Service-Oriented Network Virtualization for Composition of Cloud Computing and Networking

Qiang Duan
The Pennsylvania State University

Computer networks play a crucial role in Cloud service provisioning and network Quality of Service (QoS) has a significant impact on Cloud service performance. Therefore networking and Clouding computing systems should be modeled and analyzed as a composite service provisioning system in order to obtain thorough understanding about the user's perception of Cloud service performance. Network virtualization is one of the latest developments in the networking area, which de-couples networking services from network infrastructures. The Service-Oriented Architecture (SOA) serves as a key enabler in both network virtualization and Cloud computing; thus offering a promising basis for network and Cloud composition. The research work presented in this article investigates application of SOA in network virtualization for composing network and Cloud services, and develops modeling and analyzes techniques for evaluating performance of composite network–Cloud service provisioning. This article proposes a SOA-based network virtualization paradigm, describes a service-oriented framework for composing network and Cloud services, proposes a new approach to modeling service capabilities of composite network–Cloud service provisioning systems, and develops analysis techniques for determining the performance that can be offered by composite network–Cloud services to their end users.

Keywords: Network virtualization, the Service-Oriented Architecture, Cloud computing.

1. INTRODUCTION

One of the most significant recent progresses in the field of information technology is Cloud computing, which may change the way people do computing and manage information. Cloud computing can be defined as a large scale distributed computing paradigm that is driven by economics of scale, in which a pool of abstracted, virtualized, dynamically-scalable computing functions and services are delivered on demand to external customers over the Internet [Foster et al. 2008]. A Cloud is massively scalable and can be encapsulated as an abstract entity that delivers different levels of services to customers.

Networking plays a crucial role in Cloud computing. As described in the definition given by [Foster et al. 2008] Cloud functions and services are delivered to customers over the Internet. From a user perspective, Cloud service provisioning consists of not only computing functions provided by the Cloud infrastructure but also data communication services offered by the Internet. Measurement results obtained from recent testing of some commercial Clouds, such as Amazon EC2, have indicated that networking performance has a significant impact on the quality of Cloud services, and in many cases data communications become the bottleneck that limits Clouds from supporting high-performance applications [Jackson et al. 2010; Wang and Ng 2010]. Networks with Quality of Service (QoS) capabilities become an indispensable ingredient for Cloud service provisioning; therefore networking and Cloud computing systems should be seamlessly integrated into a composite service provisioning system in order to offer high-performance Cloud services.

However there exists a gap between the demands of Cloud computing for data communications and the services that can be offered by traditional networking systems. High-performance Cloud computing requires predictability in networking performance, coordination of both computing and networking resources, and application-driven network control and management. However, traditional networks were designed specifically to support a narrow range of precisely defined

Author's address: Qiang Duan, Information Science and Technology Department, The Pennsylvania State University Abington College, 1600 Woodland Road, Abington, PA, U.S.A.

A preliminary version of this paper was presented at the IEEE CloudPerf 2011 Workshop.

communication services. These services were implemented on fairly rigid infrastructures, with minimal capabilities for ad hoc reconfiguration. Operations, managements, and security functions in traditional networks were also specifically designed and customized to facilitate particular types of services. Therefore, network service provisioning is tightly coupled with network infrastructures, thus making the development and deployment of new network services slow and static. Networking resources are managed separately from computing resources; thus lacking an optimization of resource utilization.

Network virtualization is a significant recent development in the networking area, which is expected to play a key role in the next generation Internet. Essentially network virtualization separates network service provisioning functions from data transportation mechanisms; thus de-coupling network services from underlying network infrastructures [Chowdhury and Boutaba 2009]. Network virtualization can greatly enhance the flexibility, diversity, and manageability of network services in the future Internet; therefore may significantly improve networking performance for meeting the requirements of Cloud computing. More importantly, network virtualization enables the notion of virtualization, which is the technical foundation of Cloud computing, to be a key attribute of the next generation Internet; thus offering a promising basis for the composition of Cloud computing and networking systems. The Service-Oriented Architecture (SOA) is an effective architectural principle for heterogeneous system integration. Due to its loose-coupling interaction feature, SOA may greatly facilitate the separation of service provisioning and infrastructures in both networking and Cloud computing domains; thus serving as a key enabler for network-Cloud composition.

The research work presented in this article investigates the problem that how network virtualization facilitate integrating networking and Cloud computing into a composite service provisioning system. Specifically application of the SOA in network virtualization is first discussed and a SOA-based network virtualization paradigm is proposed. Then a service-oriented framework for network and Cloud composition is developed, in which network virtualization enables network infrastructures to be accessed as SOA-compliant services. This framework allows networking resources to be involved in Cloud computing as full participants just like computing resources such as CPU capacities and memory/disk space. Therefore, networking and Cloud computing systems can be integrated into an end-to-end service provisioning system to Cloud users through composition of network and Cloud services.

It is important for both Cloud service providers and users to obtain thorough understanding about the performance of composite network–Cloud service provisioning, which determines the actual service perception of Cloud users. Analytical modeling and analysis are necessary for achieving this objective. However currently available modeling and analysis techniques focus on either networking or computing systems; thus lacking the ability to analyze composite systems with both networking and computing functions. To tackle this challenge, a new model is proposed in this article for characterizing service capabilities of composite networking–Cloud computing systems. Based on this model, analysis techniques are developed for determining the worst-case service performance that can be guaranteed by composite network–Cloud service provisioning systems.

The rest of this article is organized as follows. Section 2 introduces the concept of network virtualization and discusses application SOA in network virtualization. Section 4 presents a service-oriented framework for network and Cloud services composition. Section 5 proposes a new approach to modeling service capabilities of composite network–Cloud service provisioning systems. Section 6 develops analysis techniques for determining the worst-case performance of composite network–Cloud services. Numerical examples are given in Section 7 to illustrate applications of the developed techniques. Section 8 draws conclusions.

2. NETWORK VIRTUALIZATION AND THE SOA

2.1 Network Virtualization

Network virtualization was first developed as an approach to building an open experimental network platform that allows researchers to create customized virtual networks for evaluating new network technologies and architecture [Anderson et al. 2005; Group 2006]. Since then the role of virtualization in the Internet has been shifted from an evaluation tool to a fundamental diversifying attribute of the internetworking paradigm [Feamster et al. 2007; Turner and Taylor 2005]. Essentially network virtualization follows a well-tested principle – separation of policy from mechanism – in the networking area. In this case, network service provisioning is separated from data transportation mechanisms; thus dividing the traditional role of Internet Service Providers (ISPs) into two entities: Infrastructure Providers (InPs) who manage the physical infrastructures, and Service Providers (SPs) who create virtual networks for offering end-to-end services by aggregating resources from (probably multiple) infrastructures [Chowdhury and Boutaba 2009].

InPs are in charge of operations and maintenance of physical network infrastructures and offer their resources to different service providers. SPs lease networking resources from multiple InPs to create virtual networks and deploy customized protocols in their virtual networks to offer services to end users. Each virtual network is composed and managed by a single SP, which synthesizes the networking resources allocated in underlying infrastructures. A virtual network is a collection of virtual nodes connected together by a set of virtual links to form a virtual topology, which is essentially a subset of the underlying physical topology. Each virtual node could be hosted on a particular physical node or could be a logical abstraction of a networking system. A virtual link spans over a path in the physical network and includes a portion of the networking resources along the path. Figure 1 illustrates a network virtualization scenario in which the service providers SP1 and SP2 construct two virtual networks by leasing resources from the infrastructure providers InP1 and InP2.

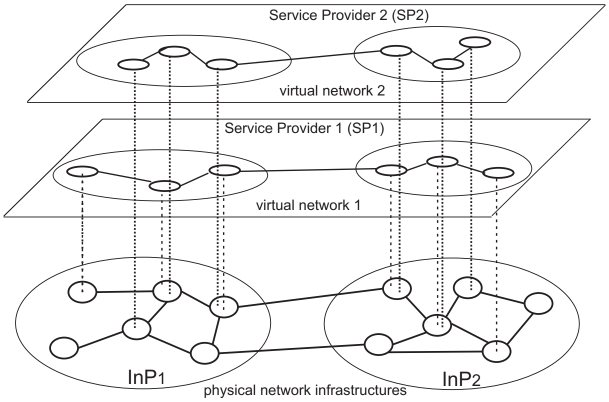

Figure. 1: Illustration of a network virtualization environment.

Network virtualization will bring a significant impact on service provisioning in the future Internet. By allowing multiple virtual networks to cohabit on a shared physical substrate, network virtualization provides flexibility, promotes diversity, and promises increased manageability in the Internet. A diversified Internet enabled by network virtualization offers a rich environment for innovations, thus stimulating the development and deployment of new Internet services. In such a environment, SPs are released from the requirement of purchasing, deploying, and maintaining physical network equipments, which significantly lowers the barrier to entry of the Internet service market. Network virtualization enables a single SP to obtain the control over network service delivery across infrastructures that belong to different network domains, which will greatly facilitate inter-domain network QoS provisioning.

2.2 The Service-Oriented Architecture

The SOA is a system architecture that provides an effective solution to coordinating computational resources across heterogeneous systems to support various application requirements. As described in [Channabasavaiah et al. 2003] the SOA is an architecture within which all functions are defined as independent services with invokable interfaces that can be called in defined sequences to form business processes. The SOA can be considered as a philosophy or paradigm to organize and utilize services and capabilities that may be under the control of different ownership domains [OASIS 2006]. Essentially the SOA enables virtualization of various computing resources in form of services and provides a flexible interaction mechanism among services.

A service in the SOA is a computing module that is self-contained (i.e., the service maintains its own states) and platform-independent (i.e., the interface to the service is independent with its implementation platform). Services can be described, published, located, orchestrated, and programmed through standard interfaces and messaging protocols. All services in the SOA are independent of each other and service operation is perceived as opaque by external services, which guarantees that external components neither know nor care how services perform their functions. The technologies providing the desired functionality of the service are hidden behind the service interface.

A key feature of SOA is the *loose-coupling* interaction among heterogeneous systems in the architecture. The term *coupling* indicates the degree of dependency any two systems have on each other. In a loosely coupled interaction, systems need not know how their partner systems behave or are implemented, which allows systems to connect and interact more freely. Therefore, loose coupling of heterogeneous systems provides a level of flexibility and interoperability that cannot be matched using traditional approaches for building highly integrated, cross-platform, inter-domain communication environments. It is this feature that makes the SOA a very effective architecture for coordinating heterogeneous systems to support various application requirements.

Though the SOA can be implemented with different technologies, Web services are the preferred environment for realizing the SOA promise of maximum service sharing, reuse, interoperability. Key Web service technologies include the technologies for service description, service publication, service discovery, service composition, and message delivery. The standard for web service description is Web Service Description Language (WSDL) [(W3C) 2007b], which defines the XML grammar for describing services as a collection of communicating endpoints capable of exchanging messages. Web service publication is achieved by Universal Description Discovery and Integration (UDDI) [OASIS 2005], which is a public directory with standard interfaces for publishing and searching service descriptions. Simple Object Access Protocol (SOAP) [(W3C) 2007a] is a XML-based messaging protocol on which web services rely to exchange information among them. Web service composition describes the execution logic of service-based functions by defining their control flows. The Business Process Execution Language for Web Services (BPEL4WS) [OASIS 2007] provides a standard for Web service composition.

3. APPLICATION OF SOA IN NETWORK VIRTUALIZATION

Application of the service-orientation idea in telecommunications and networking recently attracted attention of the research community. Some efforts in this area include Web services-based application program interface specified by Parlay X, the Open Service Environment (OSE) developed by Open Mobile Alliance (OMA) [(OMA) 2007], the optical network control architecture developed in UCLPv2 project [Grasa et al. 2007], and the network management system for the experiment platform in FEDERICA project [Szegedi et al. 2009]. Survey about applications of the SOA concept and Web service technologies in telecommunications can be found in [Magedanz et al. 2007; Griffin and Pesch 2007].

Separation between network infrastructures and network service provisioning is a key requirement of network virtualization, which allows resources provided by InPs to be accessed by SPs via standard and programmable interfaces. Therefore loose coupling interaction between InPs and SPs serves as a key enabler for network virtualization. The SOA is an effective architecture for coordinating resources in heterogeneous systems to support various application requirements. Essentially the same challenge, namely coordinating networking resources across heterogeneous network infrastructures for offering network services, is faced by network virtualization for the Internet. Therefore, the SOA offers an effective mechanism to support flexible interactions between InPs and SPs in network virtualization; thus may greatly facilitate network service provisioning in the Internet.

By following the SOA principle, networking resources and data transportation functions offered by network infrastructures can be encapsulated into *network infrastructure services* offered by InPs. Such infrastructure services can be published by InPs; and then discovered, selected, and accessed by SPs in order to build virtual networks for providing network services to end users. This is essentially the Infrastructure-as-a-Service (IaaS) paradigm applied in network virtualization environments. A SOA-based network virtualization paradigm is shown in Figure 2. The infrastructure platform consists of networking systems encapsulated into infrastructure services. Each InP compiles a description of its infrastructure service and publishes it at the service registry. By publishing an infrastructure service description, an InP can advertise the networking functions and capabilities of its infrastructure without exposing internal implementation details. The service broker searches and discovers appropriate infrastructure services for different SPs to build their virtual networks and offer network services to upper layer applications.

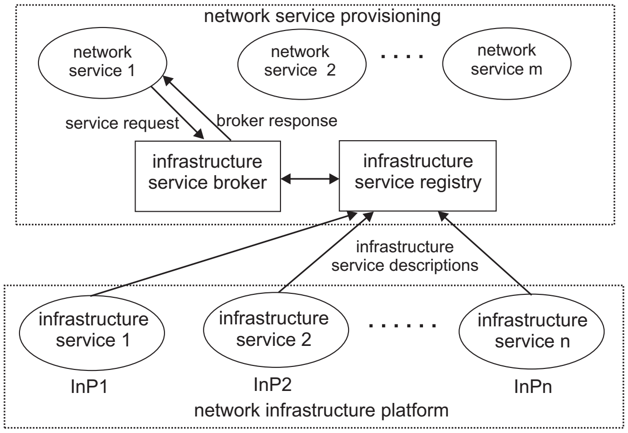

Figure. 2: A SOA-based network virtualization paradigm.

Applying the SOA principle in network virtualization ensures loose-coupling to be a key feature of the interactions between SPs and InPs. Such a network virtualization paradigm inherits the merit of SOA that enables flexible and effective collaborations across heterogeneous systems for providing services that meet diverse application requirements. SOA-based network virtualization also gives Internet service providers the ability to view their underlying infrastructures more as commodities and allows infrastructure development to become more consistent. This enables faster time to market as new network service initiatives can reuse existing services and components, thus reducing design, development, testing, and deployment time in addition to the cost and risk of new service development.

The SOA-based network virtualization paradigm provides a means to present abstracted networking capabilities to upper-layer software including Cloud computing systems. This allows for the use of networking capabilities without having to address the specific dependencies of certain types of low-level network protocols and hardware. Because of the heterogeneity of network protocols, equipments, and technologies, exposing networking capabilities to Cloud computing systems without virtualization would lead to unmanageable complexity. The abstraction of networking resources through service-oriented network virtualization can address the diversity and significantly reduce complexity of network and Cloud composition.

4. SERVICE-ORIENTED COMPOSITION OF CLOUD AND NETWORK SERVICES

In addition to serve as a key enabler for network virtualization, the SOA also forms a core element in the technical foundation for Cloud computing, especially for Cloud service provisioning. Recent research and development have been bridging the power of SOA and virtualization in the context of Cloud computing ecosystem [Zhang and Zhou 2009]. The Open Grid Forum (OGF) is working on the Open Cloud Computing Interface (OCCI) standard [OGF 2010], which defines SOA-compliant open interfaces for interacting with Cloud infrastructures. Taking a look at some of the most important Cloud providers, we can see that the SOA principle has strongly influenced Cloud service provisioning. For example Amazon, arguably the most well-known Cloud service provider, offers a complete ecosystem of Cloud infrastructure services including virtual machines (Elastic Compute Cloud EC2) and plain storage (Simple Storage Service S3). Amazon Cloud services are exposed by interfaces defined in WSDL and accessed through SOAP messaging.

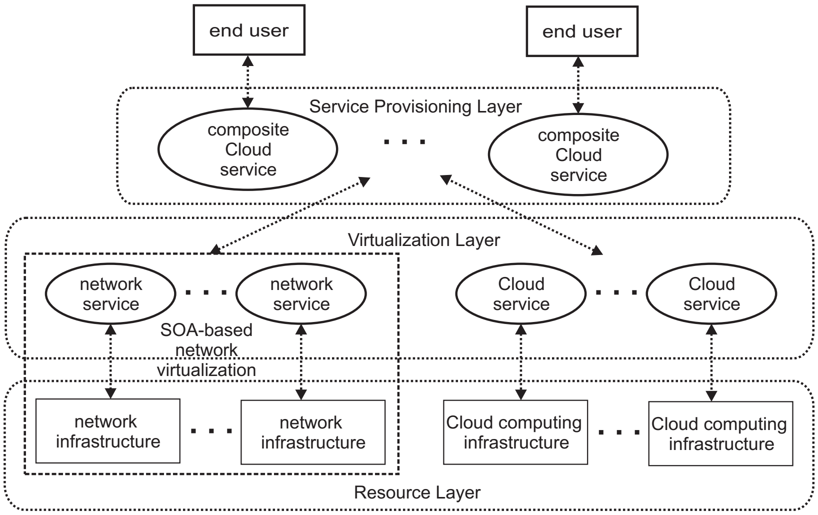

Figure. 3: The layered structure of a service-oriented framework for networking and Cloud computing composition.

Since the SOA plays a key role in both network virtualization and Cloud computing, it provides a promising mechanism for integrating networking and Cloud computing into a composite service provisioning system, which enables a holistic view of service provisioning across network and Cloud services. Figure 3 shows the layered structure of a service-oriented framework for composing networking and Cloud computing systems. At the bottom of this framework is the Resource Layer, which consists of physical infrastructures for both networking and Cloud computing. Above the Resource Layer is the Virtualization Layer. At this layer SOA-based network virtualization encapsulates resources in various network infrastructures into network services. Computational resources offered by Cloud infrastructure providers are also abstracted into Cloud services in this layer by following the SOA principle. The Service Provisioning Layer is above the Virtualization Layer. A key function of this layer is to discover and select both network services and Cloud services, and synthesizes them into composite network–Cloud services that match the requirements of end users. In this framework networking resources are virtualized, accessed, and managed through a unified mechanism as computational resources in Clouds, such as CPU capacities and memory/disk space. SOA-based network virtualization in this framework enables networking resources to be exposed as commodity service components and composed with Cloud computing resources into composite network-Cloud services. Therefore networking and Cloud computing systems are integrated into a composite service provisioning system for supporting the requirements of various applications.

SOA-based network virtualization greatly facilitates network and Cloud service composition. This new networking paradigm enables a much wider range of communication services with more attributes than what can be offered by traditional networking technologies. Networking resources, virtualized and encapsulated in SOA-compliant services, may be combined in almost limitless ways with other service components that abstract both computational and networking resources; thus greatly expanding the services that can be offered by composite network-Cloud systems. SOA-based network virtualization offers the ability to match Cloud requirements to communication services through discovering and selecting the appropriate network services and composing them with Cloud services. Through de-coupling network infrastructures from service provisioning, SOA-based network virtualization also allows new composite Cloud services to be developed and deployed without being limited by the evolution of underlying networking technologies.

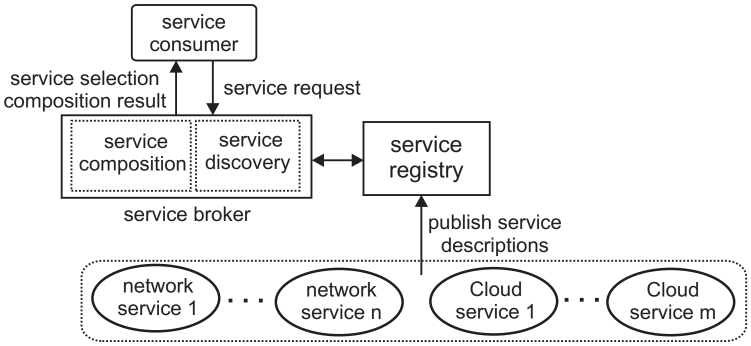

Figure. 4: A Web service-based broker system for network–Cloud service composition

Since Web services offer key technologies for realizing SOA, which is the core of this composite service provisioning framework, composite network–Cloud service delivery can also be implemented based on Web service technologies. The architecture of such a service delivery system is shown in Figure 4. In this system, both network service providers and Cloud service providers publish their service descriptions at a service registry. When a service consumer, typically is a computing application, needs to utilize a Cloud service, it sends a service request to the service broker. The service broker discovers available Cloud and network services by searching the registry, and then synthesizes the appropriate network and Cloud services into a composite service that meet the consumer's requirements. For example, multiple providers publish Cloud services like Amazon EC2, Amazon S3, Google App, etc., and network services such as AT&T Internet service and Verizon network service. When a user needs to send a data file to a Cloud for processing, it sends a request to the service broker and specifies the requirements for both data processing and transmission. The broker may select the Amazon EC2 for data processing and the Verizon network service for data transmission, and then compose these two services into one composite service that meets the requirements for both data processing capacity (offered by Amazon EC2) and data transmission bandwidth (provided by Verizon). The network service in this example provides communication between a Cloud service provider and a service consumer, instead of communications inside a Cloud infrastructure. This is the typical case considered in this article. It is worthwhile to notice that a Cloud infrastructure could be distributed across different geographic locations that are interconnected through data communication links. In such a case, a Cloud service offered based on such a distributed infrastructure is also a composite service consisting both networking and distributed computing functions.

Service-oriented composition of network and cloud services allows provisioning of network services and Cloud services, which used to be offered separately by different providers, merge into one layer of composite services provisioning. This convergence enables a new service delivery model in which the roles of traditional network service providers, like AT&T and Verizon, and computing service providers, such as Amazon and Google, merge together into one role of composite network–Cloud service provider. This new service delivery model may stimulate innovations in service development and create a wide variety of new business opportunities.

5. MODELING COMPOSITE NETWORK–CLOUD SERVICE PROVISIONING SYSTEMS

The performance of composite network–Cloud service provisioning determines the Cloud user's actual perception of service quality, which has a direct impact on the achievable performance of Cloud-based applications. Modeling and analysis helps us to obtain thorough understanding and deep insights about performance of composite network–Cloud services. Recently performance analysis on Cloud computing started attracting the research community's attention. A queueing model was developed in [Xiong and Perros 2009] for analyzing Cloud service performance. Performance analysis on a star-topology Cloud under divisible load was reported in [Ismail and Zhang 2010]. In [Chen and Li 2010] the authors proposed a queueing-based model for performance management in Clouds. However currently available modeling and analysis techniques mainly focus only on Cloud computing systems; thus lacking the ability to analyze composite systems with both networking and computing functions. Therefore, it becomes very important to develop new techniques for modeling and analyzing composite network–Cloud service provisioning systems.

Composition of networking and Cloud computing brings new challenges to system modeling and performance analysis. The main challenges come from the heterogeneity of service providers and the abstraction caused by resource virtualization. A composite network-Cloud service provisioning system consists of networking systems and Cloud infrastructures with diverse implementations. Therefore, the modeling and analysis techniques must be general and applicable to the heterogeneous networking and Cloud computing systems that may coexist in the composite service provisioning system. Both networking and Cloud computing systems are encapsulated into services through virtualization, which separates service provisioning from any specific communication and computing technologies. Therefore, the modeling and analysis techniques should be agnostic to specific networking and Cloud computing technologies. The rest of this article tackles the challenges for developing a model and analysis techniques that satisfy these requirements

A typical end-to-end provisioning system for composite network–Cloud services is shown in Figure 5, which consists of both a Cloud computing infrastructure that offers Cloud services and the communication network that provides the network service for accessing the Cloud service. If the end user wants a certain level of performance guarantee from the composite service provisioning, the end user must expect a certain level QoS from the providers of both network and Cloud services. In general, such QoS expectation can be defined in the Service Level Agreements (SLAs) between the user and service providers. The currently available commercial Cloud services such as Amazon Web Service do not include any explicit QoS guarantee in their SLAs, which leads to performance that cannot meet the requirements of high-performance computing applications [Garfinkel 2007; Jackson et al. 2010; Wang and Ng 2010]. Although the QoS expectation may vary due to the diversity of service providers, it typically includes a requirement on the minimum data transport rate for network services and the minimum processing capacity / maximal processing latency for Cloud services.

In order to analyze the composite service performance, one must understand the communication capability offered by the network service and the computing capability provided by the

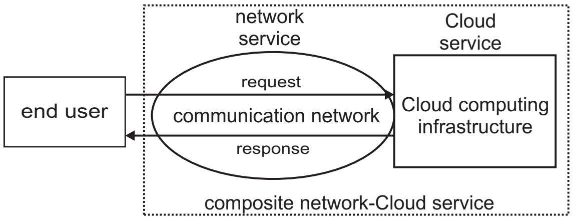

Figure. 5: A provisioning system for composite network-Cloud services.

Cloud service. The methodology taken in this article is to first develop a general capability profile that can model service capabilities of both network and Cloud services, and then compose the capability profiles of the two service components into one 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 service providers (including both network and Cloud service providers), should be independent of implementation technologies of the underlying networking and Cloud infrastructures, should also be easy to compose for modeling the composite service capability. In order to meet these requirements, the concept of *service curve* from *network calculus* theory [Boudec and Thiran 2001] is employed in this article for developing such a general and flexible capability profile.

A service component capability profile 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 convolution operation in min-plus algebra, which is defined as $x(t) \otimes y(t) = \inf_{s:0 \leq s \leq t} \{x(t-s) + y(s)\}$.

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

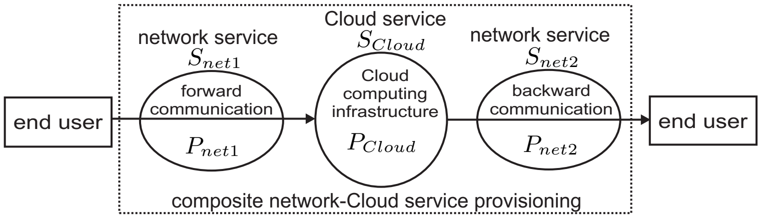

Figure. 6: Modeling composite network–Cloud service provisioning.

The service capacity of a composite network–Cloud provisioning system can be characterized by the model shown in Figure 6. Considering the general case that the two directions of data communications between the end user and the Cloud infrastructure take two network services, denoted as S_{net1} and S_{net2} respectively. The Cloud service component is denoted as S_{Cloud}. We assume that the forward network service (communication from user to Cloud), the Cloud service component, and the backward network service (communication from Cloud to user) respectively have the capability profiles $P_{net1}(t)$, $P_{Cloud}(t)$, and $P_{net2}(t)$. Assume each input message of the Cloud service component, after being processed in the Cloud, triggers one output message with the same length, then S_{net1}, S_{Cloud}, and S_{net2} form a tandem server series. 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 the convolution of the service curves of all these servers. Since the capability profile defined in (1) is essentially the service curve of a service component, the capability profile for the composite network–Cloud service provisioning system can be determined accordingly. Therefore, the capability profile for the composite service, denoted by $P_{Composite}(t)$, can be determined as,

$$P_{Composite}(t) = P_{net1}(t) \otimes P_{Cloud}(t) \otimes P_{net2}(t). \tag{2}$$

The capability profiles defined in (1) and (2) give a general approach to modeling composite network–Cloud service provisioning. In order to obtain a more tractable profile that can characterize typical network and Cloud service capabilities, this article defines a *Latency-Rate* (*LR*) profile for a service component as follows. If a service component S has a capability profile

$$\mathcal{P}[r, \theta] = \max\left\{0, r(t - \theta)\right\}, \tag{3}$$

then we say that the service component S has a *LR capability profile*, where 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 a certain amount data transport capacity (the minimum bandwidth) to a service user. Such a minimum bandwidth guarantee is described by the rate parameter r in a *LR* profile. Data communication in a network infrastructure also experiences a fixed delay that is independent with 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 Cloud computing systems. Cloud service providers typically offer a certain amount of service capacity. For example according to Amazon, 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 250Mb/s regardless of instance type. The latency and rate parameters of a *LR* profile for the Cloud service can be derived from the I/O bandwidth and processing capacity information specified by a Cloud service provider.

Suppose each service component in the composite network–Cloud system has a *LR* profile,

$$P_{net1} = \mathcal{P}[r_1, \theta_1], \quad P_{Cloud} = \mathcal{P}[r_C, \theta_C], \quad P_{net2} = \mathcal{P}[r_2, \theta_2],$$

it can be proved that the capability profile of the composite service provisioning system is

$$P_{Composite} = \mathcal{P}[r_1, \theta_1] \otimes \mathcal{P}[r_C, \theta_C] \otimes \mathcal{P}[r_2, \theta_2] = \mathcal{P}[r_e, \theta_e] \tag{4}$$

where $r_e = \min\{r_1, r_C, r_2\}$ and $\theta_e = \theta_1 + \theta_C + \theta_2$.

Equation (4) implies that if each network and Cloud service component in an composite service provisioning system can be modeled by a *LR* profile, then the service capability of the entire provisioning system can also be modeled by a *LR* profile. The latency parameter of the composite *LR* profile is the summation of latency parameters of all service components in the system, and the service rate parameter of the composite profile is the minimum service rate of all the service components.

6. PERFORMANCE ANALYSIS FOR COMPOSITE NETWORK–CLOUD SERVICES

Based on the profile-based service capability model presented in the preceding section, analysis techniques will be developed in this section for determining the worst-case performance that can be provided by composite network–Cloud services to end users. The analysis in this section focuses on the minimum data throughput and the maximum response delay (the delay between time instants when a user sends out a request to the Cloud and when the user receives the corresponding response from the Cloud), which are the performance parameters that are most significant to computing applications developed based on Cloud services.

A service provisioning system with a certain amount of service capacity may achieve different levels of performance under different loads. Therefore it is necessary to develop an approach to characterizing the traffic that end users load onto a composite network–cloud service provisioning system. Due to the highly diverse applications that will be supported by Cloud computing, a general profile that can describe traffic loads of various applications is required. Since the entry of a composite service system where user's applications load the system is at the boundary of the forward networking system, the traffic load for the forward network service is the load of the entire service provisioning system. Therefore, the concept of *arrival curve* is adopted from network calculus theory as a general load profile, which is defined as follows.

Let $R(t)$ denote the accumulated amount of traffic that arrives at the entry of a composite network–Cloud service provisioning system by 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 $0 < s < t$

$$R(t) - R(s) \leq \mathcal{L}(t - s). \tag{5}$$

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 the traffic loaded by various computing applications onto a service system.

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 $\mathcal{L}[p, \rho, \sigma] = \min\{pt, \sigma + \rho t\}$, where p, ρ, and σ are respectively called the peak rate, sustained rate, and maximal burst size for the traffic.

The capability profile of a composite network–Cloud service provisioning system gives the lower bound of the amount of service offered by the system, which essentially gives the minimum throughput guaranteed by the service system to its user. Therefore, given the capability profile $P(t)$ of an composite service provisioning system, the minimum throughput performance, denoted as $\mathcal{T}_{min}$, can be determined as

$$\mathcal{T}_{min} = \lim_{t \to \infty} [P(t)/t]. \tag{6}$$

The maximum response delay performance is associated with both the guaranteed service capacity of the system, which is modeled by the capability profile, and the characteristic of the traffic load of the system, which is described by a load profile. It can be shown by following network calculus that for a service system with a capability profile $P(t)$ under traffic described by a load profile $\mathcal{L}(t)$, the maximum delay d_{e2e} guaranteed by the system to its end user can be determined as,

$$d_{e2e} = \max_{t:t \geq 0} \{\min\{\delta : \delta \geq 0 \;\; \mathcal{L}(t) \leq P(t + \delta)\}\}. \tag{7}$$

Since the LR profile can model typical network and Cloud service capabilities and leaky bucket traffic regulator is widely deployed at the entries of QoS-enabled networks, the rest of this section focuses on analyzing end-to-end response delay of composite network–Cloud service systems with a LR capability profile and a leaky bucket load profile. Suppose the capability profile of the composite service system is

$$P_{Composite} = \mathcal{P}[r_e, \theta_e] = \max\{0_e, r_e(t - \theta_e)\},$$

then following (6) the minimum throughput performance of the composite network–Cloud service is

$$\mathcal{T}_{min} = \lim_{t \to \infty} \frac{r_e(t - \theta_e)}{t} = r_e. \tag{8}$$

Known from (4) that $r_e = \min\{r_1, r_C, r_2\}$, and the rates r_1 and r_2 in network service profiles represent the amount of bandwidth provided by the forward and backward network services;

therefore equation (8) implies that the minimum throughput performance of a composite network–Cloud service is limited by either network bandwidth or Cloud I/O bandwidth, whichever is less.

The maximum response delay guaranteed by this composite network–Cloud service under a load profile $\mathcal{L}(p, \rho, \sigma)$ can be determined by following (7) as

$$d_{e2e} = \theta_e + \left(\frac{p}{r_e} - 1\right)\frac{\sigma}{p - \rho} = \theta_\Sigma + d_C + \left(\frac{p}{r_e} - 1\right)\frac{\sigma}{p - \rho}$$

where $\theta_\Sigma = \theta_1 + \theta_2$, which is the round-trip communication latency of network services; and d_C is the computing latency of the Cloud service.

Network service latency parameter can be estimated as follows. The latency of a LR profile for a network service reflects a system property of the underlying network infrastructure that may be seen as the worst-case delay experienced by the first traffic bit in a busy period of a networking session through the infrastructure. Therefore the main elements of the latency includes link transmission delay and network processing delay for a packet, if signal propagation delay is assumed to be ignorable. Then the latency can be estimated as $\theta = L/r + L/R$, where L and R are respectively the maximum packet length and maximum link rate of the network infrastructure. Suppose the forward and backward networks services of the composite network–Cloud service provisioning system have identical link rate R and packet length L, then the maximum response delay performance of the composite service becomes

$$d_{e2e} = 2L\left(\frac{1}{r} + \frac{1}{R}\right) + d_C + \left(\frac{p}{r_e} - 1\right)\frac{\sigma}{p - \rho}.$$

7. NUMERICAL EXAMPLES

This section gives numerical examples for illustrating applications of the developed techniques for analyzing composite network–Cloud service performance. The composite service provisioning system considered in this section is shown in Figure 5, in which a Cloud user accesses a Cloud computing infrastructure through a data communication network. The service received by the user is a composite service consisting of network services provided by the communication network and a Cloud service offered by the Cloud computing infrastructure. Data transmission for request messages from the user to the Cloud and for response message from the Cloud back to the user are supported by the same network; therefore the forward and backward network service have an identical capability profile. We assume that each of the network service and the Cloud service has a LR profile. Based on the measurement results obtained in [Wang and Ng 2010; Jackson et al. 2010], the latency parameter of the Cloud service profile is assumed to be 150 μs. Traffic parameters of the load profile are 320 Mb/s, 120 Mb/s, and 200 kbits for the peak rate, sustained rate, and burst size respectively.

This section first examines the service provisioning scenario that the user accesses the Cloud infrastructure through a high-speed network with a link transmission rate up to 10 Gb/s. The maximum packet length in such a network is assumed to be 1500 bytes, which is the Maximum Transmit Unit (MTU) size for Ethernet networks. The maximum response delay performance of the composite network-Cloud service, denoted as d_{e2e}, is determined with different amounts of data transport capacity (bandwidth) offered by the network service. The obtained results are given in Table 1 and plotted in Figure 7. Table 1 also gives the ratio of the total response delay over the Cloud computing latency, i.e., d_{e2e}/d_C.

bandwidth (Mb/s)	d_{e2e} (μs)	d_{e2e}/d_C
125	500	3.33
175	372	2.48
225	301	2.01
275	256	1.70
325	212	1.41
375	203	1.35
425	197	1.31
475	192	1.28

Table 1 The maximum delay performance of a composite network–Cloud service with a high-speed network.

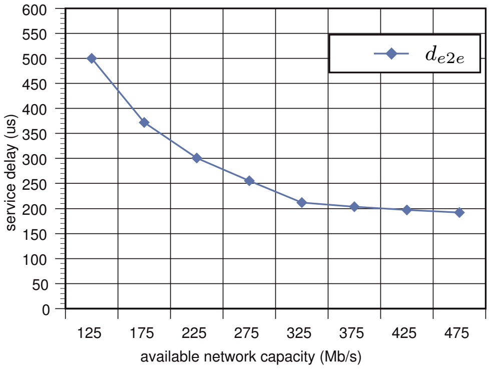

Figure. 7: The maximum delay performance of a composite network–Cloud service with a high-speed network.

Table 1 and Figure 7 show that the delay performance of the composite service first decreases significantly with the increment of available network capacity then becomes flat after the capacity value is greater than a threshold (325 Mb/s in this example). This shows that in this service scenario when data transport capacity offered by the network service is less than the peak load rate (320 Mb/s), networking system forms the bottleneck of composite service provisioning and data communications contribute a significant part of the total response delay. From the d_{e2e}/d_C ratio given in Table 1 we can see that the total response delay is between 3.3 and 1.7 times of Cloud computing latency when bandwidth is less than the peak load rate. In this case leasing more bandwidth from the network service provider can significantly improve delay performance of the composite service. The flat part of the delay curve shows that when the network service offers a data transport capacity that is higher the peak load rate, latency of the Cloud service becomes the major part of the total service delay. The d_{e2e}/d_C ratio listed in Table 1 shows that communication delay is just about 20% to 30% of the total response delay when bandwidth is greater than the load peak rate. Therefore in this case leasing more bandwidth from network service provider has a minor impact on improving delay performance of composite service provisioning.

bandwidth (Mb/s)	d_{e2e} (μs)	d_{e2e}/d_C
125	578	3.85
150	503	3.36
175	450	3.0
200	410	2.73
225	379	2.52
250	354	2.36
275	334	2.22
300	317	2.11

Table 2 The maximum delay performance of a composite network–Cloud service with a moderate speed network.

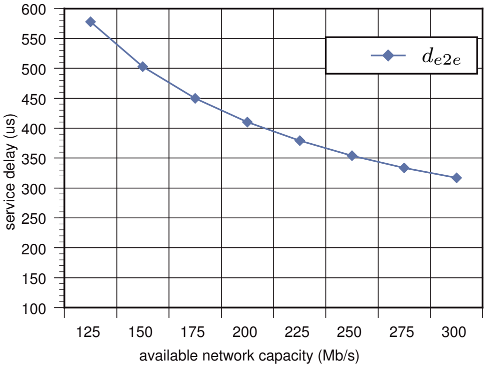

Figure. 8: The maximum delay performance of a composite network–Cloud service with a moderate speed network.

The other service scenario analyzed in this section is that the user accesses the Cloud infrastructure through a network with a moderate link rate up to 300 Mb/s. The results for the maximum response delay performance d_{e2e} with different amounts of network service capacity are given in Table 2 and plotted in Figure 8. Table 2 and Figure 8 show that the service delay performance for this scenario always decreases with the increment of network capacity. This is because the network service provider in this scenario can only offer moderate service rate up to 300 Mb/s, which is less than the peak load rate. In this case the delay performance of composite service provisioning is mainly limited by the networking system; therefore, leasing more bandwidth from the network service provider may significantly improve service delay performance. The d_{e2e}/d_C ratio given in Table 2 shows that the total response delay perceived by the user is more than 2 times (up to nearly 4 times) of Cloud computing latency. This implies that when a user accesses a Cloud infrastructure from a bandwidth constrained networking environment, such as a wireless network or a cellular communication system, the networking system may have a much bigger impact on the user's perception of service performance than the Cloud computing infrastructure does.

8. CONCLUSIONS

Computer networks play a crucial role in Cloud service provisioning and network QoS has a significant impact on Cloud service performance. Therefore networking and Clouding computing systems should be modeled and analyzed as a composite service provisioning system in order to obtain thorough understanding about the user's perception of Cloud service performance. Network virtualization, which is expected to become a key attribute in the next generation Internet, offers a promising approach to integrating networking and Cloud computing systems for composite service provisioning. The SOA, as an effective architectural principle for heterogeneous system integration, may serve as a key enabler for network and Cloud service composition. The research presented in this article investigated application of the SOA in network virtualization for composing network and Cloud services, and studied modeling and performance analysis on network virtualization for composite network–Cloud service provisioning. The main contributions of this article include a SOA-based network virtualization paradigm, a service-oriented framework for composing network and Cloud services, a mew approach to modeling service capabilities of composite network–Cloud service provisioning systems, and analysis techniques for evaluating performance of composite network–Cloud services. The modeling and analysis techniques developed in this article are general and independent of any specific network and Cloud implementation technology; thus are applicable to various heterogeneous networking and Cloud computing systems.

REFERENCES

Anderson, T., Peterson, L., Shenker, S., and Turner, J. 2005. Overcoming the Internet impasse through virtualization. *IEEE Computer 38,* 4, 34–41.

Boudec, J. L. and Thiran, P. 2001. *Network calculus: a theory of deterministic queueing systems for the Internet.* Springer Verlag.

Channabasavaiah, K., Holley, K., and Tuggle, E. 2003. Migrating to a Service-Oriented Architecture. *IMB DeveloperWorks*.

Chen, H. and Li, S. 2010. Q queueing-based model for performance management on Cloud. In *Proc. of the 6th Intl. Conference on Advanced Information Management and Service.*

Chowdhury, N. M. M. K. and Boutaba, R. 2009. Network virtualization: state of the art and research challenges. *IEEE Communications 47,* 7, 20–26.

Feamster, N., Gao, L., and Rexford, J. 2007. How to lease the Internet in your spare time. *ACM SIGCOMM Computer Communications Review 37,* 1, 61–64.

Foster, I., Zhao, Y., Raicu, I., and Lu, S. 2008. Cloud computing and Grid computing 360-degress compared. In *Proceedings of the 2008 Grid Computing Environment Workshop.*

Garfinkel, S. L. 2007. An evaluation of Amazon's Grid computing services: EC2, S3, and SQS. In *Computer Science Group Technical Report TR-08-07, Harvard University.*

Grasa, E., Junyent, G., Figuerola, S., Lopez, A., and Savoie, M. 2007. Uclpv2: A network virtualization framework built on web services. *IEEE Communications 46,* 3, 126–134.

Griffin, D. and Pesch, D. 2007. A survey on web services in telecommunications. *IEEE Communications 45,* 7, 28–35.

Group, G. P. 2006. GENI design principles. *IEEE Computer 39,* 9, 102–105.

Ismail, L. and Zhang, L. 2010. Service performance and analysis in Cloud computing. In *Proc. of the 4th Eruopean Modeling Symposium on Computer Modeling and Simulation.*

Jackson, K. R., Muriki, K., Canon, S., Cholia, S., and Shalf, J. 2010. Performance analysis of high performance computing applications on the Amazon web services Cloud. In *Proceedings of CLOUDCOM'2010.*

Magedanz, T., Blum, N., and Dutkowski, S. 2007. Evolution of soa concepts in telecommunications. *IEEE Computer 40,* 11, 46–50.

OASIS. 2005. Universal Description, Discovery and Integration (UDDI) version 3.0.2.

OASIS. 2006. Reference Model for the Service-Oriented Architecture version 1.0.

OASIS. 2007. Business Process Execution Language for Web Services (BPEL-WS) version 1.1.

OGF, O. G. F. 2010. Open Cloud Computing Interface.

(OMA), T. O. M. A. 2007. OMA Service Environment Architecture.

Szegedi, P., Figuerola, S., Campanella, M., Maglaris, V., and Cervello-Pastor, C. 2009. With evolution for revolution: Managing federica for future internet research. *IEEE Communications 47,* 7, 34–39.

Turner, J. and Taylor, D. E. 2005. Diversifying the Internet. In *Proc. of IEEE Global Communication Conference.*

(W3C), W. W. W. C. 2007a. Simple Object Access Protocol (SOAP) version 1.2.

(W3C), W. W. W. C. 2007b. Web Service Description Language (WSDL) version 2.0.

Wang, G. and Ng, T. S. E. 2010. The impact of virtualization on network performance of Amazon EC2 data center. In *Proceedings of INFOCOM'2010.*

Xiong, K. and Perros, H. 2009. Service performance and analysis in Cloud computing. In *Proc. of the IEEE 2009 Congress on Services.*

Zhang, L.-J. and Zhou, Q. 2009. CCOA: Cloud computing open architecture. In *Proc. of the 1st Symposium on Network System Design and Implementation (NSDI '09)*.

Dr. Qiang Duan is an Assistant Professor of Information Science and Technology at The Pennsylvania State University Abington College. His research interests include data communications, computer networking, the next generation Internet, Cloud computing, and Web services. He has published four book chapters, fifteen journal articles, and more than thirty conference papers in these areas. Dr. Duan is serving on the editorial boards of Journal of Network Protocols and Algorithms and International Journal of Internet and Distributed Computing Systems. He has also served on the technical program committees for numerous conferences and as a reviewer for various journals. Dr. Duan received his Ph.D. degree in electrical engineering from the University of Mississippi in 2003. He holds a B.S. degree in electrical and computer engineering and a M.S. degree in telecommunications and electronic systems.