A Two-Dimensional Architecture for End-to-End Resource Management in Virtual Network Environments

Ning Wang, University of Surrey
Yan Zhang, Simula Research Laboratory
Joan Serrat and Juan Luis Gorricho, Universitat Politècnica de Catalunya Spain
Tao Guo, University of Surrey
Zheng Hu and Ping Zhang, Beijing University of Post and Telecommunications

Abstract

In recent years, various network virtualization techniques have been proposed for flexibly supporting heterogeneous services over virtual network platforms. However, systematic views on how virtual network resources (VNRs) can be practically managed in such open environments has been missing till now. To fill the gap, we present in this article a two-dimensional architecture for end-to-end VNR management from distinct viewpoints of service providers and network providers. The horizontal dimension of VNR management allows SPs to bind VNRs rented from heterogeneous NPs to form unified end-to-end service delivery platforms. The vertical dimension of VNR management enables NPs to perform cost-efficient allocation of VNRs to requesting SPs, but without necessarily forcing themselves to collaborate with each other. Such a VNR management architecture will complement existing network virtualization platforms in accelerating the realization of virtual resource sharing in the future Internet business marketplaces.
IP/multiprotocol label switching (MPLS) backbone networks and mobile/wireless networks. In this case, it is the SP’s responsibility to “concatenate” VNR capabilities provided by individual NPs to form a global virtual network infrastructure for supporting its own services to be deployed. This dimension is known as horizontal VNR management on the SP side.

In this article, we present a holistic architecture framework for VNR management according to the two-dimensional views from both the SP and NP sides. A set of functional components and their interdependencies are specified for supporting VNR manipulation by the two types of stakeholders. With network virtualization, SPs play a more active role in controlling network resources by means of deploying customized networking mechanisms and protocols dedicated to their own virtual network platforms. This is in contrast to the conventional scenario where heterogeneous SPs only demand specific service requirements, and passively rely on the underlying NP to dimension its physical resources accordingly, typically through unilateral traffic management/engineering paradigms [10, 11, 16]. Due to the fact that an NP normally treats customer traffic in an aggregated manner, the flexibility in providing service differentiation is limited. In this article we illustrate how SPs and NPs can systematically manipulate virtualized resources according to the two-dimensional management architecture in virtual network environments. A set of key VNR operations on both the SP and NP sides are specified based on the proposed architecture and its functional components.

Architecture Overview

Figures 2 and 3 show the functional architecture for provisioning and controlling VNRs on the SP and NP sides. Fundamentally, an SP’s role is to provide specific value-added services to customers based on the VNR “segments” rented from individual underlyiing NPs. In order to offer end-to-end services in the Internet, an SP may need to rent separate VNR segments from individual interconnected NPs, and then horizontally “bind” them together to form an integrated virtual service platform on top of the physical network infrastructure. In this case, the top-level objective of SPs is to maximize their global service capability with minimum rented VNRs, typically through cost-efficient VNR provisioning and engineering operations. On the other hand, NPs rent their edge-to-edge VNR segments to incoming SPs with heterogeneous service requirements. It is the objective of NPs to attract and satisfy maximum incoming VNR requests based on available physical resources in order to maximize their revenues, typically through optimized provisioning and allocation of their own network resources. It is worth mentioning that NPs, even backbone network operators, do not need to worry about the end-to-end service requirements on the SP side, as it is the responsibility of individual SPs to bind their rented VNR segments together according to their own service targets. The benefit is that underlying network operators, who are mainly rivals in the Internet marketplace, will not be forced to cooperate with each other in order to provide end-to-end services. Instead, each of them only worries about how to appropriately allocate its own VNR segments to the requesting SPs, while each SP may apply its dedicated strategy in implementing the end-to-end service capability through its own VNR binding operation. A distinct example is for an SP to deploy its own routing protocol on top of its end-to-end virtual network infrastructure with horizontally bound VNR capabilities which covers multiple underlying NP domains. In such a scenario, flexible and customized overlay routing on the SP side is able to override the rigid BGP routing configurations and path selections in the physical network.

Horizontal VNR Management on the SP Side

We first focus on the horizontal VNR management performed by SPs. First of all, an SP needs to plan its basic service objectives, for instance, what types of services to be offered to customers and what will be the service coverage in the Internet? Such top-level service planning provides necessary guidance and policies to the VNR provisioning and binding component that is a decision maker for actual VNR requesting and engineering. VNR provisioning and binding plays the key role in resource management on the SP side on a long timescale (e.g., monthly). Based on the forecasted customer demand that can be derived from service level agreements (SLAs) with customers, the function determines the appropriate VNRs to be rented from NPs in long-term operation. According to the expected service level objectives, this component also performs offline resource engineering based on the rented VNR segments provided from the underlying NPs (e.g., the determination of virtual node locations and virtual path selections). One distinct operation task in this case is VNR binding in order to concatenate local VNR capabilities rented from individual NPs for provisioning of a virtual platform with end-to-end service capabilities. A detailed description of VNR binding will be introduced in the next section.

A complementary function to long-term VNR provisioning is adaptive VNR control, which caters for short-term or ad hoc...
hand, given the uncertainty in the VNR renting modes from and possibly even content caching capabilities. On the other such as bounded edge-to-edge delay and bandwidth support, account the common quality of service (QoS) requirements.

For instance, if an NP is able to envisage that the majority of the NP might be sufficient for accommodating changing or new customer service requirements. In this case the application service assurance monitoring component needs to capture the up-to-date service conditions to the current end users on their ongoing applications. Such information needs to be fed into the adaptive VNR control component for on-the-fly decision making, for instance, to either issue an adjusted VNR request to the NP based on the original (static) one, or simply perform admission control over excessive incoming user demands.

A special case for VNR renting is the construction of short-term virtual network infrastructures for some one-off events. For instance, a media broadcasting service provider may want to establish a global content delivery network infrastructure based on information-centric networking (ICN) technologies (e.g., the publish-subscribe paradigm) [12] in order to provide live broadcasting services for an event such as the Olympic Games or simply a music concert for a given period. In this case, there is no long-term VNR renting contract between this SP and the underlying NPs; additionally, the number of customers that will use such services might be very difficult to predict in advance. Under such circumstances, the media service provider may only request short-term virtual network resources from the underlying NPs, and at the same time perform adaptive VNR control according to the actual user demand on the service. If necessary, adjusted VNR requests are negotiated for accommodating more incoming users (e.g., in a flash crowd situation).

**Vertical VNR Management on the NP Side**

On the NP side, a key issue is how to partition its own network resources for optimized VNR engineering in order to maximize revenues with cost-efficient resource utilizations. For instance, if an NP is able to envisage that the majority of the requesting SPs will target multimedia-based content delivery services, the provisioning of the VNRs should take into account the common quality of service (QoS) requirements such as bounded edge-to-edge delay and bandwidth support, and possibly even content caching capabilities. On the other hand, given the uncertainty in the VNR renting modes from SPs (e.g., long-term vs. short term/ad hoc), it is also beneficial for an NP to have initial planning in dealing with different types of incoming VNR requests, for instance, determination of the proportion of VNRs reserved only for ad hoc or short-term requests. These issues are to be considered by the top-level VNR planning component, which provides general guidance and policies to the VNR engineering component for handling specific VNR requests. An example of such guidance can be that the maximum edge-to-edge delay associated with any VNR segment allocated to SPs should be 100 ms, as this upper bound is the outcome of the dimensioning of physical network resources.

Upon receiving a specific VNR request from an SP, the NP seeks to process it through the VNR request handling interface component which provides necessary input to the central VNR engineering component. The VNR engineering component plays an essential role in the decision-making process during network operations. Normally, long-term VNR requests should be processed in batches together on a certain timescale (e.g., at monthly intervals). Once these requests have been accepted, the corresponding VNR segments are actually allocated through the offline VNR allocation operation. The benefit of batch-based VNR request processing is that such an operation gives the NP a clear view of the global VNR availability at the top level in order to perform long-term VNR allocations. As mentioned previously, due to the uncertainty associated with the service demand requested by end users on the SP side, an SP may issue adjusted VNR requests on the fly based on their original long-term VNR requests. On the other hand, according to the guidance provided from the VNR planning component, a specific portion of VNR capabilities can be reserved for allocation upon some ad hoc VNR requests at any time. To deal with both adjusted VNR requests and brand new ad hoc VNR requests, the short-term adaptation intelligence in VNR engineering is needed for making runtime decisions on whether or not to accept them. Such decisions are typically made based on necessary inputs from both service and network monitoring functions that are responsible for reporting the up-to-date service and resource utilization conditions. As the outcome, the online VNR allocation component is responsible for enforcing the actual on-the-fly VNR allocations.

The VNR service assurance monitoring component is respon-
sible for verifying the fulfillment of the current VNR allocations, while the network resource monitoring component provides complementary information on the current network conditions and physical resource availability. For instance, in event of a network anomaly in the physical network (e.g., link or node failure), it is the responsibility of the network resource monitoring function to immediately inform the VNR engineering component for taking necessary recovery actions [13]. Depending on the severity of the fault, short-term adaptations can be applied for performing local service recovery for the affected VNR segments without disrupting the ongoing service sessions. In the event of a serious or large-scale network failure, the VNR engineering component may need to interrupt all ongoing VNR sessions and perform global network recovery in an offline manner.

Discussion of Key VNR Operations

In this section we discuss in detail the key VNR operations in the two dimensions of the VNR management paradigm, VNR requesting and binding on the SP side and VNR allocation on the NP side.

VNR Requesting

According to heterogeneous service characteristics on the SP side, the corresponding VNR requests may have very different requirements on the virtual resources to be rented. We first summarize below a set of common properties that are generally needed in all types of VNR requests. Following that, we illustrate two distinct examples of heterogeneous VNR requesting based on different types of services.

In general, the following information is necessary to be expressed in all types of VNR requests. From the functional architecture point of view, unless specifically indicated, it is the top-level service planning component on the SP side that should determine this set of generic properties.

* VNR coverage and topology: Fundamentally, an SP should specify in its VNR request the targeted geographical service coverage on top of the physical network infrastructure. This can be the case when the underlying NP has multicontinental Internet access coverage, while the requesting SP only deploys its service within a specific country or continent. In addition, the topology of the virtual network infrastructure should also be specified, for instance, the placement of virtual nodes and establishment of virtual links, optionally with dedicated capability support (e.g., storage, computing) to be associated with physical nodes and dedicated bandwidth support to be associated with the virtual links in the infrastructure. The actual optimization of the virtual network topology by taking into account specific customer demands is covered by the VNR provisioning and binding function component.

* VNR renting timescale: In all VNR requests, the timescale of planned resource renting should also be specified. Certainly this is obvious on the SP side, depending on how long the deployed services are to be provided to end users. Concerning NPs, as previously mentioned, long-term VNR requests not only help the underlying NPs to efficiently provision its VNRs to be allocated, but are also beneficial to network stability during operation. It is not difficult to infer that having excessive ad hoc or short-term VNR renting requests makes it difficult for NPs to efficiently manage the network resources to be shared.

* Runtime VNR adjustment options: Due to unpredicted service subscription demand from end users, an SP may need to adjust its VNR requests during runtime if necessary. In order to avoid potential contentions between SPs on VNRs requesting adjustments during network operation time, each SP and the NP can negotiate a priori in the initial VNR request how future runtime request adjustments can be performed. One scenario is to specify an upper limit of additional VNR capabilities that can be requested in any future runtime adjustment from the SP. If the additionally requested VNRs are still not sufficient for supporting incoming end users, the SP needs to apply its own admission control policies to avoid service disruptions to existing end users.

* Charging/tariff model: This simply indicates how SPs’ usage of the VNRs will be charged by NPs. The price of the VNRs may depend on not only the types and locations of the

Figure 3. Functional architecture for vertical VNR management architecture.
resources to be requested but also the time (peak/off-peak) at which the NP is operating.

Now we use two examples to illustrate customized VNR requesting scenarios in supporting different types of services. We consider a voice over IP (VoIP) service provider (VSP) and a content service provider (CSP) that offers video-on-demand services based on ICN technologies with various in-network content manipulating functions. Table 1 summarizes some key differences concerning the VSP and the CSP when they make VNR requests for deploying their own services. For instance, while a VSP only requires the virtual communication infrastructure for supporting end-to-end IP telephony services, a CSP may need more comprehensive in-network virtual resources such as distributed content storage space and in-network computing intelligence for on-the-fly adaptation of video content while it is being transmitted toward end users with different terminal capabilities. On the other hand, interactive communications in VoIP applications demand low delay/jitter, while bandwidth availability plays a more important role in multimedia content distribution services.

**VNR Binding**

VNR binding refers to SPs' concatenation operations on individual VNR segments rented from multiple interconnected NPs in order to provision specific end-to-end services. Such operation should be fulfilled by the VNR Provisioning and Binding function component. Figure 4a shows a simple example on horizontal virtual resource binding by a CSP whose service coverage spans two underlying NP networks. In this example both NP 1 and NP 2 allocate their own VNR substrates to the CSP upon request. It is worth mentioning that these two substrates are edge-to-edge virtual resource segments bounded within each NP's physical network. In this case, it is the responsibility of the CSP to deploy its own mechanisms to concatenate the two VNR segments together in order to have an end-to-end content delivery infrastructure. One typical example of such a binding is to enforce a CSP's own end-to-end content routing protocol across the unified virtual platform. By means of efficient VNR binding, end-to-end QoS can be engineered on the SP side for supporting corresponding services.

Take Fig. 4a as an example. If we assume that the edge-to-edge VNR delay offered from the two NPs is 50 ms each, the CSP may enforce its own path selection on top of the global virtual platform and determine the targeted end-to-end delay performance, say 100 ms. It is worth mentioning that the outcome of the VNR binding operations might not always be simply the “sum” of the concatenated VNR capabilities from individual NPs. Specifically, SPs have a high degree of flexibility in enforcing their own routing or forwarding policies to partially override the routing/forwarding decisions in the underlying network. For instance, if an SP needs to achieve load balancing over its own virtual network platform with dedicated bandwidth allocation from the NPs, it may apply intelligent virtual traffic engineering (TE) schemes, for example, to avoid local congestion by enforcing possibly longer virtual paths as compared to the direct concatenation of underlay paths. Mobility support across multiple administrative networks is another aspect to be considered in VNR binding operations. For instance, when an end user is consuming live content while moving from one physical NP to another (e.g., across two mobile operator networks), the end-to-end virtual network platform should make sure the service is not disrupted when the user is moving across network boundaries.

**VNR Allocation**

From the viewpoint of NPs, the ultimate goal of VNR allocation is to achieve maximum VNR service capability for maximizing their revenues while maintaining optimized network performances. Given the existence of heterogeneous SPs that cover a wide variety of Internet-based services, a key challenge faced by NPs is how to systematically provision different types of VNRs (bandwidth, spectrum, in-network storage, etc.) for supporting all of them. As shown in Fig. 4b, SP 1 only requires virtual communication infrastructure (e.g., virtual paths), while SP2 also needs other types of VNRs such as virtual network storage. This requires the underlying NP to appropriately provision diverse VNRs accordingly for coping with different types of requests.

Table 1. VNR renting concerns by different types of SPs with heterogeneous requirements.

<table>
<thead>
<tr>
<th>Types of VNRs</th>
<th>VNRs concerned by VSP</th>
<th>VNRs concerned by CSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virtual Nodes</td>
<td>Only virtual communication infrastructure</td>
<td>Virtual communication infrastructure</td>
</tr>
<tr>
<td>Links/Paths</td>
<td>virtual nodes/links/paths in most cases</td>
<td>Virtual in-network content storage capability</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Virtual in-network content adaptation capability</td>
</tr>
<tr>
<td>VNR QoS</td>
<td>Bounded edge-to-edge delay (&lt; 150 ms) and jitter</td>
<td>Stringent bandwidth support for video content delivery</td>
</tr>
<tr>
<td></td>
<td>Bidirectional QoS requirements for supporting interactive VoIP services</td>
<td>Mainly QoS assurance in the direction from content sources to individual consumers</td>
</tr>
<tr>
<td>VNR reliability</td>
<td>Stringent demand on communication network reliability</td>
<td>Stringent demand on communication network reliability</td>
</tr>
<tr>
<td></td>
<td>Five 9’s availability for ensuring end-to-end conversations</td>
<td>More tolerant on storage reliability due to availability of multiple content replicas</td>
</tr>
</tbody>
</table>

Table 1. VNR renting concerns by different types of SPs with heterogeneous requirements.
backup resource capabilities (e.g., alternative fast reroute [13] paths with bandwidth support) to be activated upon failures. It is certainly desirable to have shared backup resources that are able to simultaneously provide protections to the affected VNRs requested by multiple SPs [14].

Summary and Future Research Directions
In this article we have presented a holistic management architecture for providing necessary guidance to virtual network resource management on both the SP side and the NP side. The proposed two-dimensional framework sheds light on how virtual network resources should be collaboratively managed by distinct stakeholders in the Internet business market. From the SP’s point of view, the key issue is to how to optimally rent virtualized resources and horizontally bind them together in order to form an end-to-end service platform for supporting the deployed services. On the other hand, NPs are responsible for independently managing own network resources, logically partitioning them, and allocating to heterogeneous SP providers. Such vertical resource management at the NP side offers a highly flexible solution to the provisioning of physical resources for supporting a wide variety of networked services on top of the physical network infrastructure.

As our future work, we will validate the proposed two-dimensional architecture for VNR management on top of realistic virtual network platforms such as Dragon-Lab (AS number 24575) [15], which offers an ideal experimental environment at the size of a real NP network. Meanwhile, we have also envisaged a set of future research directions concerning VNR management at both the SP and NP side. From the viewpoint of SPs, a general open issue is how to holistically provision customized VNR platforms in a cost-efficient manner, mainly through joint optimization in VNR renting, virtual topology control and end-to-end service delivery path control. For VNR provisioning by CSPs, it is also interesting to investigate how content caching can be optimally achieved within their virtual network platforms. On the NP side, flexibility in VNR allocation to heterogeneous requesting SPs is an essential issue to be concerned. “Soft” VNR allocation strategies are desired, in which case VNR capabilities (e.g., bandwidth support) are not statically allocated, but instead coexisting SP virtual networks can adaptively share some common network capabilities in dynamic environments. Energy efficiency on the NP side will be another research issue to be investigated. The objective is to strategically reduce the physical network capability (through link/node sleeping or rate adaptation) when the end-user demand on top of the SP virtual networks are relatively low, for instance, during off-peak hours. In this case, the underlying NP may optimally perform sleeping or rate adaptation configurations for energy saving, but without disrupting the service performance on the SP side. Last but not least, although security aspects have not been particularly discussed in this article, it is certainly a vital issue to be addressed given the open virtual environment that allows resource sharing between different parties/applications [17]. It will certainly be our future research work to consider security and access control mechanisms on the NP side against illegitimate access of network resources and potential denial of service attacks from malicious entities.

Acknowledgement
This work is partially funded by the EU FP7 IRSES EVANS project (Grant Agreement Number PIRSES-GA-2010-269323).

References
Biographies

NING WANG (n.wang@surrey.ac.uk) is a lecturer at the Centre for Communication Systems Research (CCSR), University of Surrey, United Kingdom. He received his B.Eng. (Honors) degree from Changchun University of Science and Technology, P.R. China, in 1996, his M.Eng. degree from Nanyang Technological University, Singapore, in 2000, and his Ph.D. degree from the University of Surrey in 2004. He is the technical coordinator of the EVANS project. His research interests include virtual and overlay networks, information-centric networking, and QoS/energy-aware network optimizations.

YAN ZHANG (yanzhang@ieee.org) received a Ph.D. degree from Nanyang Technological University, Singapore. He is working with Simula Research Laboratory, Norway, and he is an adjunct associate professor at the University of Oslo, Norway. He is an associate editor or guest editor of a number of international journals. He serves as organizing committee chair for many international conferences. His current research interests include wireless networks, cognitive radio and smart grid.

JOAN SERRAT-FEMANDEZ (serrat@tsc.upc.edu) received his degree of Telecommunication Engineer in 1977, and his doctoral degree in telecommunication engineering in 1983, both from Universitat Politècnica de Catalunya (UPC), Barcelona, Spain. Currently he is a full professor at UPC where he has been involved in several collaborative projects with different European research groups, both through bilateral agreements or through participation in European funded projects. His topics of interest are in the field of autonomic network and service management.

JUAN-LUIS GORRICO (juanluis@entel.upc.edu) is an associate professor at UPC. He received his M.S. degree in telecommunications engineering in 1994 from UPC, and his Ph.D. degree in 1999 from the same university. He is an active member of the Management, Pricing and Services for Next Generation Networks research group (MAPS) in the Telematics Department of UPC. Presently he is focusing his research in the fields of network management, context awareness, and reinforcement learning agents.

TAO GUO (t.guo@surrey.ac.uk) received a Ph.D. degree in wireless networks from Newcastle University, United Kingdom, in 2009. He is currently a research fellow at the Centre for Communication Systems Research, University of Surrey. He has been working on various UK EPSRC and EU FP7 projects such as BeFEMTO, QoSMOs, and Mobile VCE Core 5 “Flexible Networks.” His current research interests include mobility and radio resource management, and network virtualization.

ZHENG HU (huzheng@bupt.edu.cn) received his Ph.D. degree in 2008 from Beijing University of Posts and Telecommunications (BUPT). Currently, he is working with BUPT, and his research interests lie in the next generation wireless network and services focusing on resource management and network control. He is also coordinating and participating several research projects such as Chinese Key program, 973 program, and 863 program.

PING ZHANG (pzh@bupt.edu.cn) received his Ph.D. degree in 1990 from BUPT. Currently, he is a professor in BUPT and director of the Key Laboratory of Universal Wireless Communications, Ministry of Education. He is Executive Associate Editor-in-Chief on information sciences of the Chinese Science Bulletin, a member of next-generation broadband wireless communication network in the National Science and Technology Major Project committee, a Member of the 5th Advisory Committee of National Natural Science Foundation of China (NSFC), and Chief Scientist of the “973” National Basic Research Program of China. His current research interests include wireless communication, cognitive radio networks, and next generation wireless networks.