technol. Sci., 2011, pp. 424–431.

[6]Amokrane, M. Zhani, R. Langar, R. Boutaba, and G. Pujolle, "Greenhead: Virtual data center embedding across distributed infrastructures," IEEE Trans. Cloud Comput., vol. 1, no. 1, pp. 36–49, Jan.–Jun. 2013.

[7]F. Larumbe, and B. Sanso, "A tabu search algorithm for the location of data centers and software components in green cloud computing networks," IEEE Trans. Cloud Comput., vol. 1, no. 1, pp. 22– 35, Jan.–Jun. 2013.

[8]Y. Deng, W. J. Braun, and Y. Q. Zhao, "M/M/1 queueing system with delayed controlled vacation," OR Trans., vol. 3, pp. 17–30,

1999. K. H. Wang and H. M. Huang, "Optimal control of an M/Ek/1 queueing system with a removable service station," J. Oper. Res. Soc.," vol. 46, pp. 1014–1022, 1995.

[9]Y. Levy and U. Yechiali, "Utilization of idle time in an M/G/1 queueing system," Manage. Sci., vol. 22, no. 2, pp. 202–211, 1975. [10]T. Naishuo, Z. Daqing, and C. Chengxuan, "M/G/1 queue with controllable vacations and optimization of vacation policy," Acta

Math. Appl. Sinica, vol. 7, no. 4, pp. 363–373, 1991. [11]Amokrane, M. Zhani, R. Langar, R. Boutaba, and G. Pujolle,

"Greenhead: Virtual data center embedding across distributed infrastructures," IEEE Trans. Cloud Comput., vol. 1, no. 1, pp. 36–49, Jan.–Jun. 2013.