Chapter 104
The Cloud Inside the Network: A Virtualization Approach to Resource Allocation

João Soares
University of Aveiro, Portugal & Portugal Telecom Inovação, Portugal

Márcio Melo
University of Aveiro, Portugal & Portugal Telecom Inovação, Portugal

Romeu Monteiro
University of Aveiro, Portugal

Susana Sargento
University of Aveiro, Portugal & Instituto de Telecomunicações, Portugal

Jorge Carapinha
Portugal Telecom Inovação, Portugal

ABSTRACT

The access infrastructure to the cloud is usually a major drawback that limits the uptake of cloud services. Attention has turned to rethinking a new architectural deployment of the overall cloud service delivery. In this chapter, the authors argue that it is not sufficient to integrate the cloud domain with the operator's network domain based on the current models. They envision a full integration of cloud and network, where cloud resources are no longer confined to a data center but are spread throughout the network and owned by the network operator. In such an environment, challenges arise at different levels, such as in resource management, where both cloud and network resources need to be managed in an integrated approach. The authors particularly address the resource allocation problem through joint virtualization of network and cloud resources by studying and comparing an Integer Linear Programming formulation and a heuristic algorithm.

INTRODUCTION

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction (Mell & Grance, 2011).

DOI: 10.4018/978-1-4666-6539-2.ch104

This is part of one of the most cited cloud computing definitions, defined by the United States National Institute of Standards and Technology (NIST). It clearly states that network is an inherent component of the cloud, not only as a mean of access to other cloud resources, but also as a resource itself. Although a definition would not be necessary to confirm this, it is interesting to highlight it, since in the cloud early stages the network component of the cloud has been neglected to a large extent. On the other hand, its importance is highly recognized today because of its fundamental role in guaranteeing performance, reliability, and security.

In today's network scenarios, the network component of the cloud has implications at two different levels: Data Center (DC) and Wide Area Network (WAN). Depending on the type of service, different Quality of Service (QoS) guarantees are required both on the DC and in the WAN. Moreover, scalability and elasticity of the cloud may suggest variations on the necessary network resources as the cloud scales up or down. However, from an administrative standpoint, DCs and WANs (which are in practice operator networks) are completely different entities, which do not cooperate on an active basis, and consequently the access to cloud services is typically done over best effort Internet.

The lack of cooperation between cloud and WAN represents a major drawback that has limited the uptake of cloud services. The current best-effort support for many cloud services is not enough as an increasingly large number of services cannot be handled in this way (e.g., Netflix, OnLive). Furthermore, looking at the enterprise market sector, network reliability is a "must have," not only from a performance perspective but also from a security one. In some cases, an independent network service that tries to fulfil the cloud service requirements can be purchased, backed up by a Service Level Agreement (SLA), connecting the user and the cloud hosting the service. This typically happens in the enterprise sector, namely through operator-managed Virtual Private Network (VPN) service models, such as Border Gateway Protocol (BGP)/Multiprotocol Label Switching (MPLS) Internet Protocol (IP) VPN (Rosen & Rekhter, 2006) or Virtual Private LAN Service (VPLS) (Kompella & Rekhter, 2007; Lasserre & Kompella, 2007). There is no reason to believe that future cloud services will require a lesser degree of reliability and performance guarantees from the network.

However, the traditional VPN model is not able to handle essential cloud properties such as elasticity and self-provisioning, which means that those properties should be also extended to network resources. Quite often, expanding or reducing cloud resource capacity, or provisioning new cloud resources, requires a corresponding reconfiguration of network resources, e.g., bandwidth assigned between two data centers, whether they are in the same geographical region or not, or between the data center and the end user. Today, the reconfiguration of network services is supposed to happen on a relatively infrequent basis and usually involves a significant amount of manual effort. In order to cope with the cloud, future network services will certainly require on-demand and self-provisioning properties. This will be the basis for an active participation of the network in the cloud computing service delivery.

Moreover, the dynamism of the cloud will often require live migration of resources (e.g., from a local enterprise data center to the cloud, or between two different sites of the cloud service provider) without interrupting the operating system or making any noticeable impact on the running application. This requires IP addressing to remain unchanged after migration, and all relevant QoS, security and traffic policies applied on network equipment (e.g., routers, switches, firewalls) to be adapted appropriately in real time.

Lately, as a result of the above mentioned aspects, attention has turned to rethinking architectural deployment of the overall cloud service delivery (Akamai, 2011; Cisco, 2009), where