Digital Communications and Networks (▮▮▮▮) ▮, ▮▮▮–▮▮▮

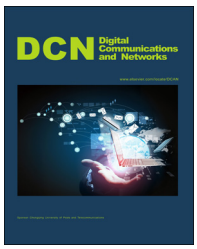
HOSTED BY

ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/dcan



Modeling and performance analysis for composite network-compute service provisioning in software-defined cloud environments

Qiang Duan

Information Sciences & Technology Department, The Pennsylvania State University Abington, PA 19001, United States

Received 18 December 2014; received in revised form 19 February 2015; accepted 16 May 2015

KEYWORDS
Cloud computing;
Composite service provisioning;
Software-defined Cloud environment;
Service modeling;
Performance analysis

Abstract
The crucial role of networking in Cloud computing calls for a holistic vision of both networking and computing systems that leads to composite network-compute service provisioning. Software-Defined Network (SDN) is a fundamental advancement in networking that enables network programmability. SDN and software-defined compute/storage systems form a Software-Defined Cloud Environment (SDCE) that may greatly facilitate composite network-compute service provisioning to Cloud users. Therefore, networking and computing systems need to be modeled and analyzed as composite service provisioning systems in order to obtain thorough understanding about service performance in SDCEs. In this paper, a novel approach for modeling composite network-compute service capabilities and a technique for evaluating composite network-compute service performance are developed. The analytic method proposed in this paper is general and agnostic to service implementation technologies; thus is applicable to a wide variety of network-compute services in SDCEs. The results obtained in this paper provide useful guidelines for federated control and management of networking and computing resources to achieve Cloud service performance guarantees.

1. Introduction

Cloud computing is a large scale distributed computing paradigm driven by economies of scale, in which a pool of abstracted, virtualized, dynamically scalable computing functions are delivered on demand as services to external customers over the Internet [1]. Networking plays a crucial role in Cloud computing. From a service provisioning perspective, the services received by Cloud end users comprise not only computing functions provided by Cloud data centers but also communications functions offered by networks. Results obtained from recent study on Cloud service performance have indicated that networking has a

E-mail address: qduan@psu.edu
Peer review under responsibility of Chongqing University of Posts and Telecommunications.

http://dx.doi.org/10.1016/j.dcan.2015.05.003

Please cite this article as: Q. Duan, Modeling and performance analysis for composite network-compute service provisioning in software-defined cloud environments, Digital Communications and Networks (2015), http://dx.doi.org/10.1016/j.dcan.2015.05.003

significant impact on quality of Cloud services, and in many cases data communications become a bottleneck that limits Clouds from supporting high-performance applications [2,3]. Therefore, networks with Quality of Service (QoS) capabilities become an indispensable ingredient for high-performance Cloud service provisioning.

For example, consider a scenario in which an application utilizes the Cloud infrastructure for storing and processing a large data set and requires upper bounded response delay. This application may use the computing capability of Amazon EC2 (Elastic Compute Cloud) and the storage capacity of Amazon S3 (Simple Storage Service). In order for the user to access EC2 and S3, network services must also be provided for data transmissions from the user to S3 virtual disk, between S3 disk and the EC2 server (if EC2 and S3 are located at different sites), and from EC2 server back to the user. Therefore, the end-to-end service offered to the user is essentially a composition of both Cloud services and network services. In order to meet the service delay requirement of the application, sufficient amount of networking resources (e.g. transmission bandwidth and packet forwarding capacity) must be provided to guarantee network delay performance in addition to the computing and storage resources offered by the Cloud infrastructure for meeting data processing and storing requirements.

The significant role that networking plays in Cloud service provisioning calls for a holistic vision of the computing and networking systems involved a Cloud environment. Such a vision requires federated management, control, and optimization of computing and networking resources for composite service provisioning. Software-Defined Networking (SDN) is one of the latest revolutions in the networking field, which decouples network control and data forwarding functions; thus enabling network control to be programmable and underlying network infrastructure to be abstracted for applications [4]. The logically centralized control plane in SDN allows upper layer applications to program the underlying network platform through a standard API; therefore can provide better support for Cloud computing. Expectation of more agile Cloud services also requires a high degree of programmability for computing infrastructure, which leads to software-defined computing and storage. Software-defined network, compute, and storage systems, when integrated together, enable an environment with fully automated provisioning and orchestration of IT infrastructures for Cloud computing, which is referred to as Software-Defined Cloud Environment (SDCE) [5].

A key to realize the SDCE notion is orchestration of heterogeneous network, compute, and storage systems for composite service provisioning. The Service-Oriented Architecture (SOA) provides an effective mechanism for heterogeneous system integration through loose-coupling interactions among system components. SOA has been widely adopted in Cloud computing via the IaaS, PaaS, and SaaS paradigms. Applying SOA in the field of networking leads to the *Network-as-a-Service* (*NaaS*) paradigm that enables encapsulation and virtualization of networking resources in the form of SOA-compliant *network services*. NaaS allows network infrastructure to be virtualized, exposed, and accessed as services that can be orchestrated with computing services in a Cloud environment to provision composite network-compute services to Cloud users [6]. Recently NaaS has been proposed as a key mechanism in SDN for achieving end-to-end QoS provisioning [7]. Therefore SOA may form a basis for service provisioning in SDCEs.

As SDCE being rapidly adopted by service providers, it becomes important to obtain thorough understanding about performance of composite network-compute service provisioning, which is the service performance actually perceived by Cloud users. Since networking has a strong impact on end-to-end performance for Cloud service provisioning, Cloud service performance should be evaluated with a holistic vision of both computing and networking aspects. For example, suppose a biology lab creates 100 GB of raw data that will be processed in Amazon EC2 Cloud. Assume that the lab obtained 10 EC2 virtual machine instances and each instance can process 20 GB data per hour, then the total processing time of the Cloud service is only 30 min. However, if the lab uses a network service that offers 200 Mb/s throughput for data transmission to the EC2 server, then even the single-trip delay for data transmission from the lab to EC2 servers will be 67 min. In this example network delay for round-trip data transmission contributes more than 80% of the total service delay; thus demonstrating the significant impact of networking on Cloud service performance.

Analytical modeling and analysis provide an effective approach to obtaining insights about end-to-end service performance. Cloud performance analysis has attracted attention of the research community and many results have been reported in the literatures. However, the significant impact of network performance on Cloud service provisioning has not been sufficiently considered in these works. On the other hand, although network performance has been extensively studied, currently available techniques typically lack the ability to analyze composite systems that consist of heterogeneous networking and computing functionalities. Therefore, system modeling and performance analysis for composite network-compute service provisioning in SDCEs are still an open problem.

Network-compute service orchestration in SDCEs brings in new challenges to service modeling and performance analysis. Key features of Cloud computing and NaaS, such as virtualization and abstraction of networking and computing resources, make service provisioning independent of system implementations; thus requiring modeling and analysis methods to be general and agnostic to network and Cloud implementations. In order to tackle this challenge, a novel modeling approach is proposed in this paper for characterizing the service capabilities of composite network-compute systems. Analysis techniques are developed based on this model for evaluating delay performance of composite network-compute services. The developed modeling and analysis techniques are general and agnostic to network and Cloud implementations; thus are applicable to composite network-compute service delivery systems in SDCEs.

The rest of this paper is organized as follows. An overview of network-compute service composition in SDCEs is given in Section 2. Related works on Cloud service performance evaluation and the new challenges brought in by network-compute service composition in SDCEs are discussed in Section 3. A new service capability model is proposed in Section 4 and applied in Section 5 to characterize capabil-

Please cite this article as: Q. Duan, Modeling and performance analysis for composite network-compute service provisioning in software-defined cloud environments, Digital Communications and Networks (2015), http://dx.doi.org/10.1016/j.dcan.2015.05.003

ities of composite network-compute service systems in SDCEs. A technique for evaluating delay performance for composite network-compute service is developed in Section 6. Section 7 provides numerical results for illustrating applications of the developed techniques. Section 8 draws conclusions.

2. Composite network-compute service provisioning in software-defined cloud environments

Virtualization is a key enabling technology for Cloud computing and the SOA forms a foundation for Cloud service provisioning. Recent research and development have been bridging the power of SOA and virtualization in the context of Cloud computing ecosystem [8]. The Open Grid Forum (OGF) is working on the Open Cloud Computing Interface (OCCI) standard, which defines SOA-compliant open interfaces for interacting with Cloud infrastructure. The SOA principle has strongly influenced the service models of most Cloud service providers. For example Amazon expose its compute and storage Cloud services via Web service interfaces.

Virtualization will play a crucial role in the next generation networks [9]. Network virtualization separates network service provisioning from data transport infrastructure. Applying SOA in network virtualization enables resources in network infrastructure to be abstracted as infrastructure services. Network service providers can construct different virtual networks by utilizing network infrastructure services to offer end-to-end network services for meeting diverse user requirements. Upper layer applications can utilize the underlying networking platform by accessing the network services through a standard abstract interface, which is essentially a *Network-as-a-Service* (NaaS) paradigm.

Software-Defined Network (SDN) represents a fundamental advancement in the field of networking. SDN technology also brings new possibility for Cloud service providers. By taking advantage of the logically centralized control of network resources, it is possible to simplify and optimize management of Cloud networks to achieve more flexible and efficient intra-data center networking and improved inter-datacenter communications. NaaS can be applied in SDN to further enhance flexibility and performance of network service provisioning. SDN and software-defined compute and storage systems form the three key components of a Software-Defined Cloud Environment (SDCE).

A framework for service-oriented SDCE architecture is shown in Fig. 1. The infrastructure layer in this framework comprises physical infrastructure for data transmission, processing, and storage. The control functions are decoupled from physical infrastructure and logically centralized at the software-defined network, compute, and storage controllers, which form the control layer of the framework. The NaaS paradigm exposes networking functionalities as SOA-compliant services in the same way as IaaS interface exposes compute and storage resources as services. The service layer supports high-level abstractions of both networking and computing systems and provides a unified mechanism to orchestrate network and compute services to form composite services, which are provisioned to user applications on the application layer. In this framework, service-oriented virtualization of both networking and computing systems enables federated control, management, and optimization of the resources in both networking and computing domains.

From a service provisioning perspective, the services offered to end users are composite services that comprise both computing functions (for data process and/or storage) and communication functions (for data transmission). A typical end-to-end provisioning system for composite network-compute services is shown in Fig. 2, which consists of Cloud infrastructure that offers compute service and the communication network that provides network services.

SDCE is an integrated service delivery environment in which network and compute services, used to be offered separately by different providers, converge into composite network-compute services. Therefore SDCE enables a new service model that allows the roles of traditional Internet service providers and Cloud service providers merge together for composite service provisioning. Such a new service environment may stimulate innovations in the development, deployment, and utilization of Cloud services; thus creating a wide variety of new business opportunities.

3. Challenges to performance evaluation of composite network-compute services

Service performance is one of the decisive factors in adoption of the new Cloud computing paradigm for various applications. Therefore, performance evaluation of Cloud services has attracted extensive research interest.

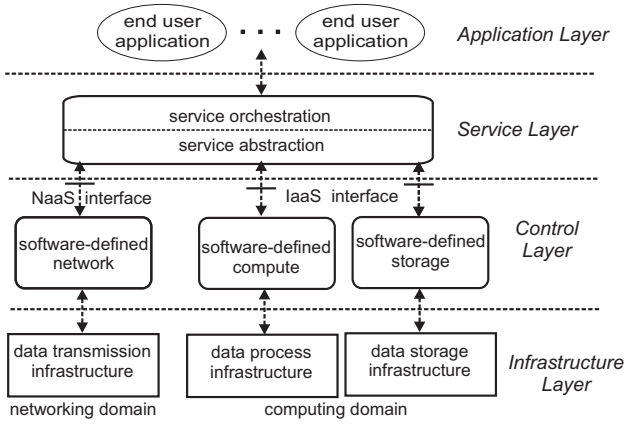


Fig. 1 A framework for software-defined cloud environment architecture.

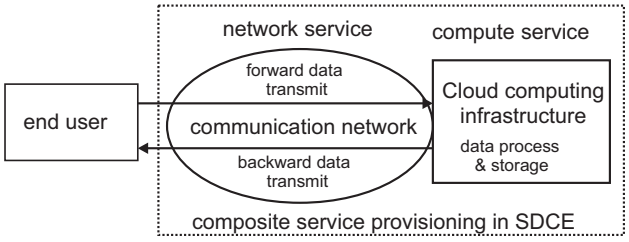


Fig. 2 Composite network-compute service provisioning in SDCE.

Please cite this article as: Q. Duan, Modeling and performance analysis for composite network-compute service provisioning in software-defined cloud environments, Digital Communications and Networks (2015), http://dx.doi.org/10.1016/j.dcan.2015.05.003

In [10] the authors developed an approach to evaluating the performance of various Cloud service offerings with different configurations. An experimental evaluation on Amazon Elastic Compute Cloud (EC2) was reported in the paper to verify this approach. Hill and Humphrey [11] have quantitatively evaluated EC2's performance for the common MPI-based scientific computing applications. The authors of [2] presented a comprehensive performance evaluation of Amazon EC2 for supporting high-performance computing using representative workloads of a typical supercomputing center. A portable and extensible framework for generating and submitting test workloads to computing Clouds was designed in [12]. The authors employed this framework to study Cloud response time in various metrics including overheads for acquiring and realizing virtual computing resources. In [13] the authors quantified the presence of many-task-computing users in real scientific computing workloads and performed an empirical evaluation of the performance of four commercial Cloud services. The authors of [14] examined the performance and cost of using Cloud computing for supporting scientific workflow applications, which consist of a set of loosely-coupled computing tasks connected through data- and control-flow dependencies. Li and his coauthors constructed a taxonomy of performance evaluation of commercial Cloud services in [15], which focuses on measurement-based Cloud service evaluation approaches.

The aforementioned research results are mainly based on measurement and testing experiments conducted on some particular types of computing Clouds that have been deployed and offered to the public, such as Amazon EC2 services. Cloud computing is a rapidly developing field in which new services as well as new implementations of existing services keep emerging. Measurement-based methods are not sufficient for determining the achievable performance of new Cloud services or services with new configuration settings. Therefore, analytical modeling and performance analysis techniques are necessary in order to obtain thorough understanding about service performance for general Cloud computing scenarios. This allows service customers and service brokers to predict the achievable service performance to discover and select the appropriate Cloud services for various applications. Analytical methods may also provide guidelines for Cloud service providers to manage and configure resources for service provisioning.

Queueing theory has been applied to develop analytical methods for evaluating Cloud service performance. Xiong and Perros [16] modeled a Cloud computing system as an open queue network consisting of two tandem servers with finite buffer space. In order to study resource allocation for meeting performance requirements of clients with different priority levels, the authors of [17] modeled a Cloud data center as an M/M/C/C queueing system, which has C servers with no buffer space. Yang et al. developed an $M/M/m/m+r$ queueing model for Cloud data centers in [18]. In this model both arrival time and service time are assumed to be exponentially distributed and service response time is broken into three independent parts: waiting, service, and execution periods. Due to the complexity of Cloud computing technologies and diversity of user requests, developing exact models for Cloud service systems is very difficult. In order to address this challenge, Khazaei and his coauthors of [19] proposed an approximate $M/G/m/m+r$ queuing model for Cloud server farms and solved it to obtain estimation of complete probability distribution of service response time and other performance indicators. This approximate model was extended in [20] to reflect burst arrival process of Cloud data centers, which leads to a $M^{[x]}/G/m/m+r$ queuing system where arrival process is a sequence of super-tasks each of which consists of a burst of tasks.

Although many research results indicate that networking has a strong influence on Cloud service performance, the aforementioned research works focus on evaluating performance of Cloud data centers without sufficiently considering the impact of network performance on Cloud service provisioning. On the other hand, although network performance and QoS have been extensively studied, little research that views networking as an integrated element of Cloud service provisioning has been reported. Currently available methods for performance analysis basically consider computing and networking systems separately; thus lacking a holistic vision for evaluating performance of composite network-compute service provisioning in a software-defined Cloud environment.

Traditional queueing analysis methods are all based on some assumptions about certain implementation mechanisms of the studied systems such as a Cloud data center or a networking system. However Cloud services and their implementation technologies are changing rapidly. In addition, resource virtualization and service orientation in Cloud computing and NaaS-based SDN decouple Cloud and network services from their implementations and make the latter transparent to applications. End users and service brokers select, orchestrate, and access services without knowledge of their detailed implementations. Therefore, traditional queuing theory-based analysis methods are not general enough for facing the challenges of evaluating composite network-Cloud service performance.

Composite service delivery systems in a SDCE consist of heterogeneous networking and computing systems with diverse implementations. Therefore, the modeling and analysis techniques for evaluating composite service performance must be general and applicable to the wide variety of systems coexisting in a SDCE. SOA for Cloud service provisioning and NaaS-based SDN enable both computing and networking resources to be abstracted as services and decouple service functions from their implementations. This requires the modeling and analysis techniques to be agnostic to the implementation technologies of networking and computing systems for composite service provisioning. Network calculus theory [21] has been applied in performance evaluation of Grid network services [22] and network virtualization [23,24]. In previous work [25,26], the author explored application of network calculus to tackle performance analysis for Cloud network services. In this paper, such a novel idea is fully developed into a systematic approach to modeling and analyzing composite network-compute service performance in order to meet the above requirements.

4. Modeling composite network-compute service provisioning systems

A typical delivery system for composite network-compute services in SDCEs is shown in Fig. 3, which consists of

Please cite this article as: Q. Duan, Modeling and performance analysis for composite network-compute service provisioning in software-defined cloud environments, Digital Communications and Networks (2015), http://dx.doi.org/10.1016/j.dcan.2015.05.003

network services for forward and backward data transmissions, compute services for data process in a Cloud data center, and data transform functions between the data transmissions and data process.

In order to receive performance guarantee from a composite service, the end user must expect a certain level QoS from both network and compute services. In general, such QoS expectation can be defined in the Service Level Agreements (SLAs) between the user and service providers. Although SLAs may vary due to the diversity of services, it typically includes a requirement on the minimum data transport rate for network services and the minimum data processing capacity for compute services.

In order to analyze the composite service performance, one must examine the communication capability offered by the network service and data processing capability provided by the compute service. The methodology taken in this paper is to first develop a general capability profile that can model service capabilities of both network and compute services, and then compose the capability profiles of the two service components into one end-to-end profile that models the service capability of the composite system. Such a capability profile should give a lower bound of the amount of service that a user can expect from the services (including both network and compute services), should be independent of implementation technologies of the underlying networking and computing infrastructures, and should also be easy to combine for modeling composite service capabilities. In order to meet these requirements, the concept of *service curve* from network calculus theory [21] is employed in this paper for developing such a general and flexible capability profile.

A capability profile for a service component can be defined as follows. Let $R(t)$ and $E(t)$ respectively be the accumulated amount of traffic that arrives at and departs from a service component by time t. Given a non-negative, non-decreasing function, $P(\cdot)$, where $P(0)=0$, we say that the service component has a capability profile $P(t)$, if for any $t \geq 0$ in the busy period of the service component

$$E(t) \geq R(t) \otimes P(t) \tag{1}$$

where $\otimes$ denotes the operation defined as $x\left(t\right) \otimes y\left(t\right) = \inf_{s:0 \leq s \leq t}\{x(t-s)+y(s)\}$.

The capability profile of a service component, which is essentially a service curve of the component, is defined as a general function of time that specifies service capability through the relation between arrival and departure traffic at the service component. Therefore such a profile is independent of the implementations of service components, thus is applicable to both network and compute service components with various implementations.

The capability profiles defined in (1) gives a general approach to modeling network and Cloud service

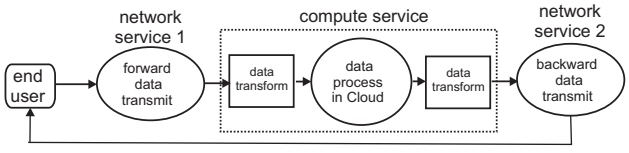


Fig. 3 A typical delivery system for composite network-compute service in SDCE.

capabilities. In order to obtain a more tractable profile that can characterize capabilities of typical network and compute services, this paper defines a *Latency-Rate* (*LR*) profile for a service component as follows. If a service component has a capability profile

$$\mathcal{P}(r,\theta) = r[t-\theta]^{+} \tag{2}$$

where $[\cdot]^{+}$ is equal to its argument if it is positive and zero otherwise; then the service component has a *LR* profile. The θ and r are respectively called the *latency* and *rate* parameters of the profile.

A *LR* profile can serve as the capability model for typical network services. The QoS expectation of a typical network service includes the minimum data transmission bandwidth guaranteed to a service user, which is described by the rate parameter r in a *LR* profile. Data communication in a network also experiences a fixed delay that is independent of traffic queuing behavior; for example signal propagation delay, link transmission delay, router/switch processing delay, etc. The latency parameter θ of a *LR* profile is to characterize this part of fixed delay of a network service.

A *LR* profile can also characterize service capabilities of typical computing systems. Cloud service providers typically offer some certain service capacity units to users. For example each type of virtual machine (called instance) in Amazon EC2 provides a predictable amount of computing capacity and I/O bandwidth. Each EC2 compute unit provides the equivalent CPU capacity of 1.0–1.2 GHz 2007 Opteron or 2007 Xeon processor. Amazon also claims an internal I/O bandwidth of 250 Mb/s regardless of instance type. The latency and rate parameters of the *LR* profile for a Cloud service can be derived from the processing capacity and I/O bandwidth information specified by its provider.

In order to represent the data transformation effect of computing function provided by Cloud infrastructure, the concepts of scaling function and scaling curve, which were originally developed in [27] as an extension to network calculus, are adopted in the model for composite network-compute service provisioning systems.

A scaling function is defined as a function $S(\cdot)$ that assigns an amount of scaled data $S(a)$ to an amount of data a. As can be seen from the definition, scaling function is a general concept for taking into account data transformation in a system model. Note that it does not model any queuing effect - a scaling function is assumed to have zero delay. Queuing related effect of Cloud computing is modeled by the service curve-based capability profile of the compute service component.

Given a scaling function S, the function $\underline{\mathcal{S}}$ is called a (minimum) scaling curve of S iff $\forall b \geq 0$ it applies that $\underline{\mathcal{S}}(b) \leq \inf_{a \geq 0}\{S(b+a)-S(a)\}$.

Applying the above defined capability profile and scaling curve, a composite network-compute service provisioning system can be modeled with capability profiles of network and compute service components and the scaling curves that represent data transform between networking and computing, as shown in Fig. 4. In this system the network services for forward data transmission (from user to data center) and backward data transmission (from data center to user) are respectively modeled by the profiles $P_{n1}(t)$ and $P_{n2}(t)$. The compute service offered by the data center is modeled by the profile $P_C(t)$. The scaling curve $\underline{\mathcal{S}}_{n2c}$ models

Please cite this article as: Q. Duan, Modeling and performance analysis for composite network-compute service provisioning in software-defined cloud environments, Digital Communications and Networks (2015), http://dx.doi.org/10.1016/j.dcan.2015.05.003

data transform from forward transmission to computing server while the scaling curve $\underline{\mathcal{S}}_{c2n}$ models data transform from computing server to backward data transmission.

5. End-to-end capability profile for composite network-compute services

This section presents a technique for integrating the capability profiles of all service components and scaling curves in a composite network-compute service system into an end-to-end profile. Such an end-to-end profile models the service capability guaranteed by the service delivery system to its end user, which forms the basis for analyzing delay performance of composite network-compute services.

It is known from network calculus theory that the service curve of a system consisting of a series of tandem servers can be obtained from convolution of the service curves of all these servers. The capability profile defined in (1) is essentially the service curve of a service component. However, due to the scaling curves for data transform between networking and computing, the convolution operation cannot be directly applied to this model. In order to solve this problem, the alternative scaled servers (Theorem 3.1 in [27]) are used to modify the system model so that an end-to-end profile can be obtained through network calculus convolution.

Considering the two systems shown in Fig. 5, system (a) consists of a server with service curve β whose output is scaled with a scaling function $\mathcal{S}$; and system (b) consists of a scaling function $\mathcal{S}$ whose output is input to a server with a service curve β_S. It is proved in [27] that given system (a), the lower bound of system (b) output function $\underline{\mathcal{S}}(R) \otimes \beta_S$ is a valid lower bound for the output function of system (a), if $\beta_S = \underline{\mathcal{S}}(\beta)$. Given system (b), the lower bound of system (a) output function $\underline{\mathcal{S}}(R \otimes \beta)$ is a valid lower bound for the output function of system (b), if $\beta = \underline{\mathcal{S}}^{-1}(\beta_S)$ where $\underline{\mathcal{S}}^{-1}$ stands for the inverse function of $\underline{\mathcal{S}}$.

This means in effect that performance bounds for systems (b)/(a) under the respective assumption are also valid bounds for system (a)/(b). Therefore, a scaling function and a service component in a model can be switched without changing performance bounds, as long as the capability profile of the service component is transformed using the scaling curve. In addition it is also shown in [27] that performance bounds obtained in the alternative systems after the above switch operation remain tight.

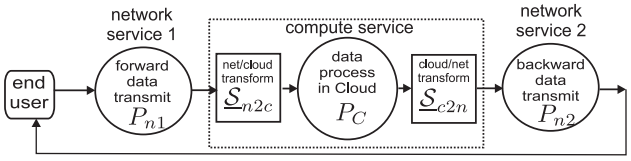


Fig. 4 A profile-based capability model for composite network-compute service provisioning.

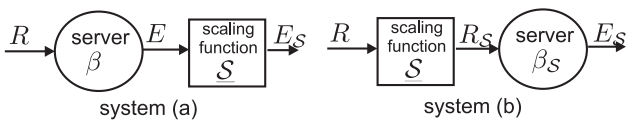


Fig. 5 Alternative service systems with a scaling function.

Applying the above alternative server method in the model shown in Fig. 4, network service 1 for forward data transmission and the scaling function for network/compute data transform can be switched without impacting the performance bound guaranteed by the system, if the capability profile of network service 1 is transformed to $P^{\mathcal{S}}_{n1} = \underline{\mathcal{S}}_{n2c}(P_{n1})$. Similarly, the scaling function for compute/network data transform and network service 2 for backward data transmission can be switched, if the capability profile of network service 2 is transformed to $P^{\mathcal{S}}_{n2} = \underline{\mathcal{S}}^{-1}_{c2n}(P_{n2})$. Then the composite service system has the alternative model as shown in Fig. 6.

Since the capability profile defined in (1) is essentially the service curve of a service component, the capability profiles for both network service components (for forward and backward data transmissions) and the compute service component can be integrated into one profile using the convolution operation defined in network calculus. Therefore, the end-to-end capability profile for the composite system, denoted by $P_{e2e}(t)$, can be determined as

$$P_{e2e}(t) = \underline{\mathcal{S}}_{n2c}(P_{n1}(t)) \otimes P_C(t) \otimes \underline{\mathcal{S}}^{-1}_{c2n}(P_{n2}(t)). \tag{3}$$

Suppose each service component in a composite network-compute system has a *LR* profile; that is, $P_{n1} = \mathcal{P}[r_1, \theta_1]$, $P_C = \mathcal{P}[r_C, \theta_C]$, and $P_{n2} = \mathcal{P}[r_2, \theta_2]$. Then the transformed profile for network service 1 is

$$P^{\mathcal{S}}_{n1} = \underline{\mathcal{S}}_{n2c}(P_{n1}) = \mathcal{P}[\underline{\mathcal{S}}_{n2c}(r_1), \theta_1] \tag{4}$$

and the transformed profile for network service 2 will be

$$P^{\mathcal{S}}_{n2} = \underline{\mathcal{S}}^{-1}_{c2n}(P_{n2}) = \mathcal{P}[\underline{\mathcal{S}}^{-1}_{c2n}(r_2), \theta_2]. \tag{5}$$

Then it can be proved that the end-to-end capability profile of the composite service provisioning system is

$$\begin{aligned} P_{e2e} &= \underline{\mathcal{S}}_{n2c}(\mathcal{P}[r_1, \theta_1]) \otimes \mathcal{P}[r_C, \theta_C] \otimes \underline{\mathcal{S}}^{-1}[\mathcal{P}(r_2, \theta_2)] \\ &= \mathcal{P}[\underline{\mathcal{S}}_{n2c}(r_1), \theta_1] \otimes \mathcal{P}[r_C, \theta_C] \otimes \mathcal{P}[\underline{\mathcal{S}}^{-1}_{c2n}(r_2), \theta_2] \\ &= \mathcal{P}[r_e, \theta_e] \end{aligned} \tag{6}$$

where

$$r_e = \min\left\{\underline{\mathcal{S}}_{n2c}(r_1), r_C, \underline{\mathcal{S}}^{-1}_{c2n}(r_2)\right\}, \quad \theta_e = \theta_1 + \theta_C + \theta_2. \tag{7}$$

Eqs. (6) and (7) imply that for a composite network-compute service provisioning system, if all service components, including both forward and backward network services and the compute service, can be modeled by *LR* profiles, then the end-to-end service capability of the composite provisioning system can also be modeled by a *LR* profile with the network/compute scaling curve in front of it and the compute/network scaling curve behind it. The latency parameter of the end-to-end *LR* profile is the summation of latency parameters of all service components

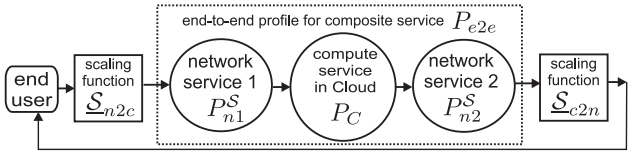


Fig. 6 Alternative model for composite network-compute service provisioning.

Please cite this article as: Q. Duan, Modeling and performance analysis for composite network-compute service provisioning in software-defined cloud environments, Digital Communications and Networks (2015), http://dx.doi.org/10.1016/j.dcan.2015.05.003

in the system. The service rate parameter of the end-to-end *LR* profile is determined by the minimum value of the Cloud service rate and the transformed service rates of forward and backward network services.

Suppose the data transform between network and compute services can be characterized by a piece-wise linear scaling curve $\underline{\mathcal{S}}(a)=\min_{i=1,\cdots N}\{p_i a+q_i\}$, then $\underline{\mathcal{S}}^{-1}(b)=\max_{i=1,\cdots N}\left\{\frac{1}{p_i}[b-q_i]^+\right\}$. Therefore, the transformed capability profile for network services 1 and 2 will be

$$P_{n1}^{\mathcal{S}}=\underline{\mathcal{S}}\left(P_{n1}\right)=\min_{i=1,\cdots N}\{p_i[r_1(t-\theta_1)]+q_i\} \tag{8}$$

and

$$P_{n2}^{\mathcal{S}}=\underline{\mathcal{S}}^{-1}(P_{n2})=\max_{i=1,\cdots N}\left\{\frac{1}{p_i}[r_2(t-\theta_2)-q_i]^+\right\} \tag{9}$$

6. Delay performance analysis for composite network-compute service provisioning

Based on the service model presented in the preceding section, this section will develop a delay analysis technique for composite network-compute service provisioning. The analysis focuses on the maximum service response delay (between the time instants when a user sends out a request to the Cloud data center and receives the corresponding response), which is a significant performance parameter for high-performance Cloud applications.

Delay analysis for a service provisioning system needs an approach to characterizing the load on the system since a system with a certain service capacity achieves different levels of delay performance under different loads. Due to the diversity in traffic generated by various Cloud applications, a general profile is needed to specify the load for composite service systems. The concept of *arrival curve* from network calculus is employed here to define a general load profile as follows.

Let $R(t)$ denote the accumulated amount of traffic that arrives at the entry of a composite network-compute service provisioning system by any time instant t. Given a non-negative, non-decreasing function, $\mathcal{L}(\cdot)$, the service system is said to have a *load profile* $\mathcal{L}(t)$ if for all time instants s and t such that $0<s<t$

$$R(t)-R(s)\le\mathcal{L}(t-s). \tag{10}$$

A load profile gives an upper bound for the amount of traffic that the end user can load on a service provisioning system. Since the profile is defined as a general function of time, it can be used to describe traffic load generated by any application.

Currently most QoS-capable networking systems apply traffic regulation mechanisms at network boundaries to shape arrival traffic from end users. The traffic regulators that are most commonly used in practice are leaky buckets. A networking session constrained by a leaky bucket controller has a traffic load profile

$$\mathcal{L}[p,\rho,\sigma]=\min\{pt,\sigma+\rho t\}, \tag{11}$$

where p, ρ, and σ are respectively called the peak rate, sustained rate, and maximal burst size for the traffic.

The maximum service delay guaranteed by a composite network-compute system to an end user is determined by two factors: (i) service capacity offered by the system to the user, which is modeled by the end-to-end capability profile; and (ii) characteristic of the traffic that the user loads the system, which is described by a load profile. The entry of the composite network-compute service system where user applications load the system is the boundary of the forward networking system; therefore the traffic load for forward data transmission is the load of the composite service system. Based on the alternative model given in Fig. 6, service delay performance is determined by the end-to-end profile $P_{e2e}(t)$ because scaling functions do not introduce any delay. Due to the network/compute scaling curve $\underline{\mathcal{S}}_{n2c}$ in front of the end-to-end profile, the actual load that determines service delay performance should be characterized by a transformed load profile $\mathcal{L}^{\mathcal{S}}(t)=\underline{\mathcal{S}}_{n2c}(\mathcal{L}(t))$, as shown in Fig. 7. Therefore, given the end-to-end capability profile $P_{e2e}(t)$ of a composite network-compute service system, the maximum service delay d_{max} guaranteed by the system to the user can be determined as

$$d_{max}=\max_{t:t\ge 0}\left\{\min\left\{\delta:\delta\ge 0\ \underline{\mathcal{S}}_{n2c}(\mathcal{L}(t))\le P_{e2e}(t+\delta)\right\}\right\}. \tag{12}$$

Suppose a composite network-compute system has a *LR* profile for each service component and a leaky-bucket load profile $\mathcal{L}[p,\rho,\sigma]$, then the transformed load profile for the system is $\underline{\mathcal{S}}_{n2c}(\mathcal{L}[p,\rho,\sigma])$. Following (6) and (11), the maximum service delay guaranteed by this system can be determined as

$$d_{max}=\theta_{\Sigma}+\left(\frac{\underline{\mathcal{S}}_{n2c}(p)}{r_e}-1\right)\frac{\underline{\mathcal{S}}_{n2c}(\sigma)}{\underline{\mathcal{S}}_{n2c}(p)-\underline{\mathcal{S}}_{n2c}(\rho)} \tag{13}$$

where $\theta_{\Sigma}=\theta_{n1}+\theta_{n2}+\theta_C$ is the total service latency including round-trip network latency and latency of the compute service.

The latency parameter of a *LR* profile for a network service reflects a system property of the network that may be seen as the worst-case delay experienced by the first traffic bit in a busy period of a networking session. Therefore, this parameter can be estimated based on link transmission delay and packet processing delay, if signal propagation delay is assumed to be ignorable; that is, $\theta\cong L/r+L/R$, where L and R are respectively the maximum packet length and maximum link rate of the network service.

Suppose the forward and backward networks services of the composite network-compute service provisioning system have identical link rate R and packet length L, then the maximum response delay performance of the composite service becomes

$$d_{max}=L\sum_{i=1,2}\left(\frac{1}{r_i}+\frac{1}{R_i}\right)+\theta_C+\left(\frac{p}{r_e}-1\right)\frac{\sigma}{p-\rho}. \tag{14}$$

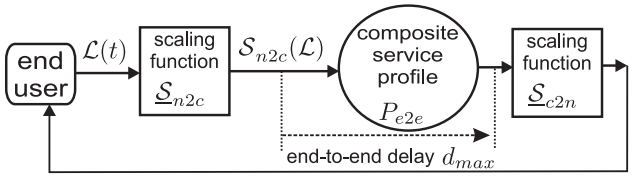


Fig. 7 End-to-end delay model for composite network-Cloud service provisioning.

Please cite this article as: Q. Duan, Modeling and performance analysis for composite network-compute service provisioning in software-defined cloud environments, Digital Communications and Networks (2015), http://dx.doi.org/10.1016/j.dcan.2015.05.003

7. Numerical results

This section gives numerical examples for illustrating applications of the developed modeling and analysis techniques. Considering the service provisioning system as shown in Fig. 2, in which an end user transmits data to a Cloud data center for processing and then receives the processed data back from the Cloud. Based on the measurement results reported in [2,3], traffic parameters of the load profile for this testing case are assumed to be 320 Mb/s, 120 Mb/s, and 200 kbits for the peak rate, sustained rate, and burst size respectively. For simplicity, forward and backward data transmissions are assumed to be provided by the same network service with a 10 Gb/s link capacity and a packet length of 1500 bytes.

A communication intensive service scenario was first examined. In this scenario data transform from forward transmission to the computing server decreases the load with a scaling factor $\mathcal{S}_{n2c} = 1/4$ and data transform from computing server to backward transmission increases the load with a scaling factor $\mathcal{S}_{c2n} = 2$. The impacts of network service rate and computing server capacity on the maximum service delay were analyzed and the obtained results are shown in Fig. 8. In this scenario we considered two cases in which latency parameters of the computing server are 150 μs and 300 μs. The obtained end-to-end delay upper bounds for both cases are denoted respectively as d^e_{c1} and d^e_{c2} in Fig. 8. From this figure we can see that both curves drop with increasing network service rate, which indicates that leasing more bandwidth from the network service providers may significantly improve end-to-end delay performance for composite service provisioning. Comparing the two curves of d^e_{c1} and d^e_{c2} shows that given the same network service rate, smaller server latency may give a tighter end-to-end delay bound but its impact is not as significant as that of increasing network bandwidth.

The relationship between the maximum service delay and available computing capacity in this scenario was also analyzed with a given network service rate. The obtained results are plotted in Fig. 9, in which d^e_{n1}, d^e_{n2}, and d^e_{n3} respectively denote the service delay bounds when network service rate is 150, 200, and 250 Mbps. The three curves in Fig. 9 are all basically flat, which implies that given a certain network service rate, increasing computing server capacity has almost no impact on the maximum end-to-end delay of this service case. Comparing the three delay curves in Fig. 9 shows that higher network service rate gives lower service delay. This is because data transmission between user and the computing server forms a service bottleneck in this scenario that determines the worst-case delay performance; therefore allocating more server capacity in data center does not help improving end-to-end delay performance. Notice that Fig. 8 indicates that less latency at the computing server does help reducing service delay to a certain degree. However, this latency parameter is mainly determined by data center implementation such as hardware and operating system speed, which cannot be easily reduced. Therefore, the above results shown in Figs. 8 and 9 tell us that for such communication intensive services, leasing sufficient bandwidth from network service providers instead of purchasing more computing service capacity is the most effective strategy for achieving an end-to-end service delay guarantee.

Then we examined a computing intensive service scenario in which data transforms between network and compute services increase the load for data processing with a scaling factor $\mathcal{S}_{n2c} = 2$ and decrease the load for data transmission with a scaling factor $\mathcal{S}_{c2n} = 1/4$. For this scenario the end-to-end delay upper bounds achieved with different network service rates were analyzed and the results are plotted in Fig. 10. The delay bounds were obtained with three computing capacity values, $r_c =$ 100, 125, and 150 Mbps, and their delay bounds are denoted as d^e_{c1}, d^e_{c2}, and d^e_{c3} respectively in Fig. 10. All three curves in this figure drop only slightly with increasing network service rate, which implies that leasing more bandwidth from network service makes little contribution to reducing end-to-end service delay in this scenario. Comparing the three delay curves shows that for a given network service rate, increasing compute service capacity may significantly improve end-to-end service delay performance in this scenario.

The end-to-end service delay bounds with various compute service capacities are given in Fig. 11, in which d^e_{n1}, d^e_{n2}, d^e_{n3} denote the delay bounds obtained with 100, 150,

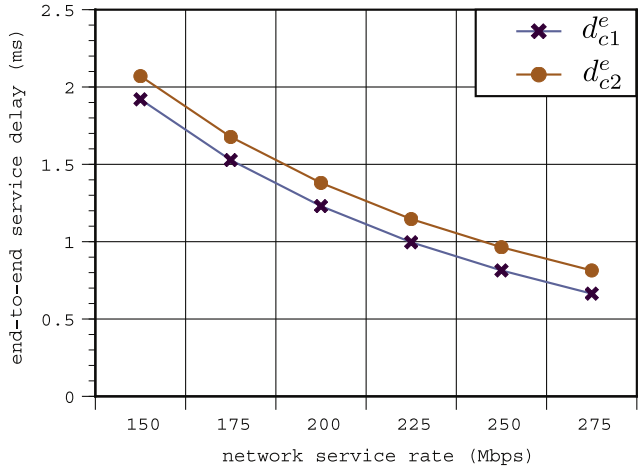


Fig. 8 End-to-end service delay vs. network service rate for a communication intensive application.

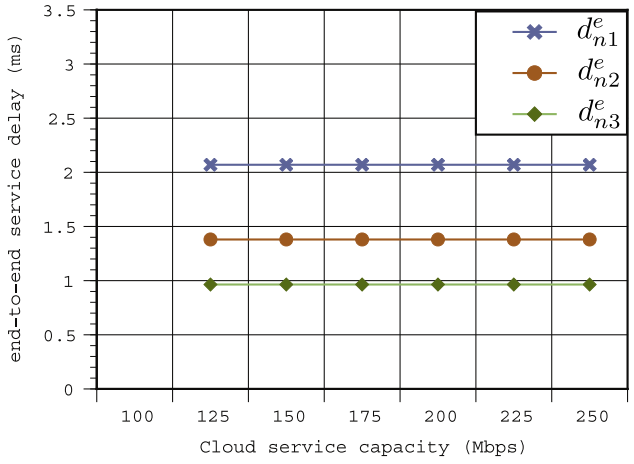


Fig. 9 End-to-end service delay vs. compute server capacity for a communication intensive application.

Please cite this article as: Q. Duan, Modeling and performance analysis for composite network-compute service provisioning in software-defined cloud environments, Digital Communications and Networks (2015), http://dx.doi.org/10.1016/j.dcan.2015.05.003

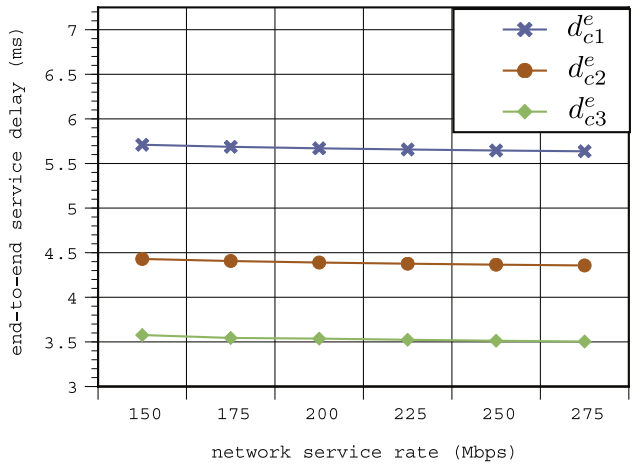


Fig. 10 End-to-end service delay vs. network service rate for a computing intensive application.

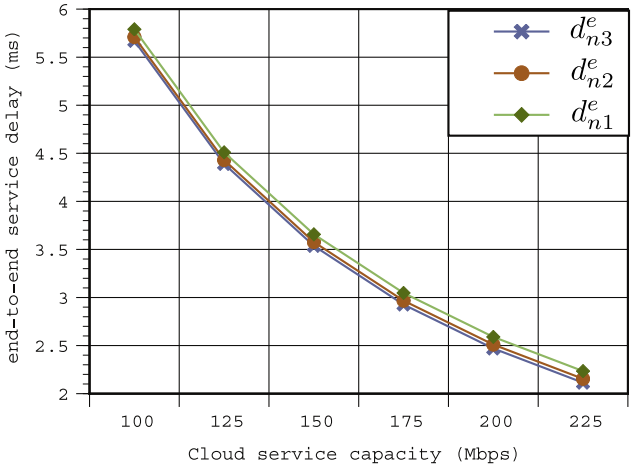


Fig. 11 End-to-end service delay vs. compute server capacity for a computing intensive application.

and 200 Mbps network service rates. This figure shows that maximum delay of the composite service decreases significantly with increasing compute service capacity, which confirms the observation we obtained from Fig. 10. Fig. 11 also shows that the three delay curves are very close to each other although not completely overlap, which implies that given the same compute service capacity, different amounts of network bandwidth do not change the delay bound much. The results shown in Figs. 10 and 11 indicate that in this scenario computing server capacity is the decisive factor for the maximum service delay and increasing network service rate only has minor contribution to delay performance improvement. This is because the compute service forms a bottleneck for this computing intensive scenario; therefore obtaining sufficient computing capacity in the data center is the key to achieving end-to-end delay guarantee for the composite service.

8. Conclusions

The crucial role of networking in Cloud computing requires a holistic vision of both networking and computing resources in a software-defined Cloud environment, which leads to composition of network and compute service provisioning. The research presented in this paper studies the problem of modeling and performance analysis for composite network-compute service provisioning systems. The main contributions made in this paper include a new approach to modeling service capabilities of composite network-compute service systems and analysis techniques for evaluating delay performance of composite network-compute services. Application of the recent development in network calculus with data scaling makes the modeling and analysis techniques developed in this paper general and agnostic to network and Cloud implementations; thus are applicable to various heterogeneous networking and computing systems coexisting in a Cloud environment for composite network-compute service provisioning. Both analytical and numerical results obtained in this paper indicate that capacities of network and compute services as well as the data transform factors between these two types of services have direct impact on delay performance of composite services. The developed model and analysis techniques provide Cloud service customers, service brokers, and service providers with an effective tool for evaluating achievable service performance in software-defined Cloud environments.

References

[1] I. Foster, Y. Zhao, I. Raicu, S. Lu, Cloud computing and Grid computing 360-degrees compared, in: Proceedings of the 2008 Grid Computing Environment Workshop, November 2008, pp. 1–10.

[2] K.R. Jackson, K. Muriki, S. Canon, S. Cholia, J. Shalf, Performance analysis of high performance computing applications on the Amazon Web services Cloud, in: Proceedings of the 2010 IEEE International Conference on Cloud Computing Technology and Science, November 2010, pp. 159–168.

[3] G. Wang, T.S.E. Ng, The impact of virtualization on network performance of Amazon EC2 data center, in: Proceedings of the IEEE INFOCOM2010, March 2010, pp. 1–9.

[4] ONF, Open Networking Foundation Software-Defined Networking (SDN) Definition, ⟨https://www.opennetworking.org/sdn-resources/sdn-definition⟩, 2013.

[5] C. Li, B. Brech, S. Crowder, D. Dias, H. Franke, M. Hogstrom, D. Lindquist, G. Pacifici, S. Pappe, B. Rajaraman, J. Rao, R. Ratnaparkhi, R. Smith, M. Williams, Software-defined environments: an introduction, IBM J. Res. Dev. 58 (March) (2014) 1–11.

[6] Q. Duan, Y. Yan, T.V. Vaslikos, A survey on service-oriented network virtualization toward a convergence of networking and Cloud computing, IEEE Trans. Netw. Serv. Manag. 9 (4) (2012) 373–392.

[7] Q. Duan, Network-as-a-Service in Software-Defined Networks for end-to-end QoS provisioning, in: Proceedings of the 2014 IEEE Wireless and Optical Communications Conference, May 2014.

[8] L.-J. Zhang, Q. Zhou, CCOA: Cloud computing open architecture, in: Proceedings of the First Symposium on Network System Design and Implementation, April 2009.

[9] N. Feamster, L. Gao, J. Rexford, How to lease the Internet in your spare time, ACM SIGCOMM Comput. Commun. Rev. 37 (1) (2007) 61–64.

[10] V. Stantchev, Performance evaluation of Cloud computing offerings, in: Proceedings of the Third International

Please cite this article as: Q. Duan, Modeling and performance analysis for composite network-compute service provisioning in software-defined cloud environments, Digital Communications and Networks (2015), http://dx.doi.org/10.1016/j.dcan.2015.05.003

Conference on Advanced Engineering Computing and Applications in Sciences, October 2009, pp. 187–192.

[11] Z. Hill, M. Humphrey, A quantitative analysis of high performance computing with Amazon's EC2 infrastructure: The death of the local cluster?, in: Proceedings of the 10th IEEE/ACM International Conference on Grid Computing, October 2009, pp. 26–33.

[12] N. Yigitbasi, A. Iosup, D. Epema, S. Ostermann, C-meter: a framework for performance analysis of computing Clouds, in: Proceedings of the Ninth IEEE/ACM International Symposium on Cluster Computing and the Grid, 2009, pp. 472–477.

[13] A. Iosup, S. Ostemann, M. Yigitbasi, R. Prodan, T. Fahringer, D. Epema, Performance analysis of Cloud computing services for many-tasks scientific computing, IEEE Trans. Parallel Distrib. Syst. 22 (6) (2011) 931–945.

[14] G. Juve, E. Deelman, K. Vahi, G. Mehta, B. Berriman, B.P. Berman, P. Maechling, Scientific workflow applications on Amazon EC2, in Proceedings of the Fifth IEEE International Conference on e-Science Workshop, December 2009.

[15] Z. Li, L. O'Brien, R. Cai, H. Zhang, Towards a taxonomy of performance evaluation of commercial Cloud services, in: Proceedings of the Fifth IEEE International Conference on Cloud Computing, June 2012, pp. 344–351.

[16] K. Xiong, H. Perros, Service performance and analysis in cloud computing, in: Proceedings of the IEEE 2009 World Conference on Services, 2009, pp. 693–700.

[17] W. Ellens, M. Zivkovic, J. Akkerboom, R. Litijens, H. Berg, Performance of Cloud computing centers with multiple priority classes, in: Proceedings of the Fifth IEEE International Conference on Cloud Computing, June 2012, pp. 245–252.

[18] B. Yang, F. Tan, Y. Dai, S. Guo, Performance evaluation of Cloud service considering faulty recovery, in: Proceedings of the First International Conference on Cloud Computing, December 2009, pp. 571–576.

[19] H. Khazaei, J. Misic, V.B. Misic, Performance analysis of Cloud computing centers using M/G/m/m+r queuing systems, IEEE Trans. Parallel Distrib. Syst. 23 (5) (2012) 936–943.

[20] H. Khazaei, J. Misic, V.B. Misic, Performance analysis of Cloud centers under burst arrivals and total rejection policy, in: Proceedings of the IEEE Globecom2011, December 2011, pp. 1–6.

[21] J.-Y.L. Boudec, P. Thiran, Network calculus: a theory of deterministic queueing systems for the Internet, in: Springer Lecture Notes in Computer Science, 2001.

[22] Q. Duan, Network service description and discovery for high-performance ubiquitous and pervasive Grids, ACM Trans. Auton. Adapt. Syst. (TAAS) 6 (1) (2011) 3.

[23] Q. Duan, Analysis on quality of service provisioning for communication services in network virtualization, J. Commun. 7 (2) (2012) 143–154.

[24] Q. Duan, End-to-end modelling and performance analysis for network virtualisation in the next generation internet, Int. J. Commun. Netw. Distrib. Syst. 8 (1) (2012) 53–69.

[25] Q. Duan, Modeling and performance analysis on network virtualization for composite network-Cloud service provisioning, in: Proceeding of the 2011 IEEE World Congress on Services, July 2011, pp. 548–555.

[26] Q. Duan, Modeling and delay analysis for converged network-Cloud service provisioning systems, in: Proceedings of the 2013 IEEE International Conference on Computing, Networking & Communications, February 2013, pp. 66–70.

[27] M. Fidler, J.B. Schmitt, On the way to a distributed systems calculus: an end-to-end network calculus with data scaling, in: Proceedings of SIGMetrics/Performance'06, 2006, pp. 287–298.

Please cite this article as: Q. Duan, Modeling and performance analysis for composite network-compute service provisioning in software-defined cloud environments, Digital Communications and Networks (2015), http://dx.doi.org/10.1016/j.dcan.2015.05.003