Cloud Server Management Method with Random Remote Backups

Kim Song-Kyoo, *Senior Member, IACSIT*

***Abstract*—A Cloud service provides the use of computing resources (hardware and software) that are delivered as a service over a network. The name comes from the use of a cloud-shaped symbol as an abstraction for the complex infrastructure it contains in system diagrams. Cloud computing entrusts remote services with a user's data, software and computation. This paper is dealing with the stochastic Cloud server management method which is focused on reliability. The (remote) backup cloud servers are hooked up by the long-haul network and replace broken main cloud severs immediately. If the Cloud servers are represented as "machines" this system can be solved by using the stochastic maintenance model with main unreliable and random auxiliary spare (remote backup) machines, subject to random breakdowns, repairs and two replacement policies: one for busy and another - for idle or vacation periods. When the repair facility is not available because of the given conditions, auxiliary machines are being used for backups. Unlike existing models, the availability of auxiliary machines is changed for activations of the system. Analytically tractable results are obtained by using a duality principle, semi-regenerative analysis, and multi-variate marked renewal processes. The results are demonstrated in the framework of optimized Cloud server allocation problems with unreliable backup servers.**

***Index Terms*—Stochastic maintenance, cloud service management, availability, stochastic optimization, closed queues, duality principle.**

I. Introduction

Optical networks provide enormous capacities in the networks and a common infrastructure over which a variety of services can be delivered. Cloud service (see Fig. 1) via Internet is one of optical network applications to interconnect computer systems with other computer systems and peripheral equipments. Even though, cloud service is operated in the Internet, the infrastructure of cloud service is identified by Fibre Channel and consists of one or more servers connected to storage devices via switching devices such as hubs, switches, and bridges. Connections between clients and storage devices are controlled by cloud servers that should be very reliable and can support high-speed connections between cloud storage and clients. Clouds operate at bit rates ranging from 200Mbps to 1Gbps over fiber optic link. While the bit rate itself is relatively modest, the cloud is attractive because of the optical layer that can support a huge number of such connections between two data centers.

Manuscript received July 17, 2013; revised September 15, 2013.

Song-Kyoo Kim is with the Asian Institute of Management, 123 Paseo de Roxas, Makati City 1260, Philippines (e-mail: SKim@aim.edu).

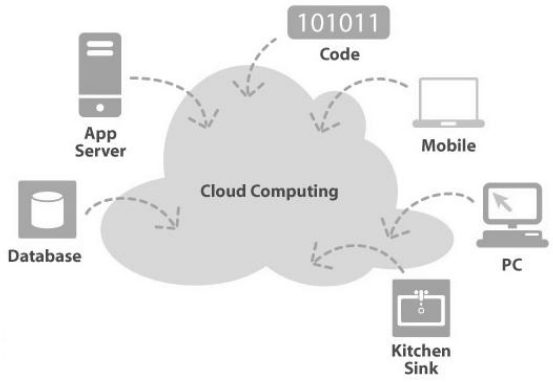

Fig. 1. Concept diagram of cloud service

Optical fibre offers much higher bandwidth than copper cables and is less susceptible to various kinds of electromagnetic interference and other undesirable effects [1]. Cloud server for resilience against disaster is a major issue of reliable network management. We assume that the remote backup Cloud servers are geometrically separated with main Cloud servers and hooked up by the long-haul networks over fiber optic links.

The stochastic model in this paper is considered to improve the availability of Cloud servers. Let us assume that we have cloud and spare cloud servers which are linked with long-haul networks. Once all $m+1$ servers become intact, the system renders a checkup. Patching applications or upgrading the operating system is a very typical pattern of maintenance. During this period, the system borrows servers with the latest version of the operating system and applications from the total quantity of from an external provider. Since the usage of external resources are random because these are also unreliable. Our model provides decision factors such as the optimal control level, the number of spare servers, failure rates, and so on, that are beneficial for the Cloud architecture. The mechanism that is mentioned above can considered as the stochastic model that consists of machines and a repair facility which fixes broken machines and goes on vacations when all machines are working properly. Unlike other existing models, the availability of auxiliary machines is changed for activations and it considers the random number of external backups for maintenance purposes beyond $M/G/1$ closed queueing system.

If each cloud server is considered as "machines", the model forms a class of closed queueing systems with the initial quantity of $m+1$ main unreliable machines, S random auxiliary reserve machines, also called "super-reserve" machines. The concept of super-reserve machines is intoduced in [2]. Main machines represent the Cloud servers and super-reserve machines mean their remote backups.

DOI: 10.7763/JACN.2013.V1.58

Main machines are subject to "exponential failures" and their repairs are rendered (in the FIFO order) by a single repair facility (also known as the "repairman") with generally distributed repair times. The random super-reserve facility is "activated" and the repairman is idle or vacationing whenever the number of main machines is restored to its original quantity. Even though the number of super-reserve machines is random, we assume that the number is fixed when the super-reserve machines are activated. During this time, dropped machines line up in the "waiting room" until the whole super-reserve facility is exhausted and one more machine breaks down. Then a repairman's busy period resumes and its cycle continues.

The classical model with unreliable machines and no backup [3] is also described in terms of closed queueing systems, i.e. those with so-called "finite sources." and the random availability of super-reserve machines makes our model different. The supplementary variable method and semi-regenerative analysis are used with additional results for semi-Markov processes and analytic solution is very appealing in optimization. The current methodology (without super-reserve facility) stems from [4]-[6] and further explored with super-reserve facility [2].

II. Stochastic Management System

The Duality principle we would like to begin with includes another model, which is more simple than the main one (Model 1) and to which we will refer as to Model 2. Model 2 is similar to Model 1, except that it does not have the backup facility and idle periods. We rather associate it with repairman's vacations, which are distributed as regular repairs. However, upon his return, the repairman brings a brand new machine, which replaces any one that breaks down during his vacation trip if any such available. Otherwise, the new machine he brings in substitutes any other machine and in both cases the old machine is disposed. Model 2 is directly connected with yet another model, which we will call Model 3. Model 3 is a regular multi-channel queueing system, in notation, (GI_0, GI)/M/m/0.

In this subsection we will describe all the above models more formally. Denote Z_t^1 the total number of intact main machines at time t in Model 1. If a repairman fixes all machines completely, the total number of main machines is restored to m. Then he goes on vacation until S+1 machines break down, after which the repairman resumes his work. In other words, a busy period begins. As we already mentioned, backup machines are replaced when main machines fail during repairman's idle period. Denote by S the random number of external backups which can be used during idle periods which have the PMF (Probability Mass Function):

$$s(n) = P\{S = n\}$$

with the mean $r = E[S]$. Let $\tau_0(=0), \tau_1, \ldots$ be the successive moments of repaircompletions. The random variable $\tau_{n+1} - \tau_n$ is supposed to have a PDF (probability distribution function).

Based on the series of mathematical calculations with duality principles, the probabilities of the cloud service system yield:

$$\pi_{m+1}^1 = E\left[\frac{P_m(n+1)}{m\mu(\bar{a} + P_m a_0 n) + P_m(n+1)}\right]$$

and

$$\pi_k^1 = E\left[(1-\phi_{m+1}^1(n))\right]\pi_k, k = 0,1,\ldots,m$$

for other status ($k=0,1,\ldots,m$). The parameters in the equation are explained by the following the stochastic calculations of Model 3. Model 3, as mentioned, is the conventional *(GI_0, GI)/M/m/0* (multi-channel) queue. Recall that such a system is characterized by the general-independent input (i.e. a renewal process), *m* parallel channels without any buffer A customer enters a free channel available with his service demand distributed exponentially with parameter μ. Inter-renewal times are distributed in accordance with the PMF *A(x)* and *A_0(x)*. Model 2, as we see it, is congruent to Model 3, while Model 1 is dual with Model 2, so also with Model 3. The latter is a classical system investigated by Takacs [9]. The stationary probabilities for the embedded process are known to satisfy the following formulas:

$$P_k = \begin{cases} \sum_{r=k}^{m} B_r \binom{r}{k} (-1)^{r-k}, k = 0,\ldots,m-1 \\ \left[a_0 \sum_{r=0}^{m} \frac{\binom{m}{r}}{a_r} \right]^{-1}, k = m \end{cases}$$

where

$$B_n = \left(a_n \sum_{r=n}^{m} \frac{{}_m C_r}{a_r} \right) \Big/ \left(a_0 \sum_{r=0}^{m} \frac{{}_m C_r}{a_r} \right)$$

$$a_r = \begin{cases} 1, r = 0 \\ \prod_{i=0}^{r} \frac{a_i}{1-a_i}, r \geq 1 \end{cases}$$

$$a(\theta) = \int_0^\infty e^{-\theta u} A(du)$$

$$a^0(\theta) = \int_0^\infty e^{-\theta u} A_0(du)$$

$$a_r = a(r\mu), a_r^0 = a^0(r\mu)$$

Now, the continuous time parameter queueing process of Model 3 is considered. By using the Kolmogorov differential equations and the semi-regenerative techniques, the limiting distribution:

$$\begin{cases} \pi_0 = 1 - \sum_{n=1}^{m} \pi_n, k = 0 \\ \pi_k = \frac{P_{k-1}}{k\mu a}, k = 1, \ldots, m \end{cases}$$

III. Optimality of the Cloud Server Management

In this section we will deal with a class of optimization problems that arise in reliability. Let us formalize a pertinent optimization problem. Let a strategy, say Σ, specify, ahead of the time, a set of acts we impose on the system, such as the choice of repair distribution, the number of main and external backups, statistics of failure rates dependent on the number of backup machines that enables us to spend more or less time on the maintenance and so on. On the other hand, a system can be subject to a set, say C, of cost functions. Denote by $\phi(\Sigma, C, t)$ the expected costs within [0, t], due to the strategy costs and define the expected cumulative cost rate over an infinite horizon:

$$\varphi(\Sigma, C) = \lim_{t \to \infty} \varphi(\Sigma, C, t)$$

Let us consider the number of external backups has a binomial distribution. For instant, the system borrows external backups but these backups are unreliable machines. The probability of each backup machine working is p (i.e., Bernoulli random variable) and the sum of backup machines yields the binomial random variables. It yields:

$$S(n) = \binom{N}{n} p^n (1-p)^{N-n}$$

and

$$\pi_{m+1}^1 = E\left[\phi_{m+1}^1(S)\right] = \sum_{n=0}^{N} \phi_{m+1}^1(n) \cdot s(n)$$

$$\pi_k^1 = E\left[1 - \phi_{m+1}^1(S)\right] \pi_k, k = 0, 1, \ldots, m$$

In addition, two primary cost functions are as follow:

$$f_1(n) = G \cdot n,\ f_2(n) = B \cdot n$$

Finally, we arrive at the following expression for the objective function:

$$\phi(\Sigma(N), C) = c \frac{P_m M_m}{PM} + r \frac{1}{PM} + (G - B)\overline{Z}_\infty^1 + B(m+1)$$

Notice that we have the parameter N varies. Even though we can not see the maximum number of external backups in the above formula, it is included in all π_k^1's. We restrict the initial strategy of this model to one, which includes only the control level N of super-reserve machines. In other words, we need to find an N such that

$$\varphi(\Sigma(n), C) = \min\{\varphi(\Sigma(n), C) : N = 1, 2, \ldots\}$$

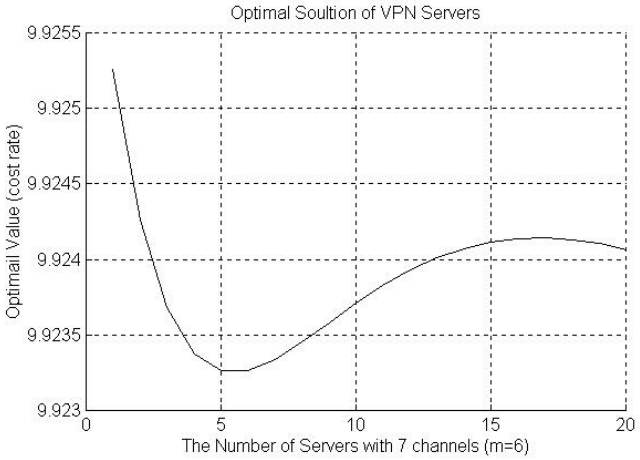

Fig. 2. Optimal solution of cloud server management

As an illustration (Fig. 2), we take c=2, r=4 and B=1. Repair time distribution is exponential with mean a=1.2 and the parameter μ=0.2. Take the total number of main machines as 5. In addition, the optimal function and N_0 that gives a minimum for the function. In other words, the control level N_0 stands for the excess level of super-reserve machines which minimizes the total cost of this cloud system. Our calculation yields that $N_0 = 5$ for which the minimal cost equals 9.9233. It means that we allocate our internal resources to 7 main (m+1) machines and obtain the decision value $N_0 = 5$ which is the number of external backups under availability p(=0.2) for a backup machine. The maximum number of external backups is needed to minimize the cost of this system.

IV. Conclusion

This paper shows the analytical approach of the cloud server management method by using the closed queueing system with flexible conditions. This approaches theoretical, but feasible to apply real-world applications such as networked server allocation management [8]. Analytically tractable results are obtained by using a duality principle, semi-regenerative analysis and multi-variate marked renewal processes. Implementation of this model and comparison between the model and the actual data will be the further research that is related with this paper.

References

[1] R. Ramaswami and K. N. Sivarajan, *Optical networks: a practical perspective*, American Press, San Diago, CA, 2002.

[2] S. Kim and J. H. Dshalalow, "A versatile stochastic maintenance model with res. and super-reserve machines," *Methodology and Computing in App. Probability*, vol. 5, no. 1, pp. 59-84, 2003.

[3] L. Takacs, *Introduction to the Theory of Queues*, Oxford University Press, New York, NY, 1962.

[4] J. H. Dshalalow, "On the Multiserver Queue with Finite Waiting Room and Controlled Input," *Adv. Appl. Prob*., vol. 17, pp. 408-423, 1985.

[5] J. H. Dshalalow, "On Single-Server Closed Queues with Priorities and State Dependent Parameters," *Queueing Systems*, vol. 8, pp. 237-254, 1991.

[6] J. H. Dshalalow, "On a Duality Principle in Processes of Servicing Machines with Double Control," *J. of Appl. Math. & Sim*., vol. 1, no. 3, pp. 245-251, 1988.

[7] E. Cinlar, *Introduction to Stochastic Processes*, Prentice Hall, Englewood Cliffs, N.J., 1975.
[8] S. Kim, "Enhanced Networked Server Management with Random Remote Backups," in *SPIE Proc. on Performance and Control of N. G. Comm, Networks*, vol. 52, no. 44, pp. 106-114, 2003.
[9] T. Kawno, S. Matsui, T. Yasue, and C. Konno, "Secure Communiction Infrastructure for Mobile," in *Hitach Review*, vol. 48, no. 1, pp. 15-20, 1999.

Song-Kyoo Kim is an associate professor of Asian Institute of Management. He had been a technical manager and TRIZ specialist of mobile communication division at Samsung Electronics. He is involved in IT industries more than 10 years. Dr Kim has received his master degree of computer engineering on 1999 and Ph.D. of operations research on 2002 from Florida Institute of Technology. He is the author of more than 20 operations research papers focused on stochastic modeling, systematic innovations and patents. He had been the project leader of several 6 Sigma and TRIZ projects mainly focused on the mobile industry.