Analysis on Quality of Service Provisioning for Communication Services in Network Virtualization

Qiang Duan
Information Science & Technology Department
The Pennsylvania State University
Abington, PA USA
Email: qduan@psu.edu

***Abstract*— The Internet is currently facing many new challenges that require fundamental changes in network architecture and service models. However the current Internet lacks the flexibility to adopt innovations in network architecture and service provisioning. To fend off this ossification, network virtualization has been propounded as a key attribute of the future inter-networking paradigm and is expected to play a crucial role in the next generation Internet. Quality of Service (QoS) provisioning for end-to-end communication services across heterogeneous network infrastructures is one of the key technical issues in network virtualization. The Service-Oriented Architecture (SOA) provides a promising approach to tackling this challenging issue. The research presented in this article investigates application of SOA in network virtualization to facilitate QoS provisioning of communication services in the future Internet. This article proposes an analytical model for SOA-based service delivery in network virtualization, and develops analysis techniques for performance evaluation and resource allocation for service provisioning in network virtualization. Resource utilization for end-to-end QoS provisioning in network virtualization is also analyzed and compared with that of the conventional inter-domain QoS mechanism in the current Internet. The modeling approach and analysis techniques developed in this article are general and applicable to the heterogeneous networking systems in the future Internet for supporting various communication services.**

***Index Terms*— Network virtualization, communication services, Quality of Service (QoS), Service-Oriented Architecture (SOA), the next generation Internet.**

I. INTRODUCTION

In a relatively short period of time the Internet has become a critical infrastructure for global commerce, media, and defense. Due to its stunning success the current Internet is facing many new challenges that require fundamental changes in network architecture and service models. Various networking systems with different architecture and implementation technologies are expected to coexist and collaborate in the future Internet in order to support highly diverse communication services and networking applications. The current Internet architecture, which lacks the œexibility for supporting collaboration across heterogeneous network domains, must be changed to meet the requirements of the next generation networking. However the signiŒcant investments in current network infrastructures together with the end-to-end design principle of IP protocol create a barrier to any disruptive innovation in Internet architecture and service models. To mitigate the ossifying force in current Internet, network virtualization has been propounded as a key architectural attribute for the future inter-networking paradigm [1] and is expected to play a crucial role in the next generation Internet.

Network virtualization decouples service provisioning functions from data transportation mechanisms; thus separating the role of traditional Internet Service Providers (ISPs) into two independent entities: infrastructure providers (InPs) and service providers (SPs). InPs deploy and operate physical network infrastructures and provide networking resources to different SPs. SPs offer communication services to end users by creating virtual networks through synthesizing networking resources obtained from InPs. A virtual network consists of virtual nodes connected by virtual links. Each virtual node could be hosted on a physical network node or could be a logical abstraction of a networking system. A virtual link utilizes a portion of resources on a physical path in network infrastructures. Figure 1 illustrates a network virtualization environment, in which a service provider constructs a virtual network by accessing networking resources from two infrastructures providers InP1 and InP2.

Network virtualization brings a signiŒcant impact on communication service provisioning in the Internet. The current Internet basically offers a best-effort commodity service that gives service providers limited opportunities to distinguish themselves from competitors. A diversiŒed Internet enabled by network virtualization provides a rich environment that motivates and facilitates developments of new Internet services. Network virtualization enables a single SP to obtain control over the entire *end-to-end* service delivery path across network infrastructures that belong to different domains, which may greatly enhance end-to-end QoS provisioning for various communication services.

Manuscript received December 15, 2010; revised June 3, 2011; accepted August 10, 2011.

doi:10.4304/jcm.7.2.143-154

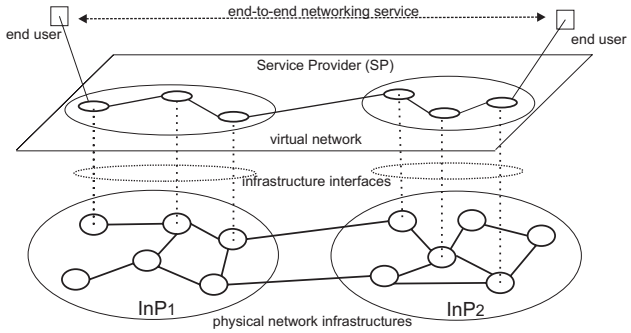

Figure 1. A virtual network and two network infrastructures in a network virtualization environment.

Recently network virtualization has attracted extensive research interest from both academia and industry. A new network architecture was proposed in [2] for diversifying the Internet, which enables various meta-networks built on top of a physical substrate. The CABO Internet architecture proposed in [3] decouples network service providers and infrastructure providers to support multiple network architectures over shared infrastructures. The GENI project sponsored by U.S. NSF employs network virtualization to build an open experimental facility for evaluating new network architectures [4]. The FP7 4WARD project sponsored by European Union has also adopted network virtualization as a key technology to allow the future Internet to run virtual networks in parallel [5]. Some standard organizations are also embracing the notion of network virtualization into their specifications. For example the Next Generation Network (NGN) architecture defined by ITU-T follows a key principle of separating service-related functions from underlying transport-related technologies [6]. An overview of more research efforts and progresses in the area of network virtualization can be found in [7].

Though many progresses have been made toward network virtualization in the Internet, this research field is still in its early stages. One of the important opening problems in this area is QoS provisioning for end-to-end communication services across the heterogeneous network infrastructures coexisting in the future Internet. A framework for end-to-end service differentiation is proposed in the AGAVE project [8], which is based on the concepts of *Parallel Internets* upon *Network Planes* of individual IP network providers. An end-to-end QoS architecture is developed in the EuQoS project for delivering different classes of service across heterogeneous network domains [9]. The QoS mechanisms proposed in both AGAVE and EuQoS projects follow the theme of DiffServ; therefore end-to-end QoS provisioning requires a universal agreement on meta-QoS-classes among all involved network domains and service providers. However, such a requirement is not realistic to network virtualization in the future Internet due to the heterogeneity of coexisting network infrastructures.

In network virtualization environments, SPs offer end-to-end communication services by allocating, synthesizing, and utilizing networking resources provided by multiple network infrastructures. These infrastructures may have various network architecture and implementation technologies, and may belong to different administration domains with diverse management policies. Therefore, interactions between SPs and underlying InPs to enable coordination across heterogeneous network infrastructures for QoS provisioning of communication services become a challenging issue in network virtualization.

The Service-Oriented Architecture (SOA), which is typically realized through Web services technologies, provides an effective architectural principle for coordinating heterogeneous systems to support highly diverse application requirements. 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) [10], the optical network control architecture developed in UCLPv2 project [11], and the network management system for the experiment platform in FEDERICA project [12]. Survey about applications of the SOA concept and Web service technologies in telecommunications can be found in [13], [14]. Applying SOA in network virtualization offers a promising approach that may greatly facilitate end-to-end communication service provisioning in the next generation Internet, which is the focus of study in this article.

In previous works the author analyzed bandwidth and delay performance for end-to-end service delivery in general network virtualization environments [15], [16]. The research presented in this article particularly investigates application of SOA in network virtualization with an end-to-end vision from a service provider s perspective. The study emphasizes end-to-end QoS guarantee as well as resource utilization for service provisioning by analyzing the minimum amount of resources an SP must expect from InPs for meeting an end-to-end QoS requirement. The main contributions include an analytical model and performance analysis technique for SOA-based service delivery in network virtualization, an approach to determining the amounts of bandwidth that an SP must acquire from InPs for QoS guarantee, and analysis on resource utilization for end-to-end QoS provisioning in network virtualization and comparison with the conventional inter-domain QoS mechanism in the current Internet.

The rest of this article is organized as follows. Section II discusses application of SOA in network virtualization and describes a SOA-based delivery system for communication services. An analytical model for SOA-based service delivery is proposed in Section III and end-to-end service performance is also evaluated in this section. Section IV develops a technique for allocating networking resources for end-to-end QoS guarantee. In Section V resource utilization for QoS provisioning in network virtualization is analyzed and compared with

that of the conventional inter-domain QoS mechanism. Numerical examples are given in Section VI to illustrate applications of the developed techniques and discuss the obtained insights. Section VII draws conclusions.

II. The SOA-Based End-to-End Delivery for Communication Services in Network Virtualization

The SOA is a system architecture initially developed by the distributed computing community as an effective solution to coordinating computational resources in multiple heterogeneous systems to support various application requirements. The SOA as described in [17] 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. 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. 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 [18]. Essentially the SOA enables virtualization of various computing resources in form of services and provides a flexible interaction mechanism among services.

Though the SOA can be implemented with different technologies, Web services provide a preferred environment for realizing the SOA promise of maximum service sharing, reuse, and interoperability. Key Web services technologies for implementing the SOA include service description, service publication, service discovery, and service composition. The related Web services specifications are Web Service Description Language (WSDL) [19], Universal Description Discovery and Integration (UDDI) [20], and Business Process Execution Language for Web Services (BPEL4WS) [21].

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 exchange, 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.

Essentially the same challenge, namely coordinating heterogeneous networking systems for providing communication services that meet highly diverse application requirements, is faced by network virtualization in the Internet. Therefore, applying the SOA principles in network virtualization may greatly facilitate end-to-end provisioning of communication services in future Internet. In a SOA-based network virtualization environment, the networking resources and capabilities of each network infrastructure are abstracted into a SOA-compliant *infrastructure service* that can be offered by the InP to SPs. Then end-to-end communication services can be constructed by SPs through composing and accessing infrastructure services.

In the virtualization-based future Internet, an end-to-end communication service delivery system is constructed by an SP through synthesizing the resources acquired from multiple network infrastructures. A key attribute of SOA-based network virtualization is the de-coupling between communication service provisioning and underlying network infrastructures. That is, an SP is able to access shares of network resources from multiple InPs without knowing the implementation details of these infrastructures. An SP leases or purchases physical resource from individual InPs via bilateral service contracts on the infrastructure services offered by the InPs. Through such an infrastructure-as-a-service paradigm, networking capabilities offered by different InPs can be encapsulated into a set of service components. The SP then assembles this set of service components to form an end-to-end communication service that meets the application requirement. Therefore, a SOA-based delivery system for communication services in network virtualization consists of a series of tandem service components, each of which is a logical abstraction of the infrastructure service provided by an InP. Figure 2 shows an end-to-end service delivery system constructed on top of n network infrastructures, denoted as $I_i, i = 1, 2, \cdots, n$. The infrastructure service provided by the i-th InP I_i to the end-to-end delivery system is virtualized as the service component S_i.

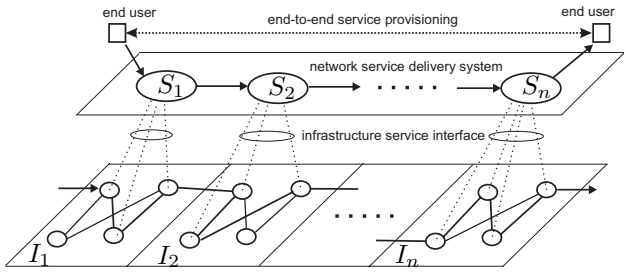

Figure 2. SOA-based end-to-end delivery for communication services in network virtualization.

Application of the SOA principle in network virtualization makes loose-coupling a key feature of both SP-InP interaction and collaboration among heterogeneous network infrastructures. Therefore, such a network virtualization paradigm inherits the merit of SOA that enables flexible and effective collaboration 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 in-

frastructure development to become more consistent. This enables faster time to market as new service initiatives can reuse existing services and components, thus reducing design, development, testing, and deployment time and cost.

It is interesting to notice that separation of network service provisioning from data transportation can be tracked back to early 90s in the Telecommunication Information Networking Architecture (TINA) [22]. TINA offers a management and control architecture of telecommunication network for supporting service delivery on top of a technology-independent transportation platform. The major goal of TINA was to overcome the limitation of service provisioning concepts in telecommunication networks, particularly in Intelligent Network (IN) and Telecommunications Management Network (TMN). The objective of network virtualization is to enable concurrent Internet architectures upon a shared physical substrate. Though TINA share a similar vision on separating service and platform and network resource abstraction with network virtualization, its realization was limited by the state-of-the-art of computing technologies in 90s, which were mainly distributed object-oriented computing technologies such as CORBA. Such distributed computing technologies are essentially tightly coupled. Therefore they lack an effective mechanism that fully supports network resource abstraction; thus having limited abilities to realize the vision of decoupling service and platform. Due to the loose-coupling feature of SOA that great facilitates resource abstraction and heterogeneous system collaboration, SOA-based network virtualization may fully realize the notion of separating service provisioning from network infrastructures with a greater level of œexibility and scalability than TINA.

In order to achieve QoS provisioning for end-to-end communication services in network virtualization environments, an SP must require a certain level of service guarantee from each of its underlying network infrastructure. In general, the quality of infrastructure service offered by an InP to the SP is deGned by a bilateral service contract. Although the contents of such a service contract may vary due to the diversity of infrastructures and service providers, it typically includes a requirement on data transport capability, such as the minimum amount of bandwidth that the SP can expect from the InP. That is, each service component in the service delivery system offers a certain level of networking capability provisioning, so that the entire delivery system can provide QoS guarantee for the end-to-end communication service.

In order to offer QoS provisioning for communication services and generate enough revenues as well, SPs in network virtualization environments must acquire sufGcient networking resources from InPs while at the same time minimize the cost of leasing infrastructure capacities. Therefore, it is signiGcant for service providers to obtain thorough understanding and deep insights about communication service performance and resource allocation for end-to-end QoS provisioning. Analytical modeling and analysis provide an effective approach to achieving this objective, and will be the focus of the study reported in the rest of this paper.

III. Modeling and Performance Analysis for End-to-End Communication Services in Network Virtualization

Network virtualization introduces some special challenges to modeling and analyzing end-to-end service provisioning. Traditional network modeling approaches were usually developed for particular network architecture. However, end-to-end services in network virtualization environments are delivered by virtual networks constructed upon infrastructures that may have different architecture and implementations. Most of the available service performance analysis techniques are based on assumptions about speciGc networking technologies, such as the data forwarding mechanisms, trafGc control schemes, and packet scheduling algorithms. However InPs in network virtualization environments may not want to make such information about their network internal details available to SPs. Therefore, the diversity of network infrastructures, de-coupling between SPs and InPs, and resource abstraction in network virtualization require a general and œexible modeling and analysis approach that is agnostic to the architecture and implementations in underlying network infrastructures.

Applying SOA in network virtualization enables SPs to access networking resources in various InPs through an infrastructure-as-a-service paradigm; thus constructing an end-to-end service delivery system by composing a series of service components. The layer of abstraction represented by service components in SOA-based service delivery offers an approach to simplifying the development of general modeling and analysis techniques that are applicable to heterogeneous network infrastructures. The methodology taken in this paper is to GrstGdevelop a proGle for modeling the networking capabilities of individual service components in a service delivery system, then compose the capability proGles of all the service components into one proGle that models the end-to-end service provisioning capability. A capability proGle gives a lower bound of the networking capacity that an SP can expect from an InP according to the infrastructure service contract between them, thus is independent with the implementation of the modeled infrastructure service. In order to develop a general capability proGle that is applicable to various network infrastructures, the notion of *service curve* from network calculus theory [23] is applied in this paper.

Let $R(t)$ and $E(t)$ respectively be the accumulated amount of trafGc that arrives at and departs from a service component by time t. Given a non-negative, non-decreasing function, $S(\cdot)$, we say that the service component guarantees a capability proGle $S(\cdot)$, if for any $t \geq 0$ in the busy period of the service component,

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

where the operator $\otimes$ denotes the convolution operation in min-plus algebra, defined as $h(t) \otimes x(t) = \inf_{s:0\leq s\leq t} \{h(t-s) + x(s)\}$.

A capability profile gives the minimum amount of networking capacity guaranteed by a service component, which is an abstraction of the infrastructure service provided by an InP to the service delivery system. Such a profile is a general function of time that specifies data transport capacity through the relation between arrival and departure traffic at a service component. Therefore a capability profile is independent of network architecture and implementations, thus is applicable to various heterogeneous network infrastructures.

This paper defines a Latency-Rate (LR) profile as a more tractable service capability description. If a service component S has a capability profile

$$S(t) = \mathcal{L}[r, \theta] = \max\{0, r(t-\theta)\}, \tag{2}$$

then we say that S has an LR profile, where the θ and r are respectively called the latency and rate parameters of the profile. an LR profile can serve as the capability model for a wide range of network infrastructures. In order to offer end-to-end QoS, an SP expects each underlying InP to guarantee a minimum amount of available bandwidth. Such a minimum bandwidth guarantee is described by the rate parameter r in an LR profile. In addition to available bandwidth, data transportation in a network infrastructure is also limited by a latency introduced by factors such as signal propagation delay, link transmission delay, and router/switch processing delay. This part of latency is a system property of an 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. The latency parameter θ in an LR profile is to characterize this aspect of the service capability offered by an InP to an end-to-end service delivery system.

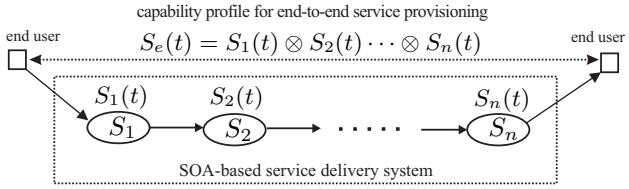

Figure 3. A Model for SOA-based end-to-end service delivery in network virtualization.

The capability profiles of all the service components in the delivery system can be composed into one profile as a capability model for end-to-end service delivery. It is known from network calculus theory that the service curve guaranteed by a series of tandem servers can be obtained through the convolution of all the service curves guaranteed by these servers. Since the capability profile defined in (1) is essentially a service curve guaranteed by a service component, the end-to-end capability profile can be determined accordingly.

For an end-to-end communication service delivery system shown in Figure 3, which consists of n service components, assume that each service component S_i has a capability profiles $S_i(t)$, $i = 1, 2, \cdots, n$, then the end-to-end capability profile for the entire system, denoted by $S_e(t)$, can be obtained as

$$S_e(t) = S_1(t) \otimes S_2(t) \cdots \otimes S_n(t). \tag{3}$$

Suppose each S_i has an LR profile; that is, $S_i(t) = \mathcal{L}[r_i, \theta_i] = \max\{0, r_i(t-\theta_i)\}$, then it can be proved by following network calculus techniques given in [23] that the capability profile of the end-to-end service delivery system is

$$S_e(t) = \mathcal{L}[r_e, \theta_\Sigma] = \mathcal{L}[r_1, \theta_1] \otimes \cdots, \otimes \mathcal{L}[r_n, \theta_n] \tag{4}$$

where $r_e = \min\{r_1, r_2, \cdots, r_n\}$ and $\theta_\Sigma = \sum_{i=1}^{n} \theta_i$.

Equation (4) implies that if each service component in an end-to-end service delivery system can be described by an LR profile, then the end-to-end service provisioning capability can be modeled by an LR profile. The latency parameter of the end-to-end profile is the summation of the latency parameters of all service components in the system, and the end-to-end service rate parameter is the minimum service rate of all the service components.

In order to analyze end-to-end performance of communication services in network virtualization, it is necessary to develop a general approach to characterizing the traffic loads generated by the various applications. The concept of *arrival curve* in network calculus is employed here as a general load profile. Let $R(t)$ denote the accumulated amount of traffic that arrives at the entry of a service delivery system by the time instant t. Given a non-negative, non-decreasing function $A(\cdot)$, if for any time instant s such that $0 < s < t$

$$R(t) - R(s) \leq A(t-s), \tag{5}$$

then we say that the service delivery system has a ***load profile*** $A(\cdot)$. A load profile gives an upper bound for the amount of traffic that a networking session can load on a service delivery system. Since the profile is defined as a general function of time, it can be used to describe the traffic load generated by various networking applications.

Currently most QoS-capable networking systems apply traffic regulation mechanisms at network boundaries to shape arrival traffic. The traffic regulators most commonly used in practice are leaky buckets. A networking session constrained by a leaky bucket loads its service delivery system a profile $A(t) = \min\{pt, \sigma + \rho t\}$, where p, ρ, and σ are respectively called the peak rate, sustained rate, and maximal burst size of the profile.

Now we focus our analysis on the maximum end-to-end delay since it is a critical performance parameter for most communication services in the Internet, such as Internet telephony and video/audio streams. It can be shown by following network calculus that for a service delivery system with a capability profile $S(t)$ under traffic described by a load profile $A(t)$, the maximum end-to-end delay guaranteed by the system, denoted as d_m^e, can be determined as

$$d_m^e = \max_{t:t\geq 0} \{\min\{\delta : \delta \geq 0 \quad A(t) \leq S(t+\delta)\}\}. \tag{6}$$

Since the LR proCle models the infrastructure services provided by typical InPs and leaky bucket traCc regulator is widely deployed in network infrastructures, we are particularly interested in analyzing the end-to-end delay performance of service delivery systems with an LR capability proCle and a leaky bucket load proCle. Suppose a service delivery system has an LR capability proCle $S(t) = \max\{0, r_e(t - \theta_\Sigma)\}$ and loaded by a networking session with a proCle $A(t) = \min\{pt, \sigma + \rho t\}$, following (6) it can be shown that the maximum end-to-end delay guaranteed by the service delivery system is

$$d_m^e = \begin{cases} \theta_\Sigma + \left(\frac{p}{r_e} - 1\right)\frac{\sigma}{p-\rho} & \text{for } p > \rho,\ r_e \geq \rho \\ \theta_\Sigma & \text{for } p = \rho,\ r_e \geq \rho \end{cases} \quad (7)$$

where θ_Σ and r_e are respectively the end-to-end latency and rate parameters of this service delivery system.

IV. Resource Allocation for QoS Provisioning of Communication Services in Network Virtualization

SPs in the virtualization-based Internet must obtain sufCient amount of networking resources from the underlying infrastructures for end-to-end QoS provisioning, while at the same time need to minimize the cost for leasing the resources in order to generate revenues. Therefore from a service provider s perspective, it is signiCant to analyze the minimum amount of networking resources that must be acquired for meeting the QoS requirements of communication services.

In network virtualization environments, networking resources are not managed directly by an SP. Instead resources are allocated by InPs in their network infrastructures and are offered to the SP in the form of infrastructure services, which are speciCed by the service contracts between the SP and InPs. The capability proCle of a service component in the SOA-based service delivery system, which models the data transport capacity guaranteed by an infrastructure service to the service system, also provides an approach to analyzing resource allocation for end-to-end QoS provisioning in network virtualization. In an LR proCle, the rate parameter describes the minimum transport capacity guaranteed by an infrastructure service; therefore from an SP s perspective, a key to resource allocation for QoS provisioning is to determine the minimum value of the rate parameter required for meeting an end-to-end QoS requirement.

The minimum bandwidth and maximum delay are the two most important performances required by most QoS-sensitive communication services. For a service that only requires a minimum bandwidth b_{req}, this requirement can be directly used as the required rate value, denoted as r_a. Equation (4) in Section III shows that the rate parameter of an end-to-end service delivery system is equal to the minimum rate of all service components in the system. This implies that the SP must request transport capacity $r_a = b_{req}$ from each underlying infrastructure. The requested capacity will be provided by each InP by allocating bandwidth that is no less than r_a. Please notice that the allocated bandwidth in a network infrastructure could include link transmission bandwidth and switch/router processing capacity as well, depending on the speciCc implementation of the infrastructure.

Analysis on resource allocation for delay performance guarantee can be started from the case of constant rate traCc load; that is when $p = \rho$. The delay analysis result obtained in Section III, as given in (7), shows that in this case the maximum end-to-end delay is equal to the total latency parameter of the service delivery system; that is, $d_m^e = \theta_\Sigma$ for $p = \rho$ and $r_e \geq \rho$. Therefore, the rate parameter r_e of a service delivery system must be no less than the sustained rate of its traCc load in order to guarantee an upper bounded maximum delay, which is a constant that is equal to the latency parameter θ_Σ of the service delivery system. This implies that under constant rate traCc load, allocating more transport capacity than the sustained traCc rate does not improve delay performance. Therefore, the required transport capacity r_a that the SP must acquire from each underlying infrastructure is equal to the traCc sustained rate; that is, $r_a = \rho$ for $p = \rho$.

For the case of variable rate traCc load (i.e., $p > \rho$), equation (7) shows that the end-to-end delay performance is impacted by the following parameters: θ_Σ, the total latency of the end-to-end service delivery system; r_e, the rate parameter that describes the available transport capacity of the service delivery system; and (p, ρ, σ), the characteristics of traCc load on the service system. For a given networking session with a certain traCc load proCle, the maximum end-to-end delay is a function of the total latency and available transport capacity.

Equation (7) shows that under both constant and variable rate traCc loads, the delay performance increases linearly with the total latency parameter θ_Σ. This implies that reducing latency in network infrastructures can significantly improve delay performance for end-to-end service delivery. However, the latency parameter is determined by some infrastructure properties, such as transmission delay and router/switch processing delay, that may not be easily reduced. Equation (7) also shows that d_m^e is a decreasing function of r_e. This means that given the total latency property of a service delivery system, the maximum delay performance can be guaranteed by acquiring sufCient transport capacity from underlying network infrastructures. The tighter is the delay requirement, the more capacity must be acquired by the SP; thus more bandwidth must be allocated to the service system by each InP in its network infrastructure.

Since d_m^e is a decreasing function of r_e and $r_e \geq \rho$ is a necessary condition for delay performance guarantee, a simple analysis on (7) shows that d_m^e achieves its maximum value $D_{max} = \theta_\Sigma + \sigma/\rho$ when $r_e = \rho$. This implies that if the end-to-end delay requirement, denoted as d_{req}^e, is greater than this maximum value; i.e., $d_{req}^e \geq D_{max}$, then the SP just needs to request each InP for a bandwidth allocation that is equal to the sustained

traffic rate; that is $r_a = \rho$ for $d^e_{req} \geq D_{max}$.

Analysis on (7) also shows that the maximum end-to-end delay approaches a lower limit with the increment of r_e. This delay limit can be determined as $D_{min} = \theta_\Sigma - \sigma/(p-\rho)$. Therefore, any end-to-end delay requirement that is tighter than this limit cannot be guaranteed, no matter how much transport capacity the SP acquires from its InPs. That is r_a does not exist for $d^e_{req} < D_{min}$.

Suppose $D_{min} < d^e_{req} < D_{max}$, then the service delivery system must guarantee that

$$d^e_m = \theta_\Sigma + \left(\frac{p}{r_e} - 1\right)\frac{\sigma}{p-\rho} \leq d^e_{req}, \tag{8}$$

from which the transport capacity that the SP must acquire to guarantee d^e_{req} can be determined as

$$r_a = \frac{p\sigma}{(d^e_{req} - \theta_\Sigma)(p-\rho) + \sigma}. \tag{9}$$

In summary for a communication service with a variable rate traffic load, the minimum transport capacity r_a that must be acquired by an SP in order to guarantee the maximum delay requirement d^e_{req} can be determined as

$$r_a = \begin{cases} \rho & d^e_{req} \geq D_{max} \\ \frac{p\sigma}{(d^e_{req} - \theta_\Sigma)(p-\rho)+\sigma} & D_{min} < d^e_{req} < D_{max} \\ \text{not exist} & d^e_{req} \leq D_{min} \end{cases} \tag{10}$$

where $D_{max} = \theta_\Sigma + \sigma/\rho$ and $D_{min} = \theta_\Sigma - \sigma/(p-\rho)$.

V. Resource Utilization for QoS Provisioning of Communication Services in Network Virtualization

In the current Internet architecture, when the end-to-end delivery path for a communication service traverses multiple network domains that belong to different ISPs (autonomous systems), no single ISP has a purview over the entire path. Therefore end-to-end QoS provisioning has to been offered through the collaboration between each pair of neighbor domains. In such a conventional inter-domain QoS mechanism, the total end-to-end delay requirement is partitioned into a set of delay budgets, one for each domain involved in the service provisioning. Each network domain has to allocate sufficient amount of bandwidth in its own network infrastructure to guarantee its delay budget. Collaboration across autonomous network domains for end-to-end QoS provisioning is still a challenging issue for Internet QoS that has not been completely solved [24].

By de-coupling service provisioning and underlying network infrastructures, network virtualization allows a single SP to offer a real *end-to-end* communication service across multiple network infrastructures managed by different InPs. Due to the loose-coupling property of SOA, its application in network virtualization provides standard and flexible service-oriented interfaces for SP-InP interactions and inter-InP collaborations. Therefore SOA-based service delivery offers a promising approach to realizing the separation of service provisioning and infrastructures, which is the key notion of network virtualization.

An end-to-end path traversing multiple Autonomous Systems (AS) with a centralized controller might look like having a similar structure as a SOA-based service delivery system in network virtualization. However, there exists some fundamental difference between these two service delivery mechanisms. Due to the dual roles of service provider and infrastructure operator played by the ISPs in the current Internet, network services are offered to end users by individual ISPs even if there is a controller for an inter-domain path. In network virtualization environments, services are provided to end users by SPs who construct and control end-to-end service delivery systems on top of network infrastructures. SOA-based service delivery enables SPs to provide network services by composing infrastructure services offered by InPs. Applying SOA in network virtualization offers an effective mechanism for supporting interactions between SPs and InPs with a level of flexibility and agility that cannot be achieved by conventional inter-AS coordination in the current Internet.

SOA-based service delivery system may also simplify service and resource management for end-to-end QoS provisioning in network virtualization. Although each InP still needs to allocate bandwidth in its infrastructure to support end-to-end service delivery, the amount of bandwidth allocation is determined by an SP with an end-to-end vision and requested by the SP as part of the infrastructure service offered by the InP. In this paper the resource allocation scheme enabled by network virtualization is referred to as *end-to-end allocation* while the resource allocation within individual domains without an end-to-end vision is called *domain-based allocation*. This section compares these two allocation schemes and analyzes the impact of network virtualization on bandwidth utilization for end-to-end QoS provisioning in the future Internet.

Without losing generality this section considers an end-to-end communication service, whose delivery path traverses the network infrastructures that belong to n different domains, I_i, $i = 1, 2, \cdots, n$, each of which is viewed as an InP in network virtualization. Following the modeling approach developed in Section III, the SOA-based end-to-end delivery system for this communication service, as shown in shown in Figure 4, consists of n service components S_i, $i = 1, 2, \cdots, n$, which respectively represents the infrastructure service provided by I_i. We assume that the service capability of each S_i can be described by an LR profile $\mathcal{L}[r_i, \theta_i]$. The end-to-end delay requirement and the end-to-end bandwidth allocation for achieving this requirement are respectively denoted as d^e_{req} and r^e_a. The delay budget for the network domain I_i is denoted as d^i_{req} and the amount of bandwidth determined by the domain-based allocation scheme for guaranteeing this delay budget is denoted as r^i_a.

We first examine bandwidth utilization for the case of constant rate traffic load, where $p = \rho$. Analysis in

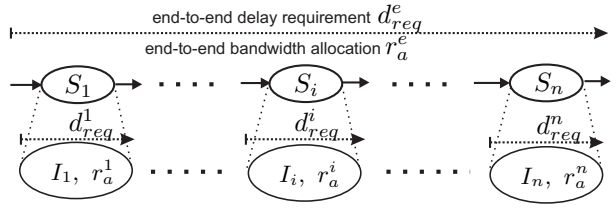

Figure 4. The end-to-end and domain-based bandwidth allocations for delay performance guarantee

Section IV has shown that the transport capacity required by the end-to-end allocation scheme is $r^e_a = \rho$. Since the delay performance and bandwidth allocation analysis given in Sections III and IV apply to both end-to-end service system and each single service component as well, we can get that for each service component S_i, the bandwidth that must be allocated in the infrastructure I_i for guaranteeing its delay budget is also equal to the sustained traffic rate; that is, $r^i_a = \rho$. Therefore the bandwidth requirements determined by both end-to-end and domain-based allocation schemes are the same for constant rate traffic loads.

For the case of variable rate traffic load and the typical situation in which the end-to-end delay requirement satisfies $D_{min} < d^e_{req} < D_{max}$, equation (10) shows that the end-to-end allocation scheme requires that

$$r^e_a = \frac{p\sigma}{(d^e_{req} - \theta_\Sigma)(p-\rho) + \sigma}. \tag{11}$$

That is, the SP will request each InP to allocate bandwidth r^e_a in its network infrastructure. Suppose the delay budget d^i_{req} for S_i satisfies $D^i_{min} < d^i_{req} < D^i_{max}$, where $D^i_{min} = \theta_i - \sigma/(p-\rho)$ and $D^i_{max} = \theta_i + \sigma/\rho$, then the transport capacity required in I_i to guarantee d^i_{req} can be determined as

$$r^i_a = \frac{p\sigma}{(d^i_{req} - \theta_i)(p-\rho) + \sigma}, \tag{12}$$

which means that the domain-based allocation scheme requires the amount of bandwidth r^i_a to be allocated in the network domain I_i.

The ratio between the amounts of bandwidth required by the domain-based allocation and the end-to-end allocation is

$$\mathcal{U} = \frac{r^i_a}{r^e_a} = \frac{(d^e_{req} - \theta_\Sigma)(p-\rho) + \sigma}{(d^i_{req} - \theta_i)(p-\rho) + \sigma}. \tag{13}$$

Let $\Delta d_e = d^e_{req} - \theta_\Sigma$ and $\Delta d_i = d^i_{req} - \theta_i$. Equation (13) shows that if $\Delta d_i < \Delta d_e$ then $\mathcal{U} > 1$; otherwise $\mathcal{U} \le 1$. This implies that for a network domain with a delay budget that is loose enough to satisfy $\Delta d_i > \Delta d_e$, the domain-based allocation scheme of the conventional inter-domain QoS mechanism may actually require less amount of bandwidth than what is required by end-to-end resource allocation in network virtualization. However given a fixed end-to-end delay requirement, a loose delay budget for one network domain means tighter delay budgets for others, which requires more bandwidth allocation in other domains. Independent ISPs in the current Internet have conflicting interests and are unlikely to sacrifice their own bandwidth resources for others benefits. Therefore it is reasonable to assume that the end-to-end delay requirement is equally partitioned among all domains when the domains have an identical latency property; that is, $d^i_{req} = d$ and $\theta_i = \theta$ for $i = 1, 2, \cdots, n$. Then $d^e_{req} = \sum_{i=1}^n d^i_{req} = nd$, and from (4) we have $\theta_\Sigma = \sum_{i=1}^n \theta_i = n\theta$. Therefore,

$$r^i_a = \frac{p\sigma}{(d-\theta)(p-\rho) + \sigma}, r^e_a = \frac{p\sigma}{n(d-\theta)(p-\rho) + \sigma}$$

and the bandwidth ratio becomes

$$\begin{aligned} \mathcal{U} = \frac{r^i_a}{r^e_a} &= \frac{n(d-\theta)(p-\rho) + \sigma}{(d-\theta)(p-\rho) + \sigma} \\ &= 1 + \frac{(n-1)(d-\theta)(p-\rho)}{(d-\theta)(p-\rho) + \sigma}. \end{aligned} \tag{14}$$

For any variable rate traffic load, $p > \rho$. The delay budget assigned to a network domain is typically larger than the latency property of its network infrastructure; i.e. $d > \theta$. Therefore, for a service delivery system consisting of more than one domain ($n \ge 2$), the ratio $\mathcal{U} > 1$.

The above analysis shows that in the considered scenarios, in order to achieve the same level of delay performance guarantee, the conventional inter-domain QoS mechanism allocates more bandwidth in each individual network infrastructure than the amount of bandwidth required by the end-to-end allocation scheme. This implies that the end-to-end service provisioning enabled by network virtualization typically improves resource utilization for QoS provisioning of communication services with variable rate traffic loads. An essential reason for the improvement in resource utilization lies in the end-to-end purview obtained by SPs in network virtualization and the more effective resource management enabled by such purview. Applying SOA in network virtualization allows SPs to take full advantage of the benefits enabled by such end-to-end purview through simplified and flexible interactions with InPs.

Equation (14) also shows that the bandwidth ratio $\mathcal{U}$ is associated with multiple parameters, including the delay requirement d, the latency parameter θ, and traffic load parameters (p, ρ, σ). Given an end-to-end service delivery system under a certain traffic load, this ratio $\mathcal{U}$ is a function of delay requirement d. The partial derivative of $\mathcal{U}$ with respect to d is

$$\frac{\partial \mathcal{U}}{\partial d} = \frac{(n-1)(p-\rho)\sigma}{[(d-\theta)(p-\rho) + \sigma]^2} > 0 \quad (n \ge 2,\ p > \rho). \tag{15}$$

This shows that $\mathcal{U}$ is an increasing function of d. Therefore, the greater value the delay requirement has, the bigger is this bandwidth ratio. This implies that end-to-end resource allocation in network virtualization achieves more improvement in bandwidth utilization for communication services with larger delay requirements.

VI. Numerical Examples

Numerical examples are given in this section to illustrate applications of the developed techniques and the insights obtained from the analysis. Suppose a network service provider SP constructs an end-to-end service delivery system by assembling resources from three network infrastructures I_i, $i = 1, 2, 3$, and the networking capability offered by each infrastructure to the service system can be modeled by an LR profile. Considering the scenarios in which the SP offers end-to-end communication services for two applications $\mathcal{A}_1$ and $\mathcal{A}_2$. Application $\mathcal{A}_1$ transmits a video stream through a networking session f_1, while $\mathcal{A}_2$ generates a networking session f_2 to deliver a flow of audio traffic. Both $\mathcal{A}_1$ and $\mathcal{A}_2$ require a maximum end-to-end delay. The traffic parameters for f_1 and f_2 are respectively ($p_1 = 5.3$Mb/s, $\rho_1 = 1.5$Mb/s, $\sigma_1 = 140$kbits) and ($p_2 = 3.2$Mb/s, $\rho_2 = 1.1$Mb/s, $\sigma_2 = 300$kbits), which are derived from the traffic characteristics reported in [25] and [26].

The maximum end-to-end delay for each networking session is calculated with various amounts of available bandwidth in the service delivery system. The obtained results are plotted in Figure 5 where d_e^1 and d_e^2 denote the delays for f_1 and f_2 respectively. This figure shows that the maximum delay performances of both networking sessions are decreasing functions of the available bandwidth. This means that the more transport capacity the SP acquires from underlying network infrastructures, the better delay performance can be guaranteed to end users. Comparing the delay curves of f_1 and f_2 shows that the two sessions has different delay values with the same amount of available bandwidth. This implies that the delay performance of a service delivery system is determined by not only the available bandwidth in the system but also the characteristics of traffic load on the system.

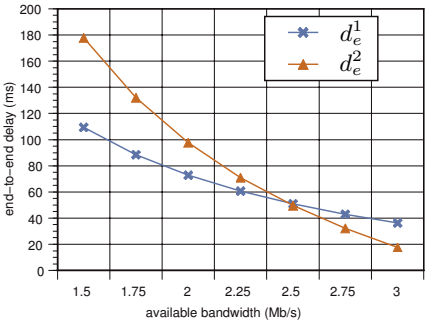

Figure 5. Delay performance of networking sessions f_1 and f_2.

Bandwidth allocation for end-to-end delay performance guarantee for the applications is also analyzed. The amounts of bandwidth that must be acquired by the SP from the network infrastructures in order to guarantee a set of end-to-end delay requirements are calculated. The results are plotted in Figure 6 where bandwidth allocations for f_1 and f_2 are denoted as r_e^1 and r_e^2 respectively. From this figure we can see that the required amounts of bandwidth for both networking sessions decrease when

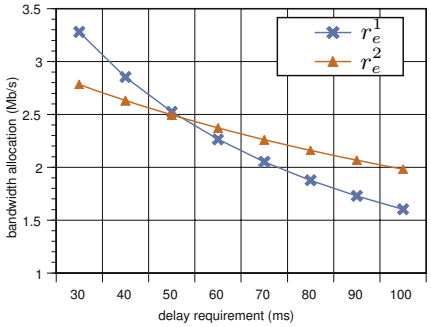

Figure 6. Bandwidth allocations for f_1 and f_2.

the delay requirement value increases. This means that more transport capacity must be acquired by the SP from each underlying network infrastructure in order to provide a tighter end-to-end delay guarantee.

Figure 6 also shows that different amounts of bandwidth are required by the two networking sessions for achieving the same delay performance. This means that bandwidth allocation is impacted by traffic load characteristics as well as the delay requirement. Comparing the two bandwidth curves shows that r_e^1 drops with the increment of delay requirement faster than r_e^2 does. This implies that for networking sessions with different traffic load characteristics, the same extent of improvement in delay performance requires different amounts of increment in bandwidth allocation. Observation on the load profile parameters of these two sessions indicates that f_1 has relatively more fluctuating traffic (bigger difference between the peak and sustained rates and shorter burst size) than f_2. Such an observation of the bandwidth curves and the load profiles tends to show that sessions with more bursty traffic loads require more bandwidth allocation for achieving a certain degree of improvement in delay performance. Impact of traffic load characteristics on resource allocation for end-to-end service provisioning in network virtualization is an interesting and important research topic, and is left by the author to future study due to the space limitation of this article.

In order to evaluate bandwidth utilization for QoS provisioning in network virtualization, the amounts of bandwidth required in an individual domain by the domain-based allocation scheme are also calculated and compared with the results of end-to-end resource allocation. The three network domains in this example are assumed to have the same latency value and the end-to-end delay requirement is divided equally into the three domains. The obtained data for networking sessions f_1 and f_2 are plotted in Figures 7 and 8, in which the domain-based allocation results for f_1 and f_2 are respectively denoted as r_d^1 and r_d^2. The bandwidth ratios for the two sessions are calculated as $\mathcal{U}_1 = r_d^1/r_e^1$ and $\mathcal{U}_2 = r_d^2/r_e^2$, and the obtained data are plotted in Figure 9.

Figures 7 and 8 all show that the amounts of bandwidth required by both the end-to-end and domain-based allocation schemes are decreasing functions of the end-to-end

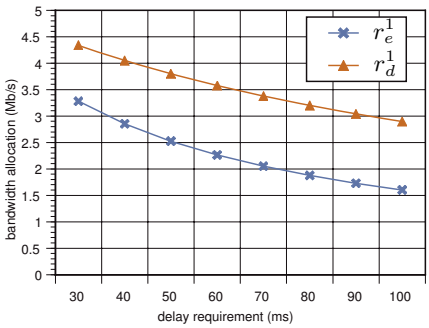

Figure 7. End-to-end and domain-based bandwidth allocations to f_1.

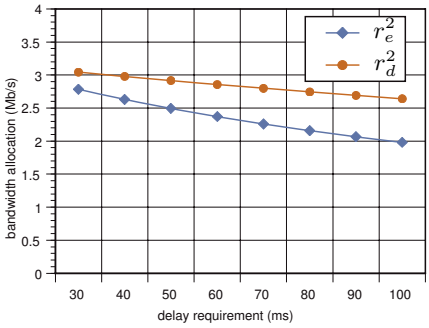

Figure 8. End-to-end and domain-based bandwidth allocations to f_2.

delay requirement. That is, more bandwidth is required by both schemes to offer a tighter delay guarantee. We can also see that for achieving the same level of delay performance, the end-to-end allocation scheme always requires a less amount of bandwidth than the domain-based allocation scheme does. This implies that in this example the end-to-end resource allocation enabled by network virtualization achieves higher bandwidth utilization compared with the domain-based allocation scheme of the conventional inter-domain QoS mechanism.

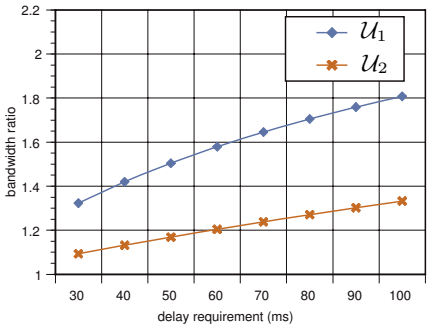

Figure 9. Bandwidth ratios for networking sessions f_1 and f_2 for achieving different delay objectives.

The bandwidth ratios $\mathcal{U}_1$ and $\mathcal{U}_2$ for achieving various end-to-end delay objectives for sessions f_1 and f_2 are given in Figure 9. A bandwidth ratio of a networking session presents the extent of bandwidth utilization improvement caused by network virtualization for the session. We can see from this Figure that both $\mathcal{U}_1$ and $\mathcal{U}_2$

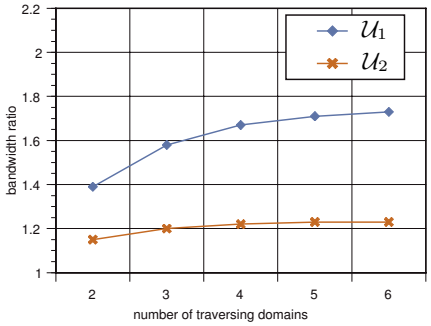

Figure 10. Bandwidth ratios for networking sessions f_1 and f_2 passing different number of infrastructures.

increase with the delay requirement, which implies that more bandwidth utilization improvement is obtained for greater delay requirement values. Comparing the curves of $\mathcal{U}_1$ and $\mathcal{U}_2$ in Figure 9 shows that for the same delay requirement, the bandwidth ratio for f_1 is greater than the ratio for f_2, which shows that traffic load characteristics also have an impact on the extent of bandwidth utilization improvement in network virtualization.

In order to evaluate the influence of the number of traversing network infrastructures (domains) on bandwidth utilization, the bandwidth ratios of the two networking sessions for achieving 60 ms end-to-end delay objective are calculated with different numbers of domains passed by the service delivery system. The obtained results are plotted in Figure 10. This figure shows that both ratios increase with the number of domains in the system, which implies that the more domains the service delivery system traverses, the bigger is the difference between the amounts of bandwidth required by the end-to-end and domain-by-domain allocation schemes. We can see from this figure that the slopes of both curves decrease and the ratios tend to approach a value with the increment of the domain number. This implies that for a given networking session and a delay objective, the impact of the traversing domain number becomes insignificant when the number is greater than a threshold (Figure 10 indicates that such a threshold is 6 for f_1 and 5 for f_2 in this example). Again the figure shows the influence of traffic load characteristics on bandwidth utilization reflected by the different ratio values of the two sessions for the same number of traversing domains.

VII. Conclusions

Network virtualization has been proposed as a key attribute of the next generation inter-networking paradigm. A key technical issue for network virtualization in the Internet lies in end-to-end QoS provisioning across heterogeneous network infrastructures. The SOA, as an effective architectural principle for coordinating heterogeneous systems to meet diverse service requirements, offers a promising approach to addressing this challenging issue.

The research presented in this article investigated application of the SOA principle in network virtualization

to facilitate end-to-end QoS provisioning. Specifically a SOA-based delivery system is proposed for end-to-end communication services in network virtualization environments. Such a service delivery system enables SPs to synthesize networking resources of InPs through an infrastructure-as-a-service paradigm. An analytical model and a performance evaluation technique are developed for this service delivery system. Then this paper examines resource allocation for QoS provisioning of communication services in network virtualization. An approach to determining the required amounts of resources for QoS guarantees of communication services is developed. Resource utilization for end-to-end QoS provisioning in network virtualization is also analyzed in this paper and compared with that of the inter-domain QoS mechanism available in the current Internet. The service-oriented modeling and analysis techniques developed in this paper are general and flexible; thus are applicable to the heterogeneous networking systems in the future Internet.

The research findings reported in this paper provide insights about the relationship between the QoS performance guaranteed by a SOA-based service delivery system in network virtualization and a set of attribute parameters regarding the service delivery system, including the amount of transport capacity acquired by the service provider for the system, the latency property of the system, and the characteristics of traffic load on the system. Such insights are very useful to service providers for determining the minimum amount of resources they must acquire from underlying infrastructures to achieve a certain level of QoS guarantee, which allows them to minimize resource costs and maximize revenues. Analysis in this paper also shows that in typical inter-domain networking scenarios, the end-to-end resource allocation enabled by network virtualization achieves higher bandwidth utilization for QoS provisioning than the conventional inter-domain QoS mechanism in the current Internet. Study in this paper indicates that the end-to-end purview of service delivery obtained by SPs in network virtualization environments allows more effective service management for QoS provisioning with improved resource utilization. Application of SOA in network virtualization may simplify SP-InP interactions and inter-InP collaborations, which enables SPs to take full advantage of the end-to-end purview through an infrastructure-as-a-service paradigm and greatly facilitates service provisioning in network virtualization.

REFERENCES

[1] N. M. Chowdhury and R. Boutaba, A survey of network virtualization,*f Computer Networks*, vol. 54, pp. 862˘876, 2010.

[2] J. Turner and D. E. Taylor, Diversifying the Internet,*f* in *Proc. of IEEE Global Communication Conference*, 2005.

[3] N. Feamster, L. Gao, and J. Rexford, How to lease the Internet in your spare time,*f ACM SIGCOMM Computer Communications Review*, vol. 37, no. 1, pp. 61˘64, 2007.

[4] G. P. Group, GENI design principles,*f IEEE Computer*, vol. 39, no. 9, pp. 102˘105, 2006.

[5] N. Niebert, S. Baucke, I. El-Khayat, M. Johnsson, B. Ohlman, H. Abramowica, K. Wuenstel, H. Woesner, J. Ouittek, and L. M. Correia, The way 4WARD to the creation of a future Internet,*f* in *Proc. of the IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications*, 2008.

[6] J. Song, M. Y. Chang, S. S. Lee, and J. Joung, Overview of ITU-T NGN QoS control,*f IEEE Communications*, vol. 45, no. 9, pp. 116˘123, 2007.

[7] N. M. Chowdhury and R. Boutaba, Network virtualization: state of the art and research challenges,*f IEEE Communications*, vol. 47, no. 7, pp. 20˘26, 2009.

[8] M. Boucadair, P. Levis, D. Griffin, N. Wang, M. Howarth, G. Pavlou, E. Mykoniati, P. Georgatsos, B. Quoitin, J. R. Sanchez, and M. L. Garcia-Osma, A framework for end-to-end service differentiation: Network planes and parallel internets,*f IEEE Communications*, vol. 45, no. 9, pp. 134˘143, 2007.

[9] E. Mingozzi, G. Stea, M. Callejo-Rodriguez, J. Enriquez-Gabeiras, G. G. de Blas, F. J. Romon-Salquero, W. Burakowski, A. Beben, O. D. M. D. L. B. J. Sliwinskic, H. Tarasiukc, and E. Monteiro, EuQos: End-to-end quality of service over heterogeneous networks,*f Computer Communications*, vol. 32, no. 12, pp. 1355˘1370, 2009.

[10] T. O. M. A. (OMA), OMA Service Environment Architecture,*f* 2007.

[11] E. Grasa, G. Junyent, S. Figuerola, A. Lopez, and M. Savoie, Uclpv2: A network virtualization framework built on web services,*f IEEE Communications*, vol. 46, no. 3, pp. 126˘134, 2007.

[12] P. Szegedi, S. Figuerola, M. Campanella, V. Maglaris, and C. Cervello-Pastor, With evolution for revolution: Managing federica for future internet research,*f IEEE Communications*, vol. 47, no. 7, pp. 34˘39, 2009.

[13] T. Magedanz, N. Blum, and S. Dutkowski, Evolution of SOA concepts in telecommunications,*f IEEE Computer*, vol. 40, no. 11, pp. 46˘50, 2007.

[14] D. Griffin and D. Pesch, A survey on web services in telecommunications,*f IEEE Communications*, vol. 45, no. 7, pp. 28˘35, 2007.

[15] Q. Duan, Modeling and analysis for end-to-end service performance in virtualization-based next generation internet,*f* in *Proc. of the IEEE 2010 Global Communication Conference (GlobeCom'10)*, 2010.

[16] , End-to-end modeling and performance analysis for network virtualization in the next generation internet,*f International Journal of Communication Networks and Distributed Systems*.

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

[18] OASIS, Reference Model for the Service-Oriented Architecture version 1.0,*f* 2006.

[19] W3C, Web Service Description Language (WSDL) version 2.0,*f* 2007.

[20] OASIS, Universal Description, Discovery and Integration (UDDI) version 3.0.2,*f* 2005.

[21] , Business Process Execution Language for Web Services (BPEL-WS) version 1.1,*f* 2007.

[22] H. Berndt, T. Hamada, and P. Graubmann, Tina: Its achievements and its future directions,*f IEEE Communications Surveys & Tutorials*, vol. 3, no. 1, pp. 2˘16, 2000.

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

[24] P. Jacobs and B. Davie, Technical challenges in the delivery of interprovider QoS,*f IEEE Communications*, vol. 43, no. 6, pp. 112˘118, 2005.

[25] M. Butto, E. Cavallero, and A. Tonietti, Effectiveness of the leaky bucket policing mechanisms in ATM networks,*f*

IEEE Journal of Selected Areas of Communications, vol. 9, no. 4, pp. 335˘342, 1991.

[26] F. H. P. Fizek and M. Reisslein, Mpeg-4 and h.263 video traces for network performance evaluation,*f* Technische Universitt Berlin, Tech. Rep. TKN-00-06, 2000, telecommunication Network Group.

Qiang Duan is currently an Assistant Professor of Information Science and Technology at the Pennsylvania State University Abington College. His research interests include data communications, computer networking, the future Internet architecture, Web services, and Cloud computing. 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 International Journal of Network Protocols and Algorithms and Journal on Internet and Distributed Computing Systems. He has also served as a technical program committee member for numerous conferences. Dr. Duan received his PhD 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. Dr. Duan is a member of the IEEE Communications Society.