Performance Evaluation of a Cloud Service Considering Hierarchical Failure Recovery

XIWEI QIU*, LIANG LUO and YANPING XIANG

School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, P.R. China

(Received on August 17, 2015; Revised on November 22 and December 8, 2015)

Abstract: In realistic scenarios of cloud computing, performance of a cloud service is indeed a random variable due to random resource failures. To make more precise evaluation, the random change of performance caused by resource failures and subsequent recovery should be captured. In this paper, we consider a cloud system that has a hierarchical failure recovery mechanism consisting of three typical kinds of repair. A theoretical modeling approach is presented to evaluate the performance metric that captures both service time of virtual machines (VM) and parse time of the cloud controller (CC) to ensure high fidelity. Numerical examples are illustrated.

Keywords: cloud computing, hierarchical failure recovery, performance evaluation.

1. Introduction

For providing efficient and stable cloud service, performance becomes an important metric that must be analyzed in detail. However, in practice, cloud service performance is significantly affected by random resource failures [1], [2]. To make more precise evaluation, random resource failures and subsequent recovery should be captured when analyzing performance. In this paper, we present a theoretical modeling approach for evaluating the performance of a cloud service. Since random failures always have multiple types and complicated phenomenon [3], the cloud system is assumed to adopt a hierarchical failure recovery mechanism consisting of three typical kinds of repair (including rapid, diagnostic, and complete repair). This mechanism with multiple repair strategies not only effectively handles most of minor failures, but also guarantees that some catastrophic failures are finally fixed. In realistic environments, it is more flexible and efficient than a single kind of repair. To ensure high fidelity of performance modeling, the CC parse time and the VM service time are fully taken into account. Markov reward models are applied to obtain an overall evaluation of the performance of the cloud service.

2. Description of the Cloud Service

Figure 1 describes a scenario of the cloud service. There are five available servers can be used for supporting the cloud service. The resource capacity of each server can support four co-located VMs at most. The CC keeps a request queue for the cloud service and requests are served on the First-Come-First-Serve (FCFS) discipline.

The service process of a request can be divided into two phases: a parse phase and a serving phase. In the parse phase, the CC analyzes the request to determine the amount of resources assigned to a VM, and then tries to find a server with adequate resources to host
it. In the serving phase, the VM is delivered to the user for providing the cloud service. After the VM successfully serves the user, the occupied resources are immediately released and can be used for serving the next request.

Figure 1: A Scenario of the Cloud Service.

3. Analysis of Resource Failures and Recovery

In this paper, the CC is assumed to be fully reliable. This is because that the CC usually has multiple redundant copies to guarantee its reliability. For a server in the resource pool, it can become unavailable due to random failures. The hierarchical recovery mechanism for removing a failure is designed to include three typical kinds of repair:

1) Rapid Repair: According to the phenomenon of the failure, the rapid repair based on previous experience is first adopted, which has a relative short mean repair time \(1/\theta_1\) and a probability \(p_r\) (\(p_r < 1\)) to successfully remove the failure.

2) Diagnostic Repair: If the failure cannot be removed by the rapid repair, a diagnostic repair with the mean repair time \(1/\theta_d\) is subsequently adopted. This kind of repair first runs a diagnosis program to locate the failure. According to the diagnostic results, the repair can be more pertinent and thus has a higher probability \(p_d\) to remove the failure.

3) Complete Repair: The most serious situation is that the failure cannot be removed in the first two repair phases, and the complete repair has to applied. This repair is usually time-consuming with mean repair time \(1/\theta_c\). However, it can guarantee that the occurred failure is perfectly removed (i.e., \(p_c = 1\)).

The Markov model for the hierarchical recovery mechanism is described in Figure 2. The failure time \(1/\eta\) and the repair times are assumed to be exponentially distribution [4]. The state 1 represents that the server is available. Without loss of generality, the inequalities \(\theta_1 > \theta_2 > \theta_3\) and \(p_r < p_d < 1\) are held. The steady probability for the server being available (denoted by \(p_1\)) can be obtained by solving the Chapman-Kolmogorov equations of the model [5]. Given \(n\) servers in the resource pool, the number of available servers is a random variable denoted by \(Y\). The probability mass function (pmf) of \(Y\) can be derived by

\[
p(y) = \Pr(Y = y) = C^n_y (p_r)^y (1 - p_r)^{n-y} (1)
\]

Figure 2: Markov Model for the Hierarchical Recovery Mechanism.
4. Evaluation of Service Performance

Suppose the CC parse time and the VM service time are exponentially distributed with rates $\mu_i$ and $\mu_r$, respectively. The arrival of requests for the cloud service is assumed to follow a Poisson process with an arrival rate $\lambda$, and the limitation of the request queue length is $L$. If there are $y$ available servers and each server supports $c$ co-located VMs at most, the cloud service can serve $x = c \cdot y$ users in parallel. The performance model of the cloud service is depicted in Figure 3.

**Figure 3:** Performance Model of the Cloud Service.

States of the model are indexed by $(i, j)$, where $i$ denotes the number of requests in the queue and $j$ represents the number of running VMs. Note that $b_i$ describes that one request has been parsed by the CC but cannot be served due to inadequate resources, which makes the CC remain ‘busy’ state and the other $k$ requests have to wait in the queue. Denote $\pi_{i,j}$ as the steady probability for the model stay at $(i, j)$, which can be obtained by using Markov reward approach [5]. Then, the conditional probability that a new arrival request is discarded due to the full queue can be calculated by

$$p_{dis}(x) = \sum_{j=0}^{\infty} \pi_{E,j} + \pi_{B,j} + x$$  \hspace{1cm} (2)

Note that $x = c \cdot y$ and the pmf of the random variable $Y$ can be obtained from (1). Thus, we can use Markov reward models (MRM) to remove the condition of $x$. For each state $x$, it has a probability $p_X(x) = p_Y(x/c)$ and a reward value $r(x) = p_{dis}(x)$. Use a Bayesian approach, the expected reward value can be calculated by

$$r = \sum_x p_X(x) r(x).$$ \hspace{1cm} (3)

The performance metric is defined as the expected throughput of the cloud service as

$$\varphi = \lambda (1 - p_{dis}).$$ \hspace{1cm} (4)

5. Numerical Examples

A simulation program has been developed to verify the proposed theoretical model. The corresponding parameters are listed as follows (the rates take the unit of per second):

1) Failure and recovery: $\eta = 0.0009$, $\theta_t = 0.01$, $\theta_r = 0.006$, $\theta_p = 0.001$, $p = 0.75$, and $p_d = 0.85$.

2) Cloud service 1: $\lambda = 2.6$, $\mu_i = 2.8$, $\mu_r = 0.2$, $L = 10$, $c = 4$, and $n = 5$. 

According to our theoretical model, the performance metrics of the cloud services are calculated as $\phi_1 = 2.3525 \text{ s}^{-1}$ and $\phi_2 = 1.3320 \text{ s}^{-1}$. The experimental data obtained by running the simulation program 150 times is shown in Figure 4. It can be seen that the experimental data fluctuates around the corresponding theoretical values, which verifies the proposed performance model.

**Figure 4**: Experimental Data of 150 Runs of the Simulation Program.

**Conclusions**

In this paper, we systematically study a theoretical model for evaluating the performance of a cloud service. To achieve more precise evaluation, random resource failures and subsequent failure recovery are captured in our model. We analyze a hierarchical recovery mechanism consisting of three typical kinds of repair, which is more flexible and efficient than a single kind of repair. Moreover, both the CC parse time and the VM service time are taken into account to establish a more realistic performance model. Finally, the proposed performance model is verified by simulation results.

**Acknowledgement**: This work was supported in part by the National Natural Science Foundation of China under Grant 61170042, in part by the Fundamental Research Funds for the Central Universities under Grant ZYGX2011Z001, and in part by the Innovational Team Project of Sichuan Province under Grant 2015TD0002.

**References**


Xiwei Qiu is currently pursuing the Ph.D. degree in computer science with the School of Computer Science and Engineering, University of Electronic Science and Technology of China (UESTC).

Liang Luo received his Ph. D. degree in computer science from Beihang University in 2014. He is a Lecturer with the School of Computer Science and Engineering, UESTC.

Yanping Xiang received her Ph.D. degree in system engineering from the National University of Singapore in 2003. She is a Professor with the School of Computer Science and Engineering, UESTC.