Queuing Theoretic Approach To Server Allocation Problem In Time-delay Cloud Computing Systems

Taichi Kusaka, Takashi OKUDA, Tetsuo IDEGUCHI and Xuejun TIAN
Graduate School of Information Science and Technology, Aichi Prefectural University
Nagakute-cho, Aichi 480-1198 JAPAN.
Tel: +81-561-64-1111(Ext.3404), Fax: +81-561-64-1108, E-mail : im101005@ist.aichi-pu.ac.jp

Abstract— **Cloud computing is a popular computing model to support processing large volumetric data using clusters of commodity computers. It aims to power the next generation data centers and enables application service providers to lease data center capabilities for deploying applications depending on user QoS (Quality of Service) requirements. Because cloud applications have different composition, configuration, and deployment requirements, quantifying the performance of resource allocation policies and application scheduling algorithms, is important in cloud computing environments for different application and service models under varying load, network time- delay and system size. To obtain quantifying, the authors apply VCHS（Various Customers, Heterogeneous Servers）queuing systems.**

I. Introduction

Cloud Computing has emerged as a hot computing model to support processing large volumetric data using clusters of commodity computers[1].

Cloud Computing is the long dreamed vision of computing as a utility, where users can remotely store their data into the cloud so as to enjoy the on-demand high quality applications and services from a shared pool of configurable computing servers. By data outsourcing, users can be relieved from the burden of local data storage and maintenance.

Fig.1 shows a Cloud Computing model for enabling convenient, on-demand network access to a shared pool of configurable computing servers.

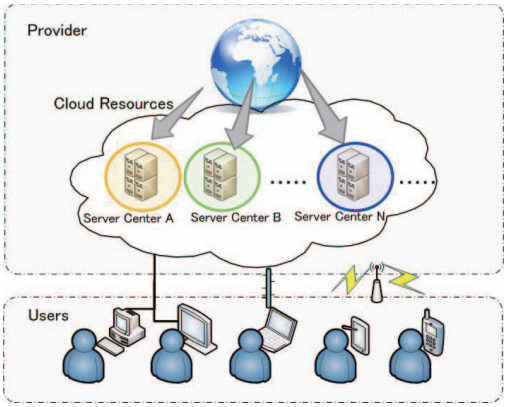

Fig. 1. Cloud Computing

A client job directed to a web site or other services hosted by a distributed pool of servers.

Cloud Computing aims to power the next generation data centers and enables application service providers to lease data center capabilities for deploying applications depending on user QoS (Quality of Service) requirements[2]. Because cloud applications have different composition, configuration, and deployment requirements, quantifying the performance of server allocation strategies and application scheduling algorithms, it is important to model different application and service under varying load, network time-delay and system size in Cloud Computing environments.

Mostly, time-delay on Cloud Computing is caused by handling routers and switches between different networks.

In order to provide Cloud Computing services economically, it is needed to optimize server allocation under the assumption that the required server can be taken from a shared some server pool.

To obtain quantifying, we will apply VCHS（Various customer, Heterogeneous Servers）queueing systems.

The outline of this paper is as follows. In Section II we will study VCHS queuing theory In Section III, we will introduce our performance evaluation model and in Section IV shows numerical examples. We will summarize at the end of this paper in Section V.

II. VCHS queuing model

Fig.2 shows VCHS(Various customer, Heterogeneous Servers) queueing model[3]. VCHS queueing model has two conditions; one is there are various types of jobs, the second is servers has various types of processing abilities. The purpose of using this model is to search optimal server allocation strategies, which minimizes the system waiting time of jobs, by evaluating performance measure(i.e. CPU utilization, the mean response time, throughput).

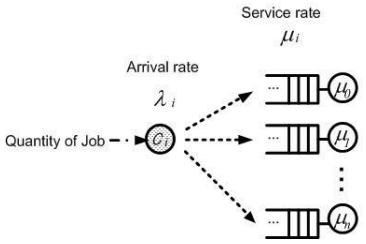

Fig. 2. VCHS queuing model

In this paper, we represent 2 server allocation strategies, strategy α and strategy β.

strategy α allocate to optimum processing speed server.

strategy β are combined with 3 components below:

① number of waiting jobs
② quantity of job
③ processing speed of servers

We try to use weight-points for quantizing above the 3 components. Jobs are allocated to servers according to the smallest total component points.

978-0-9836283-0-9

This paper was peer reviewed by subject matter experts for publication in the Proceedings of ITC 2011

Here, we will explain server allocation algorithm by using the queuing model below(show Fig.3).

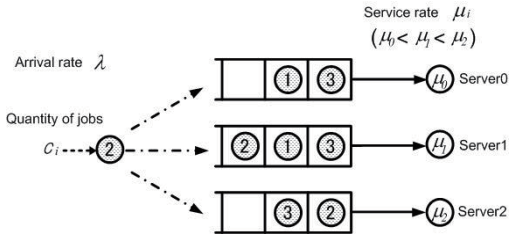

Fig. 3. example model

In this model, *Server0* has 2 waiting jobs and each quantity of job is 3 and 1. *Server1* has 3 waiting jobs and each quantity of job is 3, 1, and 2. *Server2* has 2 waiting jobs and each quantity of job is 2 and 3. When we use strategy β each waiting job has $\underline{3}$ weight-points and each quantity of job has $\underline{1}$ point. The service rate of server is not considered.

Each server's weight-points are below, and underline number is each weight-points

Server0 $(\underline{3} \times 2) + (\underline{1} \times 3 + \underline{1} \times 1) = 10$
Server1 $(\underline{3} \times 3) + (\underline{1} \times 3 + \underline{1} \times 2 + \underline{1} \times 1) = 15$
Server2 $(\underline{3} \times 2) + (\underline{1} \times 2 + \underline{1} \times 3) = 11$

According to the results above, the arrival job will be allocated to Server0 which has the smallest total points.

III. DELAY TIME MODELING

In this paper, we will use VCHS queueing model to model Cloud Computing system because it has various types of jobs and processing abilities of services[4].

But in the case of applying server allocation to Cloud Computing system, it is necessary to know that delay times until job arrives at allocated servers effects waiting time of jobs[5][6][7]. To evaluate performance of the server allocation system on time-delay Cloud Computing, we use the model that expands VCHS queue shown in Fig.4.

We assume that the jobs with some quantity C_i arrives at a server allocation queue. The jobs are allocated to servers by server allocation strategies. The system have delay time ΔT of which the jobs gets to the servers. The delay times ΔT are assumed to be independent conditional distribution random variables having mean $1/t$. The delay time ΔT have either uniform, normal, logarithmic normal or exponential distribution. The service time for each servers are assumed to be independent conditional distribution random variables having each mean μ^{-1}.

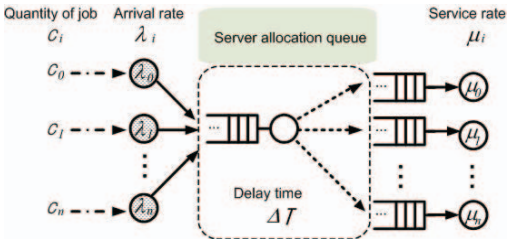

Fig. 4. evaluation model

IV. NUMERICAL EXAMPLE

In order to evaluate the performance of server allocation for Cloud Computing, we use the model shown in Fig.2. In this paper, we computed the mean time until job completes its services under the assumption of 3 types of jobs, that each quantity of job has $1 \sim 3$ in accordance with uniform distribution over $(1, 3)$, arrival rate $\lambda = 0.5$, and 3 types of servers; Server0, Server1 and Server2. Each server's rates are $\mu_0 = 1.0, \mu_1 = 2.0, \mu_2 = 4.0$.

We use CSIM20 that is a library of routines, for use with C++ programs, which allows you to create discrete-event simulation models[8].

TABLE I shows the waiting time assumption that there is no delay time and delay time ΔT are either

- regular
- normal
- log-normal
- exponential

random variable.

TABLE I
WAITING TIME ON NON DELAY-TIME SYSTEM

	mean waiting time	95%confidence interval
Non delay time		
α	2.269	2.259∼2.327
β	1.774	1.765∼1.783
Regular distribution		
α	7.767	7.757∼7.778
β	7.859	7.882∼7.909
Normal distribution		
α	5.267	5.256∼5.277
β	5.097	5.086∼5.108
Log-normal distribution		
α	3.267	3.236∼3.300
β	2.904	2.872∼2.935
Exponential distribution		
α	5.265	5.253∼5.277
β	5.164	5.151∼5.177

V. CONCLUSION

In this paper, we have applied VCHS queueing model of time-delay Cloud Computing system and evaluated it. As a result, the system waiting time of 2 server allocation strategies varies whether they have delay time or not.

REFERENCES

[1] P.Mell and T. Grance, The NIST, Definition of Cloud Computing v15, National Institute of Standards and Technology, 2009.
[2] M.Armbrust et al, Above the clouds:A berkeley view of cloud computing, Tech.Rep. UCB-EECS-2009-28, 2009.
[3] Y. Nonaka, S.Nogami etc al, "On a Evaluation for Queue Selection by Simulation", *Journal of Reliability Engineering and System Safety*, IN2008-175, pp.255-260, 2009(in Japanese).
[4] N. Sato and K.S.Trivedi, "Stochastic modeling of composite web services for closed-form analysis of their performance and reliability bottlenecks", in ICSOC 2007.
[5] Donald Gross and Carl M. Harris, *Fundamentals of Queueing Theory*, Wiley-Interscience, :3rd edition, 1998.
[6] Sheldon M.Ross, *Introduction to Probability Models*, Academic Press, 2006.
[7] Shaler Stidham Jr, *Optimal design for queueing systems*, CRC Press, 2009.
[8] Mesquite Software, `http://www.mesquite.com/`.