Modeling and Analysis on Network Performance for Cloud Service Delivery with Multiple Paths

Jun Huang, Juan Lv
School of Communication and Information Engineering
Chongqing University of Posts and Telecommunications
Chongqing, China 400065
{xiaoniuadmin, lvjuanjkl}@gmail.com

Qiang Duan
Information Sciences & Technology Department
The Pennsylvania State University Abington, PA 19001
qduan@psu.edu

ABSTRACT
Networking plays a crucial role in Cloud service delivery and network Quality of Service (QoS) has a significant impact on Cloud service performance. Cloud services for data-intensive computing applications typically require high throughput data transmission, which in some cases may not be achievable with a single end-to-end path in the network. Therefore multi-path data transmission in wide area networks for supporting Cloud service delivery is necessary. Performance evaluation on multi-path communications for Cloud service delivery is still an opening challenging issue. In this paper, we propose a novel model of networking systems for Cloud service delivery with multiple paths and develop techniques for analyzing delay and backlog performance of such systems. We employ network calculus in this paper to address some new challenges brought in by multi-path networking for Cloud service delivery. Applying network calculus makes our modeling and analysis applicable to inter-domain heterogeneous networking systems for supporting a wide variety of Cloud services. Numerical results are also given in the paper to illustrate applications of the developed modeling and analysis techniques.

Categories and Subject Descriptors
C.2.3 [Computer–Communication Networks]: Network Operations; C.4 [Performance of Systems]: Modeling Techniques

Keywords
networking, Cloud service delivery, multiple paths, delay, backlog

1. INTRODUCTION
Cloud computing is a large scale distributed computing paradigm in which a pool of virtualized computing resources are delivered on demand as services to users over the Internet. Networking plays a crucial role in Cloud computing. Recent study has shown that network performance has a significant impact Cloud service quality, and in many cases data communications become a bottleneck that limits high-performance applications in Clouds [6, 12]. Thorough understanding of network QoS in a Cloud environment forms a foundation for obtaining deep insight about Cloud service performance. Therefore, analysis on network performance for Cloud service delivery becomes an important research problem.

Data intensive Cloud services typically require networking for high throughput data transmission. If service delivery relies only on a single network path and there is no path between the service provider and user that meets the throughput requirement, then this Cloud service may not be supported. Therefore, a networking system that utilizes multiple paths for data transmission would be desirable to Cloud service provisioning. A variety of existing networking technologies support multi-path data transmission. One of the most well-known multi-path routing techniques is Equal-Cost Multi-Path (ECMP) routing that is currently supported by OSPF and RIP. In Multi-Protocol Label Switching (MPLS) networks, multiple paths between the source and destination gateways can be setup by using a signaling protocol, e.g., Constraint-Based Routing Label Distribution Protocol (CR-LDP) or Resource Reservation Protocol-Traffic Engineering (RSVP-TE). Some emerging networking technologies, for example Network Virtualization (NV) and Software-Defined Networking (SDN), also support establishing multiple paths for Cloud service delivery. In addition, mobile Cloud computing is becoming an important extension of the regular Cloud paradigm, in which wireless mobile devices are used for accessing services hosted in Cloud infrastructure. Due to advance of wireless communications, a single mobile device may simultaneously use several different types of wireless access networks; thus enabling multiple paths of data transmission for Cloud service delivery. Therefore, networking systems with multipath data transmission are expected to be widely applied in Cloud environments and particularly important for service delivery in mobile Cloud computing.

Comparing to data center networking inside Cloud infrastructure, networking for Cloud service delivery to end users, which is typically achieved through data transmission across wide area networks, faces some special challenges. Data communications for Cloud service delivery to an end user
is operated by an Internet Service Provider (ISP), which is independent to Cloud service providers. Also, such a data communication session may need to traverse multiple network domains that are managed by different ISPs, which leads to the complex inter-domain QoS problem. In addition, the resource abstraction caused by NV and separation of data and control planes enabled by SDN decouple network service provisioning from specific data forwarding mechanisms, which further complicates the evaluation of network performance for Cloud service delivery.

Although some previous works on Cloud networking performance have attempted to address the above challenges, currently available results only consider single-path data transmission for Cloud service delivery [3, 4]. The performance of multi-path networking for Cloud service delivery is still basically an unexplored problem, which is the focus of our research presented in this paper. Our main contributions made in this paper include a novel model for networking systems with multiple paths for Cloud service delivery and the analysis techniques for evaluating delay and backlog performance of such systems. Application of network calculus in this paper enables the proposed modeling and analysis techniques to address some challenging issues in Cloud networking, including inter-domain heterogeneous networking, network resource virtualization, and data-control planes decoupling. The developed techniques also allow us to compare performance of multi-path and single-path networking for Cloud service delivery; thus offering guidelines for network operation and management to improve Cloud service performance.

The rest of the paper is organized as follows. We propose a system model for multi-path networking systems for Cloud service delivery in Section 2 and then develop analysis techniques for evaluating delay and backlog performance of such systems in Section 3. We summarize the related works in Section 5. Numerical experiment results are reported in Section 4. We draw conclusions in Section 6.

2. SYSTEM MODEL

Evaluating performance of multi-path networking for Cloud service delivery requires a system model that can address some special challenges. First, each of the parallel paths in such a networking system may traverse multiple network domains that are managed by different ISPs and implemented with various technologies; therefore, the model should be flexible and applicable to inter-domain networking across heterogeneous domains. Second, emerging technologies such as NV and SDN for Cloud networking enables network resource virtualization and data-control plane separation, which allows network service provisioning decoupled from any specific data forwarding mechanisms; therefore, performance evaluation requires a general model that is agnostic to the underlying implementations of service delivery.

In order to develop a general and flexible model for multi-path networking systems across heterogeneous domains for Cloud service delivery, we employ the concept of service curve from network calculus theory [2] to characterize the service capability of each component in the analyzed networking system. Although network domains may have diverse implementations for data transmission of the Cloud service delivery, each domain is expected to provide a certain level of QoS that is specified in the form of Service Level Agreement (SLA) between the domain and the service user.

![Model of single-path networking for Cloud service delivery.

The service curve concept in network calculus offers a flexible method to characterize such QoS guarantee provided by each domain and such a method is independent with domain implementations. Network calculus also provides the technique to obtain the service curve of an end-to-end system based on the service curves of individual domains in the system; thus offering a promising solution to analyzing performance of inter-domain networking.

The service curve of a network node is defined as follows. Let \( R(t) \) and \( E(t) \) respectively be the accumulated amount of traffic of a flow that arrives at and departs from the node by time \( t \). We say that the node offers to the flow a service curve \( \beta(t) \) if and only if \( \beta(t) \) is wide sense increasing, \( \beta(0) = 0 \), and for all \( t \geq 0 \),

\[
E(t) \geq R(t) \otimes \beta(t) \tag{1}
\]

where \( \otimes \) denotes the operation defined as \( x(t) \otimes y(t) = \inf_{s \leq t \leq s'} \{ x(t-s) + y(s) \} \).

We employ the arrival curve concept from network calculus to specify the traffic load of data transmission for Cloud service delivery. Let \( R(t) \) be the amount of traffic in a flow that arrives at a node by time \( t \), if for any time instants \( s \leq t \),

\[
R(t) - R(s) \leq \alpha(t-s), \tag{2}
\]

where \( \alpha(t) \) is a wide sense increasing function for all \( t \geq 0 \); then \( \alpha(t) \) is called an arrival curve of the flow.

A service curve-based model for a single-path networking system is shown in Figure 1. In this system the path passes \( k \) nodes, \( n_1, n_2, \ldots, n_k \), and each node offers a service curve \( \beta_i, i = 1, 2, \ldots, k \), to the traffic flow for Cloud service delivery. Each node in this model is a general service component that represents the network service offered by a network domain for delivering a Cloud service.

Figure 2 shows a multi-path networking system for Cloud service delivery. Suppose there are \( m \) paths, denoted as \( p_1, p_2, \ldots, p_m \), available for Cloud data transmission. The gateway node \( n_0 \) merges the traffic flow \( f \) for a Cloud service to \( m \) sub-flows, \( f_1, f_2, \ldots, f_m \), one on each path. The aggregate node \( n_m \) merges the traffic from all \( m \) paths before delivering them to the destination. Suppose the arrival curve for flow \( f \) to be \( R(f) \) and that the arrival curve of the subflow \( f_i \) be \( R_i(t) \), then \( R(f) = \sum_{i=1}^{m} R_i(t) \). Suppose there are \( c_i \) nodes on path \( p_i \) and the \( j \)-th node on this path is denoted as \( n_{ij} \). The node \( n_{ij} \) offers a service curve \( \beta_{ij} \) to the subflow \( f_i \). The gateway node offers a service curve \( \beta_f(t) \) to the flow \( f \). The service capacity offered by the aggregate node to the total traffic is modeled by a service curve \( \beta_\lambda(t) \).

3. PERFORMANCE ANALYSIS

In this section we develop techniques to evaluate the maximum delay of end-to-end networking for Cloud service delivery, which has a significant impact on Cloud service performance. We will also study the backlog upper bound in
the networking system, which gives the required buffer space in the networking system to avoid packet loss. We assume that each node in the service delivery networking system is an Latency-Rate (LR) server [11], which is a general server model defined as follows.

If a node guarantees a traffic flow a service curve

$$\beta_{R,T} = R[t - T]^{+} = R(t - T), \text{for } t > T$$  \hspace{1cm} (3)

then the node is called a LR server with R as the service rate and T as the latency parameter of the server. For a broad range of LR servers the latency parameter of the server can be determined as

$$T = L/R + L/C,$$  \hspace{1cm} (4)

where L the maximum packet length for the flow, R is the service rate the server offers to the flow, and C is the total capacity of the output link of the server.

We assume that the arrival traffic entering the service delivery system is constrained by a leaky bucket regulator. Then the arrival curve for the gateway node is $\alpha(t) = M + \rho t$, where M is the maximum burst size and $\rho$ is the sustain rate of the traffic flow.

### 3.1 Delay and Backlog Performance of Single Path Service Delivery

As shown in Figure 1, there are k nodes on the end-to-end Cloud service delivery path and each node $n_i$, $i = 1, 2, \cdots , k$, offers a service curve $\beta_i = R_i[t - T_i]^{+}$. Then the service curve guaranteed by the end-to-end path for Cloud service delivery can be determined as

$$\beta_{\text{net}} = \beta_1 \otimes \beta_2 \otimes \cdots \beta_k = R_{\text{net}}[t - T_{\text{net}}]^{+} \hspace{1cm} (5)$$

where $R_{\text{net}} = \min\{R_1, R_2, \cdots , R_k\}$ and $T_{\text{net}} = \sum_{i=1}^{k} T_i$.

Equation (5) implies that if each network node on the single path of Cloud service delivery is a LR server, then the service capability of the end-to-end path can be modeled by a LR server. The latency parameter of the end-to-end LR server is equal to the sum of the latency parameters of all individual nodes on this path. The end-to-end service rate will be the minimum rate of all individual nodes on the path.

Network calculus theory shows that if a service offers a service curve $\beta(t)$ to a traffic flow constrained by an arrival curve $\alpha(t)$, then the maximum delay for the flow at this server, $D_{\text{max}}$, can be determined as

$$D_{\text{max}} = \sup_{0 \leq s \leq t} \{\inf \{d \geq 0 : \alpha(s) \leq \beta(s + d)\}\}. \hspace{1cm} (6)$$

The maximum backlog of the flow at this server will be

$$B_{\text{max}} = \sup_{0 \leq s \leq t} \{\alpha(s) - \beta(s)\}. \hspace{1cm} (7)$$

Therefore, for a Cloud service traffic flow with an arrival curve $\alpha(t) = M + \rho t$, we can obtain that the maximum delay performance of the single path networking system is

$$D_{\text{max}} = T_{\text{net}} + M/R_{\text{net}} \hspace{1cm} (8)$$

and the maximum backlog at the entry node of the end-to-end path is

$$B_{\text{max}} = M + \rho T_{\text{net}}. \hspace{1cm} (9)$$

Following (4), the latency parameter of the end-to-end server model is $T_{\text{net}} = L/R_{\text{net}} + L/C$. Then

$$D_{\text{max}} = L(1/R_{\text{net}} + 1/C) + M/R_{\text{net}} \hspace{1cm} (10)$$

and

$$B_{\text{max}} = M + \rho L(1/R_{\text{net}} + 1/C). \hspace{1cm} (11)$$

#### 3.2 Delay Performance for Multi-Path Service Delivery

The model of multi-path networking for Cloud service delivery is shown in Figure 2. Suppose the gateway node and the aggregate node offer a service curve $\beta_{ij} = R_i[t - T_i]^{+}$ and $\beta_a = R_a[t - T_a]^{+}$ respectively. Assume that there are $m$ paths, $p_i$, $i = 1, 2, \cdots , m$, from the gateway node to the aggregate node, and there are $c_i$ nodes on path $p_i$. If the service curve offered by the $j$-th node on the path $p_i$ is $\beta_{ij} = R_{i,j}[t - T_{i,j}]^{+}$, then the service curve guaranteed by the entire path $p_i$ will be

$$\beta_{p_i} = \beta_{i,1} \otimes \beta_{i,2} \otimes \cdots \otimes \beta_{i,c_i} = R_{i}[t - T_{i}]^{+} \hspace{1cm} (12)$$

where $R_{i} = \min\{R_{i,1}, R_{i,2}, \cdots , R_{i,c_i}\}$ and $T_{i} = \sum_{j=1}^{c_i} T_{i,j}$.

The end-to-end delay $D_{e}$ of data transmission across the multi-path networking systems for Cloud service delivery consists of three parts, the delay at the gateway node $D_g$, the delay of multi-path transmission $D_p$, and the delay at the aggregation node $D_a$; that is, $D_e = D_g + D_p + D_a$.

At the gateway node $n_g$, given the arrival curve $\alpha(t) = M + \rho t$ of the traffic flow and the service curve $\beta_a(t) = R_a[t - T_a]^{+}$, then following (6) we can get the maximum delay at this node

$$D_g = \sup_{s \geq 0} \{\inf \{d \geq 0 : \alpha(s) \leq \beta_g(s + d)\}\} \hspace{1cm} (13)$$

$= T_g + M/R_g$

The departure process from the aggregate node forms the arrival process for multi-path data transmission. Based on Theorem 1.4.2 in [2], the departure traffic from node $n_a$ is upper bounded by $M + \rho T_g + \rho t$, which is actually the arrival curve for the multi-path data transmission system. The total arrival traffic splits into $m$ subflows, $f_i$, $i = 1, 2, \cdots , m$, one
for each paths. Suppose $f_i$ has an arrival curve $\alpha_i(t) = b_i + r_i t$, then

$$\sum_{i=1}^{m} \alpha_i(t) = M + \rho T_g + \rho t \quad (14)$$

$$\sum_{i=1}^{m} b_i = M + \rho T_g \quad (15)$$

$$\sum_{i=1}^{m} r_i = \rho \quad (16)$$

For each path $p_i$, given the arrival curve $\alpha_i(t)$ of the sub-flow on this path and the end-to-end service curve $\beta_{net}^i$ guaranteed by the path to this subflow, the maximum delay on this path is

$$D_e^i = \sup_{s \geq 0} \{ \inf \{ d \geq 0 : \alpha_i(s) \leq \beta_{net}^i(s + d) \} \} \quad (17)$$

Therefore, the delay upper bound for multi-path data transmission can be determined as

$$D_e = \max_{1 \leq i \leq m} \{ D_e^i \} = \max_{1 \leq i \leq m} \{ T_{net}^i + b_i / R_{net}^i \} \quad (18)$$

The node $n_a$ aggregates traffic from all $m$ paths before delivering them to the final destination for Cloud service delivery. The arrival traffic from path $p_i$ at $n_a$ is constrained by $\alpha_i^a(t) = b_i + r_i T_{net}^i + r_i t$. Then the arrival curve for the total traffic at $n_a$ is

$$\alpha_a(t) = \sum_{i=1}^{m} \alpha_i^a(t) = \sum_{i=1}^{m} (b_i + r_i T_{net}^i) + \sum_{i=1}^{m} r_it \quad (19)$$

The service curve offered by the aggregate node is $\beta_a(t) = R_a [T_a - T_a^+]$, therefore the maximum delay at this node is

$$D_a = \sup_{s \geq 0} \{ \inf \{ d \geq 0 : \alpha(s) \leq \beta_a(s + d) \} \} \quad (20)$$

Therefore the maximum end-to-end delay of multi-path networking for Cloud service delivery is

$$D_{max} = D_e + D_a \quad (21)$$

$$= T_a + \frac{M}{R_a} + T_a + \max_{1 \leq i \leq m} \{ T_{net}^i + b_i / R_{net}^i \}$$

$$+ \sum_{i=1}^{m} (b_i + r_i T_{net}^i) / R_a \quad (22)$$

### 3.3 Backlog Performance for Multi-Path Service Delivery

Due to the burst arrival traffic at each node on the path for Cloud service delivery, buffer space is important resource that must be allocated to avoid packet loss. In this subsection we analyze the upper bound of traffic backlog at different nodes of the multi-path networking system, which provides a guideline for buffer space allocation. We are particularly interested in the buffer space requirements at the gateway node and the aggregation node.

For the gateway node $n_g$, given the arrival traffic constrained by $\alpha(t) = M + \rho t$ and the service curve offered to Cloud service $\beta_g = R_g [T_a - T_a^+]$, following (7) we get that the maximum backlog at this node is

$$B_g = \sup_{s \geq 0} (\alpha(s) - \beta_g(s)) = M + \rho T_g \quad (23)$$

At the aggregation node $n_a$, the total arrival traffic to this node is the aggregation of all sub-flows $f_i, i = 1, 2, \cdots, m$. As given in (17), the arrival curve for this node is

$$\alpha_a(t) = \sum_{i=1}^{m} (b_i + r_i T_{net}^i) + \sum_{i=1}^{m} r_i t = M_a + \rho a t. \quad (24)$$

Given the service curve guaranteed by this node, $\beta_a = R_a [T_a - T_a^+]$, the upper bound of traffic backlog at the node can be determined as

$$B_a = \sup_{s \geq 0} \{ \alpha_a(s) - \beta_a(s) \} \quad (25)$$

$$= M_a + \rho_a T_a = \sum_{i=1}^{m} (b_i + r_i T_{net}^i) + T_\rho \sum_{i=1}^{m} r_i$$

### 4. Numerical Experiments

In this section we give numerical experiment results to illustrate applications of the developed analysis techniques and discuss the insights we obtained about delay and backlog performance of multi-path networking for Cloud service delivery. We consider three Cloud services A, B, and C, whose traffic loads are constrained by leaky-bucket traffic regulators. The leaky-bucket parameters are respectively $M_A = 200$ kbits and $\rho_A = 120$ Mbps for service A, $M_B = 120$ kbits and $\rho_B = 120$ Mbps for service B, and $M_C = 200$ kbits and $\rho_C = 120$ Mbps for service C. The maximum packet length in the network is assumed to be 1500 bytes. For the single-path transmission case we assume 1 Gbps total link capacity of the path. For multi-path case, we consider three parallel paths between the gateway and aggregate nodes, which have link capacities of 77 Mbps, 124 Mbps, and 99 Mbps.

We first analyzed the maximum delay and backlog performance in the single-path service delivery case. Figure 3 shows the relation between end-to-end delay bounds for the three Cloud services and the available network service rate on the path. Figure 4 gives the backlog upper bounds of these services with different network service rates on the path. These two figures show that for a given Cloud service, both delay and backlog upper bounds drop with increasing network service rate. Comparison between Figures 3 and 4 indicates that for all three Cloud services increase in network service rate is more effective on improving delay performance than on reducing backlog upper bounds. From Figure 3 we can see that the delay curves of services A and B overlap and is lower than the delay curve of service C. Services A and B have different sustain rates but the same burst size while services A and C have the same sustain rate but different burst sizes. Therefore, Figure 3 implies that given network service rate the delay upper bound of a Cloud service is determined by the burst size of the service traffic but is not impacted by its sustain rate. Similarly Figure 4 shows that for a given network service rate the backlog upper bound of a Cloud service is mainly determined by traffic burst size.

When study performance of multi-path networking for Cloud service delivery, we considered two policies for traffic distribution: equal distribution and weighted distribution. In equal distribution the gateway node distributes its delivery rate equally to the $m$ parallel paths. Therefore, the traffic parameters $b_i$ and $r_i$ in (14), which are the sustain rate and burst size of the traffic flow on path $p_i$, will be

$$b_i = (M + \rho T_g) / m \quad (26)$$

and $r_i = \rho / m, \quad i = 1, 2, \cdots, m. \quad (27)$$
In weighted distribution traffic is distributed by the gateway node to each path according to a weight that is equal to the ratio of the available service rate on the path to the total service rate of all paths. That is, the weight for path $p_i$ is

$$w_i = \frac{R_{inet}^i}{\sum_{i=1}^{m} R_{inet}^i}.$$  

Then the traffic parameters of the sub-flow on path $p_i$ is

$$b_i = w_i(M + \beta T_g) \quad \text{and} \quad r_i = w_i \rho_i, \quad i = 1, 2, \cdots, m. \quad (23)$$

The equal and weighted traffic distribution policies can be realized by various technologies. For example packet-by-packet Round-Robin scheduling can be used to distribute load equally to all parallel paths. Weighted distribution can be implemented with Weighted Round-Robin or Weighted Fair Routing scheduling schemes [9].

The maximum end-to-end delay upper bounds of multi-path networking for Cloud service delivery with various network service rates are plotted in Figure 5. Figure 5 (a) and (b) show delay performance of service A with different service rates at the gateway node and the aggregate node respectively. We can see that for both equal distribution and weighted distribution, delay upper bound drops with increase in the service rate at both gateway and aggregate nodes. These two figures also show that given the service rate at either gateway or aggregate node, weighted traffic distribution always achieves a lower delay bound than what equal distribution does. This is because weighted distribution makes traffic load of each sub-flow match the available bandwidth of its path so that avoid overloading any path to cause long delay.

Figure 5 (c) and (d) give the delay upper bounds of all three services with different service rates at the gateway node and aggregate node respectively. Traffic distributions of all services follow the weighted policy. In addition to verify the impact of increasing service rate on improving delay performance of all services, these figures also show the influence of service traffic parameters on delay performance. We notice that in each figure the delay curves of services A and B are very close and are much lower than the curve of service C. This implies that traffic burst size of a service is the decisive factor of the maximum delay performance of the service while its sustain rate only has very limited contribution that is almost ignorable.

We also analyzed the maximum traffic backlog of services A, B, and C at both gateway and aggregate nodes. Figure 6 (a) shows variation of the backlog upper bounds of all services at gateway and aggregate nodes with increasing gateway service capacity. Figure 6 (a) gives the backlog of all services at the aggregate node with different service rates. These figures indicate that increasing service rate at the gateway node can reduce backlog upper bounds of all services at both gateway and aggregate nodes. Increase in aggregate node service capacity, although decreasing the backlog at this node, has no influence on the backlog at the gateway node. Comparing the backlog curves of different services given in Figure 6 (a) and (b) tells us that the traffic burst size of a service has more significant impact on its backlog, at both gateway and aggregate nodes, than what its sustain rate has. Another observation we got from Figure 6 (a) is that for each service, its backlog at the aggregate node is always larger than that at the gateway node. This is because transmitting a traffic flow through multiple nodes increases its burst size, thus generating a larger backlog upper bound at the aggregate node. The results plotted in Figure 6 (a) and (b) are all obtained with weighted traffic distribution. To examine influence of traffic distribution policies on backlog performance, we show in Figure 6 (c) the backlog bounds of service A with equal and weighted distribution. This figure shows that weighted distribution causes less amount of backlog than what equal distribution does.

We also compared the maximum delay and backlog performance achieved by single path and multi-path networking systems. We first considered a case in which a given total amount of service capacity (300 Mbps in our experiments) is split equally among multiple paths. That is, we compare the performance achieved by a single path with 300 Mbps bandwidth, by two parallel paths with 150 Mbps each, by three parallel paths with 100 Mbps each, etc. The obtained results are given in Figure 7, where Figure 7 (a) gives the end-to-end delay upper bounds of services A, B, and C when delivered through different numbers of paths, while Figure 7 (b) shows the maximum backlog of these services for different numbers of paths. From both figures we can see that delay and backlog upper bounds of all services increase with the number of paths. The results indicate that using multiple paths to achieve the same amount of bandwidth that can be provided by a single path will increase service delay and cause larger traffic backlog (thus consuming more buffer space). This implies that if there exist a single path in the network that meets the bandwidth requirement for Cloud service delivery, it is preferable to use the single path than splitting traffic to multiple paths.

Then we examined a case where using multiple paths to gain more bandwidth, which is the typical reason why multi-path networking is applied. We assume each single path offers 120 Mbps bandwidth; that is, we increase the total bandwidth by 120 Mbps each time a path is added in the
a Cloud service, using parallel paths for data transmission to existing single path can provide the required bandwidth for transmission. Results obtained from this case tell us that when no single path is supported, the multi-path system can still improve end-to-end delay performance for service delivery. Figure 8 (b) shows that increasing the number of paths will improve end-to-end delay performance for service delivery. Authors of [1] studied delay performance of Cloud service delivery. Although networking has a significant impact on Cloud service performance, the aforementioned research on modeling and analysis focuses on evaluating performance of Cloud data centers without sufficient investigation on network performance for Cloud service delivery. In [13] a Cloud computing system is modeled as an open queuing network with exponentially distributed service time and inter-arrival time. Yang and coauthors of [14] model a Cloud data center as a $M/M/m/m$ queueing system, in which both inter-arrival and service times are assumed to be exponentially distributed and the service response time is broken into independent waiting, service, and execution periods. In [8] an $M/G/m+r$ queueing system with single task arrivals is developed for analyzing Cloud service performance. Study on Cloud service performance under burst arrivals is reported in [7], where a Cloud center is modeled as a $M^{[s]}/G/m/m+r$ queueing system. Performance of Cloud computing centers with multiple priority classes is analyzed in [5] by using the birth-death process theory.

5. RELATED WORK

Modeling and analysis on Cloud service performance has attracted extensive research interest. In [13] a Cloud computing system is modeled as an open queuing network with exponentially distributed service time and inter-arrival time. Yang and coauthors of [14] model a Cloud data center as a $M/M/m/m$ queueing system, in which both inter-arrival and service times are assumed to be exponentially distributed and the service response time is broken into independent waiting, service, and execution periods. In [8] an $M/G/m+r$ queueing system with single task arrivals is developed for analyzing Cloud service performance. Study on Cloud service performance under burst arrivals is reported in [7], where a Cloud center is modeled as a $M^{[s]}/G/m/m+r$ queueing system. Performance of Cloud computing centers with multiple priority classes is analyzed in [5] by using the birth-death process theory.

Although networking has a significant impact on Cloud service performance, the aforementioned research on modeling and analysis focuses on evaluating performance of Cloud data centers without sufficient investigation on network performance for Cloud service delivery. Authors of [1] studied delay performance of Cloud service over virtual networks.
However, only network propagation delay was considered, which is just a small portion of the end-to-end delay for Cloud services. On the other hand, network service performance has been extensively studied and a wide variety of technologies for multi-path data transmissions have been developed and evaluated [10]. However, little research in this area is done with Cloud service delivery as an objective. Therefore currently available analysis methods treat Cloud computing and networking systems separately and performance of multi-path networking for Cloud service delivery is basically an unexplored problem.

In addition, traditional queuing analysis is based on assumptions about certain implementation mechanisms of the studied system, such as a Cloud data center or a networking system. However, resource virtualization and service orientation, which are key features of Cloud computing, decouple Cloud and network services from their implementations. The same services could be realized by different technologies that are transparent to service users. Therefore, traditional queuing model is not sufficient to face the challenge of evaluating converged network-Cloud service performance.

In order to tackle this challenging problem, a novel modeling approach is proposed in [3, 4] by exploiting network calculus to develop a general profile for characterizing service capabilities of both network and Cloud service components. This model offers a promising direction toward a general performance evaluation method that is applicable to heterogeneous network and Cloud systems that are virtualized and provided as services. However, research presented in [3, 4] only considered data communications through a single path between the service user and provider. In this paper, we apply network calculus to study performance of multipath networking systems for Cloud service delivery.

6. CONCLUSION

Networking plays a crucial role in Cloud service delivery and network QoS has a significant impact on Cloud service performance. Multiple network paths could be employed to alleviate the required bandwidth for data-intensive Cloud services. In this paper, we propose a novel model of networking systems with multiple paths for Cloud service delivery and develop analysis techniques for evaluating the maximum end-to-end delay and backlog performance of such systems. We applied network calculus in our modeling and analysis to address some new challenges brought by inter-domain networking, network virtualization, and separation of data and control planes. Results obtained in this paper indicate that multi-path networking systems not only enable Cloud service delivery that cannot be supported by single path transmission but also may improve end-to-end delay performance for Cloud service provisioning.

For future work, we plan to investigate the impact of variations in traffic characteristics on delay and backlog performance of Cloud service delivery with general multi-path, i.e., non-disjoint multi-path, data transmission.

ACKNOWLEDGEMENT

This work is supported by NSFC (Grant No. 61309031), Program for Innovation Team Building at Institutions of Higher Education in Chongqing (Grant No. KJTD201310), NSF of Chongqing (Grant No. cstc2013jcyjA40026), S&T Research Program of Chongqing Municipal Education Commission (Grant No. KJ130523), and CQUPU Research Fund for Young Scholars (Grant No. A2012-79).

7. REFERENCES