

# Modeling Approach Combining Performance and Reliability for Mobile Cloud System

Han Xu<sup>a</sup>, Haiqing Wang<sup>b</sup>, Liang Luo<sup>a,\*</sup>, Xiwei Qiu<sup>a</sup>, Sa Meng<sup>a</sup>, and Xun Guo<sup>a</sup>

<sup>a</sup>University of Electronic Science and Technology of China, Chengdu, 610000, China

<sup>b</sup>The 28th Research Institute of China Electronics Technology Group Corporation, Nanjing, 210007, China

---

## Abstract

In recent years, cloud computing has been widely used to improve the quality of mobile services in wireless networks. The ability of cloud computing gives the new mobile cloud system (MCS) the power to provide high-performance services. However, these services may be interrupted by random resource failures, such as network failures and virtual machine (VM) failures, which extremely affect the service performance in real-life scenarios. In this paper, we present a theoretical modeling approach to evaluate the service performance of the MCS, which takes into consideration the impact of reliability on service performance. The queue theory and the Markov chain are first used to model the performance and the reliability. In reliability modeling, the network failure and VM failure are considered. Then, a Bayesian method is applied to connect the performance and the reliability, and a correlation metric is proposed. Finally, numerical examples are illustrated.

**Keywords:** quality of service; cloud radio access network; reliability modeling; performance modeling

(Submitted on May 16, 2018; Revised on June 20, 2018; Accepted on July 21, 2018)

© 2018 Totem Publisher, Inc. All rights reserved.

---

## 1. Introduction

In recent years, mobile interconnection has become increasingly popular. More and more smart devices are connecting to the Internet and producing huge amounts of service requests, which poses a great challenge to existing wireless networks. In addition, an increasing number of computing-intensive and data-intensive mobile services has been developed to meet business requirements, which puts forward a higher demand for the network system. To address these challenges, cloud computing has been combined with wireless networks. This combination creates two amazing technologies: the cloud radio access network (C-RAN) and mobile cloud computing (MCC). The MCS is a typical system, formed from the integration between the C-RAN and the MCC. Nowadays, the C-RAN and MCC are well-integrated and have made great achievements. However, this combination also brings great challenges, such as the difficulty of evaluating the quality of service (QoS) that results from the enhanced sophistication of scheduling. The increasing number of influence factors is the main reason behind these challenges. Therefore, it is important to comprehensively analyze these factors and evaluate their impact on the MCS.

Some researchers have built an evaluation model to evaluate the impact of different influences factors on different QoS metrics of the MCS, such as performance, energy, and price. Wang et al. [1] present a unifying framework for optimizing the power-performance tradeoff of MSP by jointly scheduling network resources in the C-RAN and computation resources in the MCC to minimize the power consumption of MSP while still guaranteeing the QoS for mobile users. Tong et al. [2] design the edge cloud as a tree hierarchy of geo-distributed servers to deploy cloud servers and propose a workload placement algorithm. Simulations prove that the authors' work can efficiently utilize cloud resources to serve peak loads from mobile users and achieve better performance. Cai et al. [3] study the topology configuration and rate allocation problem in the C-RAN with the objective of optimizing the end-to-end performance of MCC users in next-generation wireless networks. They also prove that the design and operation of future mobile wireless networks can be significantly affected by cloud computing through performance experiments. However, they fail to consider reliability. In fact, different types of failures and errors may occur in the process of service providing, which makes the service unreliable. For example,

\* Corresponding author.

E-mail address: [luolianng@gmail.com](mailto:luolianng@gmail.com)

in the C-RAN, the wireless transmission of the signal makes the signal susceptible to interference. In addition, the mobility of devices may also cause signal instability. Due to these two factors, request errors may occur in the C-RAN [3]. Moreover, network failures and VM failures may occur in the MCC, which may cause the request to retransmit, delay, or fail. If any one of these failures occurs, it may have a significant impact on the QoS metric, such as the service performance. Thus, reliability is an indispensable factor for precisely evaluating the service performance of the MCS.

A solid mathematical background supports the theoretical modeling ability to catch the complexity and randomness of factors. Theoretical modeling is one of the most appropriate ways to conduct an evaluation. Therefore, in this paper, we build a theoretical modeling approach to evaluate the service performance of the MCS, which takes into consideration the impact of reliability on service performance. Two reliability factors are considered, including the VM failure (VM reliability), and the network failure (network reliability). The Markov model, the queuing theory, and the Bayesian approach are used to model and connect the reliability and the service performance.

The main contributions of the current study can be listed as follows:

- A theoretical modeling approach is proposed, which models the service performance of the MCS and takes into consideration the impact of reliability on service performance in the MCS.
- A large set of influence factors, such as the number of VMs, the request arriving rate, and the service rate, are considered in this approach to improve its fidelity. Two reliability factors, the VM failure and the network failure, are also considered.
- A correlation metric that can precisely evaluate the service performance of the MCS is presented.
- The impact of reliability on service performance of the MCS is revealed through experiments.

The rest of this paper is organized as follows. The physical structure and the logical structure of the MCS are described in section 2. In section 3, the performance and the reliability modeling approach are given, and the correlation metric is proposed. Section 4 gives numerical experiments. Finally, section 5 provides a conclusion for this paper.

## 2. The Structure of the MCS

In this part, the physical structure of the MCS is first given. Then, the logical structure of the MCS and the parameter definition of it are introduced.

Figure 1 shows the physical structure of the MCS, which consists of the C-RAN and the MCC. The MCC is the combination of cloud computing, mobile computing, and wireless networks [4]. In the MCC, virtualization technology has been widely used. The deployment of numerous VMs brings powerful computation and store abilities to the MCC. The C-RAN is a centralized, cloud computing-based architecture for radio access networks, which consists of several remote radio heads (RRHs), fronthaul links, and a BBU pool [5-6]. RRHs receive the user requests and transport these requests to the BBU pool through fronthaul links. Finally, the BBU pool will send these requests to the MCC [7]. By integrating the C-RAN with the MCC, the MCS can not only handle the increasing mobile traffic, but also enhance the capabilities of mobile devices [1].

Figure 2 shows the logical structure of the MCS. As shown in Figure 2,  $a_m kb \cdot s^{-1}$  is defined as the requests transmitting rate between user  $m$  and RRH  $m$ .  $b_m kb \cdot s^{-1}$  is defined as the bandwidth of the fronthaul link between RRH  $m$  with BBU pool. Assume that the TCP protocol is used for requests transmitting (multiple requests form a TCP packet) in the C-RAN.  $I_{tcp}$  represents the link frame size of one TCP packet. Then, the average throughput of C-RAN  $\eta_m$  can be calculated by applying the TCP transmission model proposed in [8]. This TCP transmission model will be specifically introduced in the next section. Here, we ignore the transmission delay between the MCC and the C-RAN. Thus, the requests arriving rate in the MCC is approximately equal to the  $\eta_m$ . In the MCC,  $B_k$  is defined as the execution capability of the  $k$ -th VM. The  $B_k$  means that the  $k$ -th VM can execute  $B_k$  TCP packets per unit time. For convenience, we quote the content in [9] and assume that the number of VM are limited and  $NOS_m$  represents the number of VMs that were assigned to perform requests from user  $m$ . In [9], it is mentioned that VMs that handle the same type of tasks usually have the same configuration in cloud. Finally, the MCC will send the results back to user  $m$ .

As mentioned before, network failures and VM failures may occur in the MCS. Here, some parameters are defined to describe these failures.  $\alpha$  is defined as the failure rate of the network. The failure rate  $\beta$  and the repair rate  $\gamma$  are defined to describe VM failures, which will be further discussed in section 3.

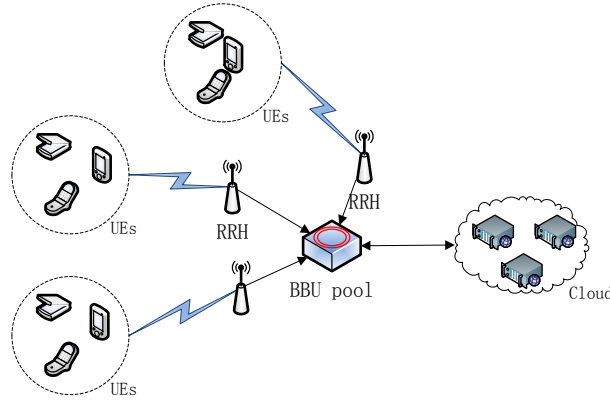


Figure 1. The physical structure of the MCS

In the following, the performance-reliability modeling for the MCS will be given.

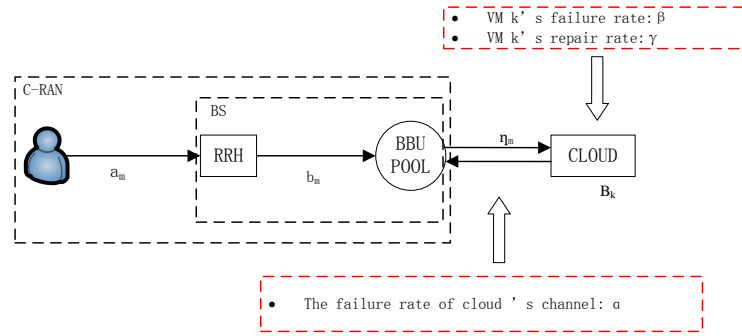


Figure 2. The logical structure of the MCS

### 3. Performance-Reliability Analysis

In this part, the method to calculate the average throughput of C-RAN is first given. Then, the performance and reliability modeling for the MCS are introduced. Finally, a new performance metric is proposed, which is a correlation metric that takes into account the impact of reliability on service performance.

#### 3.1. The Average Throughput of C-RAN

Here, we quote the approach in [3, 8] to calculate the average throughput of C-RAN. Padhye et al. [8] develop a model for TCP connections in RAN and present the expression to calculate the average throughput of RAN. Applying this approach, the average throughput of C-RAN  $\eta_m$  is calculated by

$$\eta_m \approx \min\left\{\frac{W_{max}}{RTT_m}, \frac{1}{RTT_m \sqrt{\frac{2bp_{e,m}}{3}} + T_0 \min\{1, 3V\}}\right\} \quad (1)$$

In Equation (1),  $W_{max}$  represents the maximum congestion widow.  $T_0$  is the initial time-out for the TCP sender.  $V$  is defined as:

$$V = \sqrt{\frac{2bp_{e,m}}{8}} p_{e,m} (1 + 32p_{e,m}^2)$$

Where  $b$  is the number of packets that are acknowledged by a received ACK, and its number is typically 2 [8]. These parameters are predefined, except for  $RTT_m$  and  $p_{e,m}$ . The  $RTT_m$  and the  $p_{e,m}$  are the round trip time (RTT) and the request error rate, respectively. The method of calculating the  $RTT_m$  and  $p_{e,m}$  are given as follows.

Calculating  $p_{e,m}$ : In this approach [8], the hybrid automatic repeat request (HARQ) is used in the link layer. Assaad et al. [10] propose that the number of requests transmitted follows a Gaussian distribution  $N(\mu_m, \sigma_m^2)$  for HARQ. Thus, we have

$$\mu_m = \frac{1 + p_{1,m} - p_{1,m}p_{2,m}}{1 - p_{1,m}p_{2,m}} \quad (2)$$

$$\sigma_m^2 = \frac{p_{1,m}(1 - p_{1,m} + p_{1,m}p_{2,m})}{1 - p_{1,m}p_{2,m}} \quad (3)$$

In Equation (2) and Equation (3),  $p_{1,m}$  represents the probability of error after decoding the information block by forward error correction.  $p_{2,m}$  represents the probability of error after soft combining two successive transmissions of the same information block. It is clear that the request error occurs when the number of requests transmitted exceeds the bandwidth of fronthaul  $b_m \text{ bytes} \cdot s^{-1}$ . Therefore, according to the approach in [3],  $p_{e,m}$  can be obtained as

$$p_{e,m}(b_m \cdot (l_{tcp})^{-1}) = Q\left(\frac{b_m \cdot (l_{tcp})^{-1} - \mu_m}{\sigma_m}\right) = \frac{1}{\sqrt{2\pi}} \int_{\frac{b_m \cdot (l_{tcp})^{-1} - \mu_m}{\sigma_m}}^{\infty} e^{(-\frac{t^2}{2})} dt \quad (4)$$

Where  $Q(\cdot)$  is the well-known Q-function.

Calculating  $RTT_m$ : Pathak et al. [11] point out that RTT consists of  $T_{fronthaul}$  and  $T_{backhaul}$ . As shown in Figure 2,  $RTT_m$  is the time between the user and the BBU pool. Thus, we have

$$RTT_m = T_{fronthaul} + T_{backhaul} \quad (5)$$

$$T_{fronthaul} = \mu_m \frac{l_{tcp}}{a_m} \quad (6)$$

Where  $T_{fronthaul}$  represents the average transmission time of a TCP packet over wireless links between the user and the RRH. In (6),  $\mu_m$  represents the number of requests in the wireless link and  $\frac{l_{tcp}}{a_m}$  represents the time of transferring one request.  $T_{backhaul}$  is an predefined value, which is decided by the wireless network cloud in the C-RAN [3].

Now, the  $\eta_m$  is solved. Note that the formula derivation in this part has been proven in [8]. In the next section, we introduce the performance and reliability modeling.

### 3.2. Performance and Reliability Modeling in MCS

**Performance modeling:** The queuing theory has been widely used to simulation random events execution in real input and output systems. In this paper, the M/M/S queuing model is used to describe the execution of the request in the MCC. M/M/S means that the arrival of requests follows the Poisson distribution and the service time of requests follows the exponential distribution. Therefore, the arriving time of requests follows the Poisson distribution with parameter  $\lambda = \eta_m / l_{tcp}$ , and the service time of requests follows the exponential distribution with parameter  $\mu_i = i \times B_k$ . The  $\mu_i$  is the service rate when there are  $i$  available VMs in the MCC,  $i \leq L$ .  $L$  represents the maximum length of the request queue in the MCC. We also define  $T_d$  as the due time for each request. If the sojourn time of one request exceeds its due time  $T_d$ , the time-out failure occurs. Then, we can build the birth-death process for requests and calculate the performance metric of the MCS. The birth-death process of requests is shown in Figure 3.

Here,  $\delta_m(x)$  is defined as the performance metric for the MCS.  $\delta_m(x)$  represents the completing rate of requests per unit time when there are  $x$  available VMs in the MCC,  $x \leq NOS_m$ . It is clear that the request effective arrival and successful execution are two factors that ensure the completion of the request. Then, we have

$$\delta_m(x) = \zeta_e(1 - p_{time-out}) \quad (7)$$

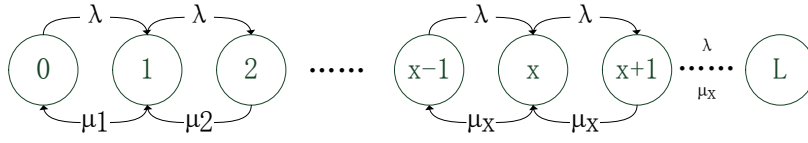


Figure 3. Birth-death process of request

$(1 - p_{time-out})$  represents the probability of successfully executing the request during  $T_d$ .  $\zeta_e$  denotes the effective arrival rate of new requests without abandonment, which can be obtained as

$$\zeta_e = \lambda(1 - q_L) \quad (8)$$

$\lambda$  represents the arrival rate of requests.  $q_L$  represents the probability that the request queue is full. When the request queue is full, the new arrival request will be abandoned. Thus,  $(1 - q_L)$  represents the probability that the request arrives effectively without abandon. Denote  $q_i$  as the probability when there are  $i$  requests in the queue. Solving the birth-death process shown in Figure 3,  $q_i$  can be derived as

$$q_0 = [1 + \sum_{i=1}^{x-1} \frac{(\lambda)^i}{\prod_{r=1}^i \mu_r} + \sum_{i=x}^L \frac{(\lambda)^i}{\mu_x^{i-x} \prod_{r=1}^x \mu_r}]^{-1} \quad (9)$$

$$q_i = \frac{(\lambda)^i}{\prod_{r=1}^i \mu_r} q_0, \quad 1 \leq i < x \quad (10)$$

$$q_i = \frac{(\lambda)^i}{\mu_x^{i-x} \prod_{r=1}^x \mu_r} q_0, \quad x \leq i \leq L \quad (11)$$

According to Equation (11),  $q_L$  is solved. We have

$$q_L = p_{block} = \frac{(\lambda)^L}{\mu_x^{L-x} \prod_{r=1}^x \mu_r} q_0 \quad (12)$$

$p_{block}$  denotes the probability of the blocking failure occurring.

For  $(1 - p_{time-out})$ ,  $p_{time-out}$  is the probability that the request is not completed during the due time  $T_d$ . Therefore, we have

$$p_{time-out} = \Pr\{T_s > T_d\} = 1 - \int_0^{T_d} f_s(t) dt \quad (13)$$

$T_s$  represents the sojourn time of a request, from when the request is put into the request queue until it is completed.  $T_s$  equals the sum of the waiting time  $T_w$  and the execution time  $T_e$ . Denote  $f_s(t)$  as the probability density function (pdf) of the sojourn time  $T_s$ . Then,  $f_s(t)$  can be obtained as

$$f_s(t) = \sum_{i=0}^x \Pr(i) \cdot f_e(t) + \sum_{i=x+1}^{L-1} \Pr(i) \cdot f_{i-x+1}(t) \otimes f_e(t) \quad (14)$$

$\sum_{i=0}^x \Pr(i) \cdot f_e(t)$  shows that the new arrival request is executed immediately.  $\sum_{i=x+1}^{L-1} \Pr(i) \cdot f_{i-x+1}(t) \otimes f_e(t)$  represents the new arrival request being executed until the  $i - x + 1$  request in front of it is executed, where  $\otimes$  is the convolution operator of the two functions.  $\Pr(i)$  is the probability that there are  $i$  requests in the queue when a new request arrives. Therefore,  $\Pr(i)$  is a condition probability, which can be computed by

$$Pr(i) = Pr\{i | i < L\} = \frac{Pr\{i, i < L\}}{Pr\{i < L\}} = \frac{q_i}{1 - q_L}, i = 0, 1, \dots, L - 1 \quad (15)$$

As mentioned above, the service time of requests follows the exponential distribution. Use  $f_e(t)$  as the pdf of the execution time  $T_e$  with parameter  $\mu_r$ . Then,  $f_e(t)$  can be calculated by

$$f_e(t) = \mu_i \cdot e^{-\mu_i t} \quad (16)$$

$f_{i-x+1}(t)$  is used as the pdf of the waiting time  $T_w$ .  $f_{i-x+1}(t)$  belongs to the Erlang distribution and represents the pdf of the time when the newly arrived request starts to be processed until  $i - x + 1$  requests in front of it have been completed. Then, we have

$$f_{i-x+1}(t) = \frac{(\mu_x t)^{i-x}}{(i-x)!} \mu_x \cdot e^{-\mu_x t}, t \geq 0, x < i < L \quad (17)$$

Now,  $\delta_m(x)$  is solved. Next, the reliability modeling approach will be introduced.

**Reliability modeling:** In this part, the VM reliability and the network reliability are taken into consideration. The reliability modeling approach is also given.

For VM reliability, the failure rate and repair rate of VM are used to describe it. Xie et al. [12] point out that the VM failures can be assumed to follow Poisson processes, which can be explained as being either within the operational phase or in a steady state after a long-time run. Trivedi et al. [13] claim that it has also been widely accepted that the hardware repair time is in accordance with the exponential distribution. Thus, the approaches shown in [12-13] are quoted in this paper. The failure rate of the VM is defined as  $\beta$ , and the repair rate of the VM is defined as  $\gamma$ . As mentioned above, the VMs in the MCC are the same. Therefore, it is reasonable to assume that each VM in the MCC has the same failure rate and repair rate. Figure 4 shows the state transition model of the number of available VMs in the MCC.

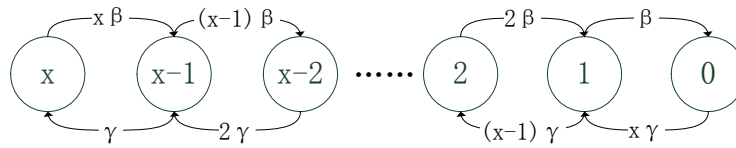


Figure 4. The state transition model of the number of available VMs in the MCC

The state  $x (x = NOS_m, NOS_m - 1, \dots, 1, 0)$  represents the number of available VMs. Use  $\pi_x$  as the steady probability for the state  $x$ . Note that  $\pi_x$  is one of the metrics to evaluate the reliability of the MCC. Then,  $\pi_x$  can be solved by applying Chapman–Kolmogorov equations. We have

$$\pi_0 = \left[ \sum_{i=0}^{NOS_m} C_{NOS_m}^i \left( \frac{\gamma}{\beta} \right)^i \right]^{-1} \quad (18)$$

$$\pi_x = C_{NOS_m}^x \left( \frac{\gamma}{\beta} \right)^x \pi_0, (x = 1, 2, \dots, NOS_m) \quad (19)$$

Then, the average number of available VMs in the MCC can be calculated through  $\pi_x$ . Use  $E(VM)$  as the average number of available VMs in the MCC. We have

$$E(VM) = \sum_{x=0}^{NOS_m} x \cdot \pi_x \quad (20)$$

For network reliability, the status of the link is one of the most important metrics [14]. Link anomalies can cause

receiving failures and output failures in the MCC. Quoting approaches in [14], the failure rate of the link is used to describe the network reliability.  $\alpha$  is defined as the failure rate of the link between the C-RAN and the MCC.  $RN(t)$  represents the network reliability. Then, we have

$$RN(t) = \exp\{-\alpha \cdot t\} \quad (21)$$

In Equation (21),  $t$  is the total communication time for transferring each request in the MCS.  $RN(t)$  represents the probability that the link between the C-RAN and the MCC is available during time  $t$ .

### 3.3. Correlation Metric

In this part, a correlation metric is proposed, which is a new performance metric for the MCS.  $E(\delta_m(x))$  is used as the correlation metric, where the impact of reliability on service performance is taken into consideration.  $E(\delta_m(x))$  represents the throughput ability of the MCS, which equals the number of completed requests per unit time. Applying the Bayesian theory to combine the performance and reliability, we have

$$E(\delta_m(x)) = RN(T_d) \cdot \sum_{x=1}^{NOS_m} \delta_m(x) \cdot \pi_x \quad (22)$$

Let us briefly explain the  $E(\delta_m(x))$ . The request completion must satisfy the following two conditions at the same time: the link is reliable during the request due time  $T_d$  and the request is completed during the request due time  $T_d$ . For link reliability,  $RN(T_d)$  in Equation (22) represents the probability that the link is reliable during the request due time  $T_d$ . For request completion,  $\delta_m(x)$  represents the completing rate of requests per unit time, which is affected by the number of available VMs  $x$ .  $x$  ranges from 0 to  $NOS_m$ , with different probabilities  $\pi_x$ . Therefore, taking all the circumstances into consideration and applying the Bayesian theory, the request completion rate of the MCS is obtained and equals  $\sum_{x=1}^{NOS_m} \delta_m(x) \cdot \pi_x$ . Finally, Equation (22) is calculated by combining the link reliability and the request completion function.

In the next section, some experiments are illustrated to verify the correctness of our modeling approach.

## 4. Experiment

In this section, two simulations are first designed to verify the performance-reliability modeling approach. Then, numerical examples are made to reveal the impact of reliability on performance.

### 4.1. Verification

Two simulations simulate the whole executing process of requests in the MCS, including the transmission process in the C-RAN and the MCC and the execution process in the MCC. Table 1 shows the parameters assignment. Note that the parameters assignment used in this example are only for illustration. Other parameters assignments can also be implemented in a similar way for more realistic scenarios.

Table 1. The parameters assignment

Parameters	Value
$m$	1
$a_m$	$6195.2\text{kb} \cdot \text{s}^{-1}$
$I_{tcp}$	512kb
$\alpha$	$0.0006\text{s}^{-1}$
$\beta$	$0.0004\text{s}^{-1}$
$\gamma$	$0.0015\text{s}^{-1}$
$L$	12
$T_d$	6s

For convenience, we assume that only one user sends requests to the MCS. Therefore,  $m = 1$ . These simulations can also be easily extended to a simulation that contains several users. The transmission process for each user is independent, and the C-RAN will send requests to the MCC after gathering requests from all users. Therefore, the request arrival rate in

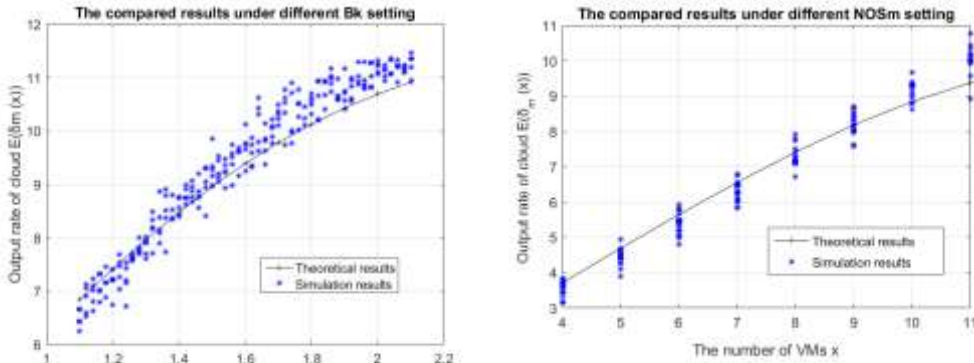
the MCC is approximately equal to the sum of the sending rate of each user under the allowable bandwidth in the C-RAN.

Two simulations are designed. In the first simulation, we specify  $B_k$  with a range of  $1.1s^{-1}$  to  $2.1s^{-1}$  and fix other parameters. In the second simulation, we specify  $NOS_m$  with a range of 20 to 30 and keep other parameters unchanged. Some brief introductions are given for these two simulations:

- In the first simulation,  $NOS_m$  is constant and  $NOS_m = 8$ . The simulation runs  $5 \times 55$  times, where 55 represents  $B_k$  having 55 possible values (the value of  $B_k$  range from  $1.1s^{-1}$  to  $2.1s^{-1}$  plus  $0.02s^{-1}$  for each time) and 5 represents repeat times.
- In the second simulation,  $B_k$  is constant and  $B_k = 1.2s^{-1}$ . The simulation runs  $50 \times 10$  times, where 10 represents  $NOS_m$  having 10 possible values and 50 represents repeat times.
- There are 3000 loops for each time. Here, each loop represents 1 second in the real environment. This means that the system runs 3000 seconds each time.
- One random number generator is given. At each loop, the generator will generate a random number. This random number ranges from 0 to 1. If the value of the request error or the link failure is less than this random number, the error or failure is happening and the request will be retransmitted in the next loop.
- When the sojourn time of the request in the queue exceeds the maximum sojourn time  $T_d$ , this request is a failure and it will be abandoned.
- After 3000 loops, we will count the number of complete requests and calculate the request completion rate. Then, we will adjust the value of  $B_k$  or  $NOS_m$  and begin the next run.

In these two simulations, the results are compared with the theoretical results calculated by the performance-reliability modeling approach. Both the simulations and theoretical calculations are configured with the same parameters. Figure 5 shows the compared results between the simulations and theoretical calculations.

The comparison results in Figure 5 show that the theoretical calculation results closely match the simulation results, which proves the accuracy of the performance-reliability modeling approach. In the next section, the performance-reliability model approach is used to study the impact of reliability factors on service performance.



(a) The compared results under different  $B_k$  setting (b) The compared results under different  $NOS_m$  setting  
Figure 5. (a) The compared results under different  $B_k$  setting; (b) The compared results under different  $NOS_m$  setting

#### 4.2. Numerical Examples

Now, we run experiments to observe changes of the service performance caused by reliability factors. Network reliability has a negative correlation with service performance. Therefore, we only run experiments to show the impact of VM reliability on service performance. The VM failure rate  $\beta$  is changed from  $0.00015s^{-1}$  to  $0.00035s^{-1}$ . Other parameter settings are shown in Table 1 and Table 2.

Table 2. The parameters setting

Parameters	Value
$NOS_m$	5
$B_k$	$2.5s^{-1}$

Figure 6 shows the value change of the service performance.



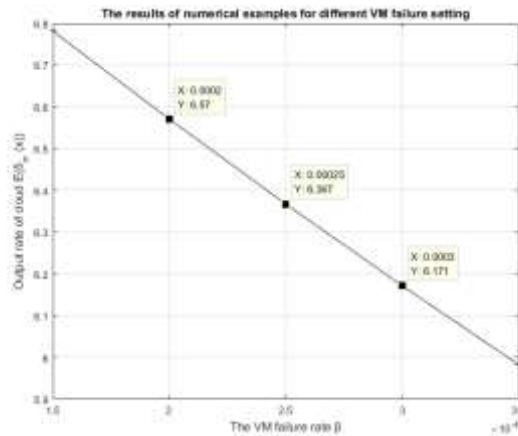


Figure 6. The value change of the service performance caused by different VM failure value

Figure 6 shows that for each increase of 0.05% of the VM failure rate, the service performance (output capability) decreases by 3%. The experimental results prove that reliability is an unavoidable factor for the MCS.

## 5. Conclusions

In this paper, the reliability-performance modeling approach is proposed for the MCS. Two reliability factors are considered: the network reliability and the VM reliability. We first complete the performance and reliability modeling for the MCS. Then, a correlation metric is obtained by using the Bayesian method, which reflects the impact of reliability factors on the service performance. In the experimental part, verifications are first made for the reliability-performance modeling approach. Then, we alter the value of the VM failure rate and observe the changes of the service performance in the MCS. The results show that for each increase of 0.05% of the VM failure rate, the service performance (output capability) decreases by 3%. This proves that reliability does have a significant effect on service performance and it is an unavoidable factor for making more precise performance evaluations. In our future work, we will focus on designing resource scheduling strategies for optimizing the service performance in the MCS. Moreover, clouds are often geo-distributed. Different transmission distances, network latency, network bandwidth, and other factors result in a different requests reception rate for each cloud. Therefore, we can try to extend our approach to this circumstance for appropriate scheduling.

## Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 61602094).

## References

1. X. Wang, K. Wang, S. Wu, S. Di, and K. Yang, "Dynamic Resource Scheduling in Cloud Radio Access Network with Mobile Cloud Computing," in *Proceedings of IEEE/ACM 24th International Symposium on Quality of Service (IWQoS)*, pp. 1-6, 2016
2. L. Tong, Y. Li, and W. Gao, "A Hierarchical Edge Cloud Architecture for Mobile Computing," in *Processings of IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications*, pp. 1-9, 2016
3. Y. Cai, F. R. Yu, and S. Bu, "Cloud Radio Access Networks (C-RAN) in Mobile Cloud Computing Systems," in *Processings of IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 369-374, 2014
4. A. U. R. Khan, M. Othman, S. A. Madani, and S. U. Khan, "A Survey of Mobile Cloud Computing Application Models," *IEEE Communications Surveys Tutorials*, Vol. 16, pp. 393-413, 2014
5. C. M. R. Institute, "C-RAN White Paper: The Road Towards Green RAN," (<http://labs.chinamobile.com/cran>)
6. X. Rao and V. Lau, "Model-based Evaluation: From Dependability to Security," *IEEE Transactions on Signal Processing*, Vol. 63, pp. 1056-1065, 2015

7. T. Q. S. Quek, M. Peng, O. Simeone, and W. Yu, "Cloud Radio Access Networks: Principles, Technologies, and Applications," *Cambridge University Press*, 2017
8. J. Padhye, V. Firoiu, D. F. Towsley, and J. F. Kurose, "Modeling TCP Reno Performance: A Simple Model and its Empirical Validation," *IEEE/ACM Transactions on Networking (ToN)*, Vol. 8, pp. 133-145, 2000
9. D. Gonzales, J. M. Kaplan, E. Saltzman, Z. Winkelman and D. Woods, "Cloud-trust-a Security Assessment Model for Infrastructure as a Service (IaaS) Clouds," *IEEE Transactions on Cloud Computing*, 2015
10. M. Assaad and D. Zeghlache, "Comparison Between MIMO Techniques in UMTS-HSDPA System," *IEEE Eighth International Symposium on Spread Spectrum Techniques and Applications*, pp. 874-878, 2004
11. A. Pathak, Y. A. Wang, C. Huang, A. Greenberg, and Y. C. Hu, "Measuring and Evaluating TCP Splitting for Cloud Services," *PAM*, pp. 41-50, 2010
12. M. Xie, Y. S. Dai, and K. L. Poh, "Computing System Reliability: Models and Analysis," Springer Science & Business Media, 2004
13. K. S. Trivedi, "Probability & Statistics with Reliability, Queuing and Computer Science Applications," John Wiley & Sons, 2008
14. Y. S. Dai, Y. Pan, X. Zou, "A Hierarchical Modeling and Analysis for Grid Service Reliability," *IEEE Transactions on Computers*, Vol. 56, pp. 681-691, 2007