Modeling and Performance Analysis of Scheduling System for Cloud Service Based on Stochastic Network Calculus

Yanbing Liu
School of Computer Science and Technology
Chongqing University of Posts and Telecommunications
Chongqing 400065, China
liuyb@cqupt.edu.cn

Le Lang
School of Computer Science and Technology
Chongqing University of Posts and Telecommunications
Chongqing 400065, China
l.angelia@163.com

Abstract—**In the background of cloud service is widely used, performance of the network has become more and more important, therefore the ability to guarantee the QoS of cloud networks is an important and essential part . To meet the demand for the QoS of cloud service network,we need an effective scheduling model to reach the user's requirements , and a tool to calculate the theoretical bounds of QoS parameters. As a tool for network analysis which is still evolving, stochastic network calculus is the theory to analysis the stochastic network service systems. We propose an efficient scheduling method of multiple priorities and differentiated services to reach users' expectations, and use stochastic network calculus to analyze the QoS parameters of cloud service network,it is proved that this scheduling method can meet the needs of different users of the network.**

Keywords-scheduling system;stochastic network calculus;performance analysis.

I. INTRODUCTION

In the widely used of cloud computing, numerous users request services from the cloud server, in the process of the service, the performance of the communication network plays an important role, even the key to guarantee the quality of users' experience of the service. Especially in the context of high-performance network applications, network performance has become the bottleneck of cloud service, QoS capabilities of the network of cloud service is an essential and important part .

To meet the demands for the QoS of cloud service network,we need an effective scheduling model to reach the user's different requirements , and a tool to calculate the theoretical bound of QoS parameters. Network calculus is a tool for QoS research created by the Chang and Cruz, and developed and looking forward to further improve by Agrawal and Le Boudec. Network Calculus is a recently developed tool for network performance analysis based on minimum plus algebra and maximum plus algebra. Network calculus provides a new way to model and analyze the service of data in a packet switched network.We can analyze the traffic backlog and transmission delay of the data flows in the network based on network calculus. However, as the deterministic network calculus analyze the worst-case network performance, which usually make low use of network resources[1], so we use stochastic network calculus to model and analyze .

Stochastic network calculus in which the probability calculation were introduced in to describe and analyze the statistical multiplexing characteristics between network data flows. Performance analysis based on stochastic network calculus can provide a certain QoS guarantees of data flows while also effectively improve the utilization of network resources, which can make up for the lack of deterministic network calculus theory effectively. Stochastic network calculus describe the service characteristics of the network data flows and nodes by probabilistic model, so the theoretical complexity is very high, and the application of the theory is also with a great difficulty.

In this paper, we build an effective scheduling model to meet the QoS requirements of different users in cloud service network, and use stochastic network calculus to analyze and model network services from an overall view of stream and time, and calculate the performance boundary of the system.

II. RELATED WORK

With the rapid development of cloud service, diversified requests to cloud service network has brought unprecedented challenges to network performance. When the cloud server provides services to users, not only needs high network bandwidth , and should satisfy the performance indicators like delay, the bandwidth and the queue length and others of different services. Scheduling strategy based on service differentiation can effectively protect the fairness and the QoS of cloud services, define criteria for service flows, so that all incoming traffic flows can meet the QoS requirements which is the necessary condition to guarantee the QoS .

Scheduling strategy of service differentiation is an important part to guarantee the QoS. How to schedule

978-1-4799-4419-4 /14/$31.00 ©2014 IEEE

service requests and allocate resources effectively is an important part of network QoS guarantees. For the past the realization of network access and scheduling is mainly by means of conventional theoretical tools like queuing theory, stochastic processes, although some of these theories can get some good results in specific applications, but also exhibit some drawbacks. For example, the object oriented of queuing theory analysis is a single packet or a data request, rather than the data stream in a period of time, which is lack of the study of overall performance of the system.

They established a non-preemptive with different priorities M/G/1 queuing model,and get approximate expected values of every service in this model by corresponding strategies and algorithms in paper [2].

A deterministic model was proposed in paper [3] to analyze the key indicators of cloud computing in hybrid network, which can get the delay boundaries, backlog of the model, and the stream reception area.

It studied the quality of service in differentiated services network with different scheduling and buffer management mechanism in paper [4] to ensure the maximum use of network bandwidth.

In the real-time network, network calculus theory was used to analyze the quality of service based on fixed priority and non-preemptive scheduling mechanism in paper [5]. They all provides a guideline of research to scheduling strategy in the papers [3,6,7,8] and [9] .

In the references [10,11] ,they presented a service model based on non-preemptive and multi-priorities , but the impact on the service curve of arrival time interval was not considered. And only two or three priorities of service was considered in the papers [12,13,14] .

In summary, most of the network model analysis is based on deterministic network calculus theory to schedule, or only introduced two or three priorities of service, so we will use stochastic network calculus in this study which is more universal to analyze multi-priorities preemptive scheduling model.We will analyze the multi-priorities service combine with non-preemptive and preemptive scheduling mechanism in the article.

III. SYSTEM MODEL AND DEFINITIONS

A. *Scheduling Model*

As the service provided by a cloud server is limited, in Figure 1 different users are divided into *n+1* priorities, the priority is defined as *1~n* with the priority level from high to low, while there is a preemptive priority as VIP user V, all the users will request to the cloud server for services. When the user with the priority of *i* is getting service from the server, there are several situations below: a. another user whose priority is lower than *i* requests for service, the cloud server will complete the current service for user *i*, then serve the lower priority user; b. another user whose priority is higher than *i* requests for service, the cloud server will serve the high-priority user after sending the current packet; c.when VIP user requests to the cloud server, as priority V is preemptive to the other users, the cloud server will immediately stop the current service and provide services to the user V.

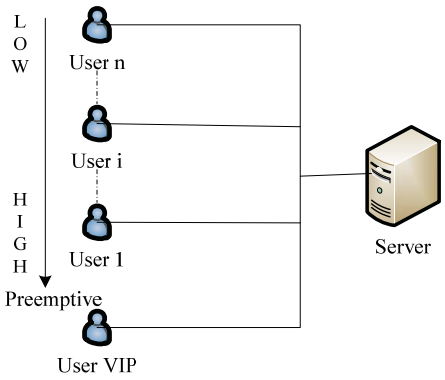

Figure 1. System Model

B. *Stochastic Network Calculus*

Stochastic network calculus in which the probability calculation is introduced to describe and analyze multiplexing characteristics between network data flow. Performance analysis based on stochastic network calculus can provide guarantees of QoS for data flow， while effectively improve the utilization of network resources to compensate for the lack of the deterministic stochastic network calculus theory.

In network calculus, the process is defined as a function of time t, the cumulative process is defined as $A(t)$, detachment process is defined as $A^*(t)$, and the service process as $S(t)$. And in the stochastic network calculus, data flow is described by statistical envelopes, it sets boundaries of arrival process which may violate the bound. That means data may not only change along with the time t, but also with other parameters, so we define that the arrival process $A(t)$ is up bounded by $(\sigma(\theta), \rho(\theta))$, and for all $s,t \geq 0$:

Definition 1:v.b.c Stochastic Arrival Curve: If for all $0 < s \leq t$,there exists

$$P\left\{\sup_{0\leq s\leq t} A(s,t)-\alpha(t-s)>x\right\}\leq f(x). \qquad (1)$$

We say that data flow A has v.b.c stochastic arrival curve α with boundary function f ,defined by

$$A\sim_{vb}\langle f,\alpha\rangle. \qquad (2)$$

Then the arrival curve is $\alpha(t)=\rho(\theta)\cdot t+\sigma(\theta)$,and it is bounded by $f(x)=e^{-\theta x}$ or $A\sim_{ta}\left\langle e^{-\theta x},\rho(\theta)+\sigma(\theta)\right\rangle$.

Definition 2:Weak Stochastic Service Curve:If for all $t,x\geq 0$, $P\{A\otimes\beta(t)-A^*(s)>x\}\leq g(x)$, then the weak

stochastic service curve of data flow A in the system is β bounded by g ,defined as

$$S \sim_{\text{ws}} \langle g, \beta \rangle . \tag{3}$$

Definition 3:Delay Bound: The delay bound at time t is defined as

$$D(t) = \inf\{d \geq 0 : A(t) \leq A^*(t+d)\} . \tag{4}$$

IV. PERFORMANCE ANALYSIS

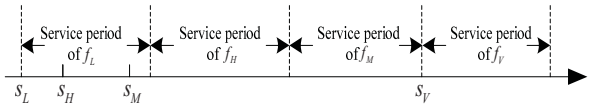

Figure 2. An example for analysis

We use Figure 2 as an example to analyze the service performance of users with different priorities in cloud service,when the orders of arrival changed, the analysis is also applicable.

We consider that a packet can only be serviced after its last bit have reached, and assume that all sequences are empty at the time 0, and the packet in a data stream is served according to FIFO (first-in-first-out), and there is no loss in this network.

In the figure s_L , s_M , s_H , s_V respectively represent as the arrival time of low priority user, the middle priority user, the high priority user and VIP user. When the low priority users was receiving service from the cloud server, the high and middle priority users arrived sequentially, they would enter the waiting queue at the completion of the current service packet, the cloud server started to serve the higher priority users, then the VIP user arrived during this time, so the server terminated the user's service of middle priority immediately, and began to serve the VIP user.

Let firstly analyze the data flow of high priority f_L ,the total output of the system is $\rho(\theta)\cdot(t-s_L)$, therefor we have

$$\begin{aligned} R^*_L(t) - R^*_L(s_L) = \rho(\theta)\cdot(t-s_L) - [R^*_H(t) - R^*_H(s_L)] - [\\ R^*_M(t) - R^*_M(s_L)] - [R^*_V(t) - R^*_V(s_L)] \end{aligned} \tag{5}$$

As at the specific time s_L , the output flow equal to the input flow, but for the mutative time t, $R^*(t) \leq R(t)$.

$$\begin{aligned} 0 \leq R^*_H(t) - R^*_H(s_L) &= R^*_H(t) - R_H(s_L) \\ &\leq R_H(t) - R_H(s_L) \leq \rho_H(\theta)\cdot(t-s_L) \end{aligned} \tag{6}$$

Similarly,

$$\begin{aligned} 0 \leq R^*_M(t) - R^*_M \leq \alpha_M(t-s_L) \\ 0 \leq R^*_V(t) - R^*_V \leq \alpha_V(t-s_L) \end{aligned} \tag{7}$$

Through the formula above, we have

$$R^*_L(t) - R_L(s_L) = R^*_L(t) - R^*_L(s_L) \geq K_L \cdot (t-s_L) \tag{8}$$

$$\begin{aligned} R^*_L(t) &\geq R_L(s_L) + K_L \cdot (t-s_L) \\ &= (R_L \otimes K_L)(t) \end{aligned} \tag{9}$$

In which

$$\begin{aligned} K_L(t-s_L) = [\rho(\theta)\cdot(t-s_L) - \alpha_H(t-s_L) \\ -\alpha_M(t-s_L) - \alpha_V(t-s_L)]^+ \end{aligned} . \tag{10}$$

According to the uncertainty of signals described in stochastic network calculus we can get

$$P\{R_L \otimes K_L(t) - R^*_L(t) > 0\} \leq g(0) . \tag{11}$$

So the weak stochastic service curve of low priority user flow f_L is $S \sim_{\text{ws}} \langle 0, K_L \rangle$.

Assuming the arrival curves of user flow f_M , f_H and f_{v} are $\alpha_M = r_{\rho_M,\sigma_M}$, $\alpha_H = r_{\rho_H,\sigma_H}$, $\alpha_V = r_{\rho_V,\sigma_V}$ r espectively, we can get the service rate of f_L is $R_L = \rho(\theta) - \rho_H(\theta) - \rho_M(\theta) - \rho_V(\theta)$,and the latency

$$T_L = \frac{\sigma_M(\theta) + \sigma_H(\theta) + \sigma_V(\theta)}{\rho(\theta) - \rho_M(\theta) - \rho_H(\theta) - \rho_V(\theta)} . \tag{12}$$

Similarly we can get the weak stochastic service curve of f_M as $S \sim_{\text{ws}} \langle 0, K_M \rangle$,in which

$K_M(t-s_M) = [\rho(\theta)\cdot(t-s_M) - L^L_{\max} - \alpha_H(t-s_M) - \alpha_V(t-s_M)]^+$.

The service rate of f_M is $R_M = \rho(\theta) - \rho_H(\theta) - \rho_V(\theta)$, and the latency is $T_M = \frac{L^L_{\max} + \sigma_H(\theta) + \sigma_V(\theta)}{\rho(\theta) - \rho_H(\theta) - \rho_V(\theta)}$.

The weak stochastic service curve of high priority user f_H is $S \sim_{\text{ws}} \langle 0, K_H \rangle$,in which

$$K_H(t-s_H) = [\rho(\theta)\cdot(t-s_H) - L^L_{\max}]^+ . \tag{13}$$

The service rate of f_H is $R_H = \rho(\theta)$,and the latency $T_H = \frac{L^L_{\max}}{\rho(\theta)}$.

For preemptive priority user VIP,The weak stochastic service curve of f_V is $S \sim_{ws} \langle 0, K_V \rangle$,

$$K_V(t-s_V) = [\rho(\theta)\cdot(t-s_V)]^+ . \tag{14}$$

The service rate of f_V is $R_H = \rho(\theta)$,and the latency $T_H = 0$.

From the analysis above, we can get the weak stochastic service curve of f_i with many priorities in one cloud server:

$$\begin{aligned}\beta_i(t) &= K_i(t-s_i) \\ &= [\rho(\theta) - \sum_{j<i}\rho_j(\theta) - \rho_V(\theta)][\\ &\quad t - \frac{\sum_{j<i}\sigma_j(\theta) + L'_{\max} + \sigma_V(\theta)}{\rho(\theta) - \sum_{j<i}\rho_j(\theta) - \rho_V(\theta)}]^+\end{aligned} \tag{15}$$

And the service rate of them

$$R_i = \rho(\theta) - \sum_{j>i}\rho_j(\theta) - \rho_V(\theta) . \tag{16}$$

So we can get the delay bound of cloud user i by definition 3:

$$\begin{aligned}D_i(t) &= \sup_{t\ge 0}\{\inf\{d \ge 0 : \alpha_i(t) \le \beta_i(t+d)\}\} \\ &= \inf\{d \ge 0 : \sigma_i(\theta) \le [\rho(\theta) - \sum_{j<i}\rho_j(\theta) - \rho_V(\theta)]\cdot \\ &\quad d - \sum_{j<i}\sigma_j(\theta) - L'_{\max} - \sigma_V(\theta)\} \\ &= \frac{\sigma_i(\theta)}{\rho(\theta) - \sum_{j<i}\rho_j(\theta) - \rho_V(\theta)} + \\ &\quad \frac{\sum_{j<i}\sigma_j(\theta) + L'_{\max} + \sigma_V(\theta)}{\rho(\theta) - \sum_{j<i}\rho_j(\theta) - \rho_V(\theta)}\end{aligned} \tag{16}$$

And the delay bound of user VIP is $D_V(t) = \frac{\sigma_V(\theta)}{\rho(\theta)}$.

Figure 3 shows that when we set the arrival rate of the users with different priorities the same value, we can see the VIP user's delay bound is floating with small margin and maintaining a low value which means the cloud server provide it with high QoS. And the delay bound of high priority user is always less than the delay bound of lower priority users.As the service rate from cloud server decreasing, the change is more and more various, but when the service rate reached high values as 5Mpbs, the delay bound of users will not change that much.

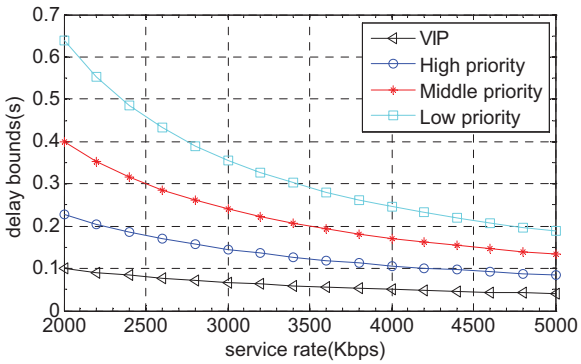

Figure 3. Delay bounds of different users

We also can get the backlog of user whose priority is *i* is:

$$B_i(t) = \sigma_i(\theta) + \rho_i(\theta)\cdot[\frac{\sum_{j<i}\sigma_i(\theta) + L'_{\max} + \sigma_V(t)}{\rho(\theta) - \sum_{j<i}\rho_i(\theta) - \rho_V(t)}] . \tag{17}$$

As the VIP user is preemptive to normal users,so its backlog only relate to its own,we can get $B_V(t) = \sigma_V(\theta)$.

V. CONCLUSION

This paper presents a scheduling model with preemptive and non-preemptive for a multi-priorities services to meet the QoS of different users, which can not only provide normal users service with different priorities, but also provide special VIP users with preemptive service. We also apply the theory of stochastic network calculus to analyze the network performance curves and delay bounds, and get the service rate, latency function and delay, which proved that the scheduling model can effectively provide differentiated services to meet the needs of different users.

We only use the weak stochastic arrival curve of stochastic network calculus as the basis for further analysis, in the future we will consider some more complex network, and apply the model to some specific network environment.

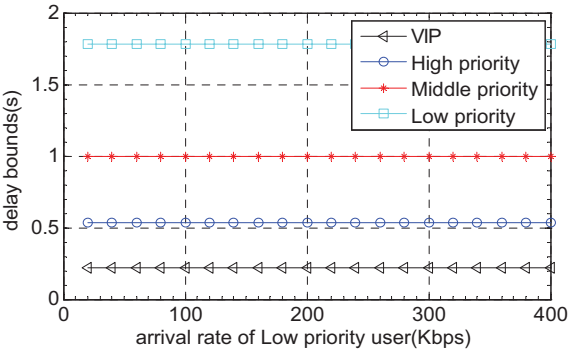

Figure 4. The influence of low priority user's arrival rate

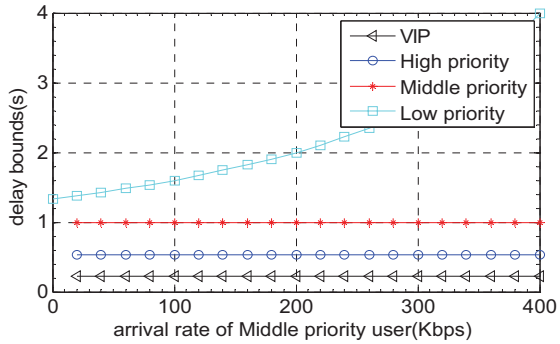

Figure 5. The influence of middle priority user's arrival rate

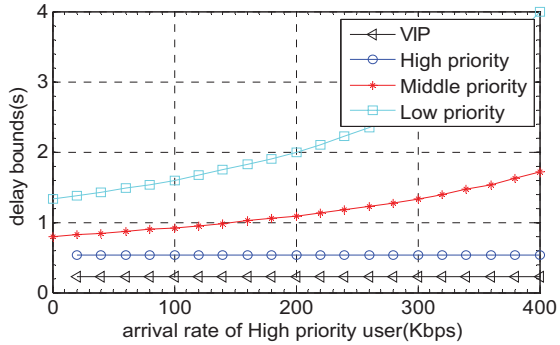

Figure 6. The influence of high priority user's arrival rate

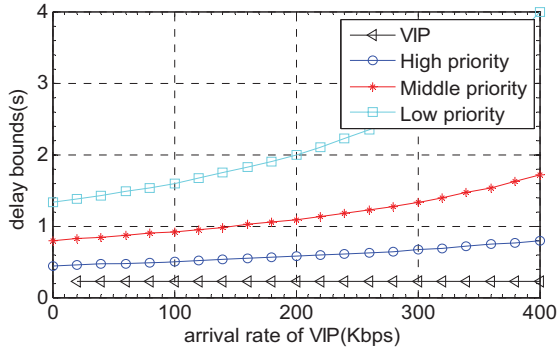

Figure 7. The influence of VIP user's arrival rate

ACKNOWLEDGMENT

Here and now, I would like to extend my sincere thanks to all those who have helped me make this thesis possible and better.This research would not have been possible without the support and substantive contributions of professor Liu Yanbing, he also played an important role in indicating a bright road in my future writing.

REFERENCES

[1] Jiang Y. Stochastic network calculus for performance analysis of Internet networks—An overview and outlook[C]//Computing, Networking and Communications (ICNC), 2012 International Conference on. IEEE, 2012: 638-644.

[2] Li L. An optimistic differentiated service job scheduling system for cloud computing service users and providers[C]//Multimedia and Ubiquitous Engineering, 2009. MUE'09. Third International Conference on. IEEE, 2009: 295-299.

[3] Duan Q. Modeling and performance analysis on network virtualization for composite network-cloud service provisioning[C]//Services (SERVICES), 2011 IEEE World Congress on. IEEE, 2011: 548-555.

[4] Musa I K, Nejabati R, Simeonidou D. Performance analysis of hybrid network for cloud datacenter[C]//Computer Science and Electronic Engineering Conference (CEEC), 2012 4th. IEEE, 2012: 154-159.

[5] Wu Z, Lv T, Wang X, et al. The buffer size assignment of AFDX based on network calculus[C]//Reliability, Maintainability and Safety (ICRMS), 2011 9th International Conference on. IEEE, 2011: 1319-1323.

[6] S. Fan, Y. Zhao, and H. Sun, Upper Delay Bound Research about LR-PANs Network Based on Network Calculus[C], In Proc. WiCOM 2010, 9: 23-25.

[7] Duan Q. Modeling and performance analysis on network virtualization for composite network-cloud service provisioning[C]//Services (SERVICES), 2011 IEEE World Congress on. IEEE, 2011: 548-555.

[8] Musa I K, Nejabati R, Simeonidou D. Performance analysis of hybrid network for cloud datacenter[C]//Computer Science and Electronic Engineering Conference (CEEC), 2012 4th. IEEE, 2012: 154-159.

[9] Boudec J L, Thiran P. Network Calculus:A theory of Deterministic Queuing System for the Internet[M]. Berlin:Springer-Verlag, 2002.

[10] Qian Y, Lu Z, Dou Q. Qos scheduling for nocs: Strict priority queueing versus weighted round robin[C]//Computer Design (ICCD), 2010 IEEE International Conference on. IEEE, 2010: 52-59.

[11] Sofack W M, Boyer M. Non preemptive static priority with network calculus[C]//Emerging Technologies & Factory Automation (ETFA), 2011 IEEE 16th Conference on. IEEE, 2011: 1-8.

[12] S. Fan, Y. Zhao, and H. Sun, Upper Delay Bound Research about LR-PANs Network Based on Network Calculus[C], In Proc. WiCOM 2010, 9: 23-25.

[13] Masrur A, Chakraborty S, Färber G. Constant-time admission control for deadline monotonic tasks[C]//Proceedings of the Conference on Design, Automation and Test in Europe. European Design and Automation Association, 2010: 220-225.

[14] Yang K, Jia X, Ren K. DAC-MACS: Effective Data Access Control for Multi-Authority Cloud Storage Systems[J]. IACR Cryptology ePrint Archive, 2012, 2012: 419.