End-to-End Service Quality for Cloud Applications

Karsten Oberle, Davide Cherubini, Tommaso Cucinotta

Bell Laboratories, Alcatel-Lucent
firstname.lastname@alcatel-lucent.com

Abstract This paper aims to highlight the importance of end-to-end service quality for cloud services, with a focus on telecom carrier-grade services. In multi-tenant distributed and virtualized cloud infrastructures, the enhanced level of resource sharing raises issues in terms of performance stability and reliability of cloud services, threatening the possibility to offer precise service levels in end-to-end scenarios.

Technology-wise, we have today some basic building blocks that may enable cloud infrastructures to exhibit stable and predictable performance to customers. However, one of the major obstacles that keep hindering the potential for a worldwide deployment of these technologies, is the fact that, in many distributed and cloud computing scenarios, there is not merely a single business entity responsible for the service delivery, but we may have multiple different, unrelated business entities with contrasting and competing requirements, interacting for the provisioning of end-to-end cloud services to customers and finally end users. It is the case, for example, of multiple cloud providers, storage providers and network service providers, that may be involved for the delivery of a distributed cloud service to a community of end users.

In this context, setting up proper Service-Level Agreements (SLAs) among the involved players, for delivering strong QoS guarantees to customers, may become overly challenging. However, the main problems arising in such interactions may be mitigated by a thoughtful intervention of standardization.

The paper reviews some of the most important efforts in research and industry to tackle end-to-end service quality and concludes with some recommendations for additional required research and/or standardization effort required to be able to deploy mission critical or interactive real-time services with high demands on service quality, reliability and predictability on cloud platforms.

1 Introduction

Over the last few years virtualization and Cloud Computing technology found their commercial success (e.g. Amazon EC2). This is surely tightly connected with the continuous and steep evolution Information and Communication Technologies (ICT) have been recently undergoing. The wide availability of high-speed network connections is causing an inescapable shift towards distributed computing models where processing and storage of data can be performed mostly in cloud computing data centers.

Cloud Computing introduces a novel model of computing that brings several technological and business advantages: customers (a.k.a., *tenants*) can rent cloud services on a pay-per-use model, without the need for big investments for resources that have to be designed for peak workloads, whilst being at risk of remaining under-utilized for most of the time; providers may offer cloud services for rental, hosting them on big multi/many-core machines, where the big infrastructure investments may easily be amortized over hundreds or thousands of customers.

This *multi-tenancy* nature of cloud infrastructures constitutes one of the major levers over which a high level of efficiency in the management of the hosted services may be achieved. Indeed, by recurring to virtualization technologies, which allow for easy and seamless migration of virtual machines (VMs) among physical hosts, a provider may manage the physical infrastructure with a high efficiency. Physical resources may easily be shared among multiple tenants whenever appropriate.

Unfortunately, this enhanced level of resource sharing brings a number of disadvantages and challenges as well. Sharing the physical infrastructure leads to an increased level of temporal interference among the hosted services. As a consequence, one of the critical issues emerging in cloud infrastructures is the stability in the performance level of the hosted services.

Cloud providers are not the only ones to which the nowadays observable unstable and unreliable performance of cloud services should be attributed. As it is well known, the Internet, over which most of the cloud offerings are accessible nowadays, is entirely designed and deployed according to best-effort paradigms. Indeed, the Internet has always been multi-tenant by its nature.

However, the requirements of cloud customers are very likely to evolve quickly, as cloud technology is being more and more known and used worldwide. Many enterprise applications that might take tremendous advantages from the cloud model cannot be hosted on nowadays infrastructures due to their stringent performance requirements that cannot be met in nowadays cloud computing infrastructures, accessible and interconnected through the best effort Internet. Think of virtual desktop, Network Function Virtualization (NFV), professional on-line multimedia editing and collaborative tools, on-line gaming, just to mention a few.

Furthermore, virtualization is becoming increasingly interesting for telecom operators (a.k.a. *telcos*) who are increasingly willing to switch from hardware-based to software-based solutions.

Some of the world leading telecom operators have initiated [2] in early 2013 a new standards group for virtualization of network functions at ETSI [3]. Aim is to transform the way network operators architect networks by evolving standard

IT virtualization technology to consolidate main network equipment types onto industry-standardized high-volume servers, switches and storage, which could be located in data centers, network nodes and in end-user premises [2]. This potentially offers some benefits, such as:

- Reduced CAPEX, lowering equipment cost
- Reduced OPEX
- Reduced time to market for new telecom services
- Increased scalability
- Reduce entry level/barrier for new players, and geographically targeted services
- Multi tenancy, multi user, multi services, telecom/network operator resource sharing/pooling

Virtualization and cloud technologies allow for an unprecedented degree of flexibility [2] in the management of the physical resources. However, they also introduce further variability and unpredictability in the responsiveness and performance of these virtualized network functions, which are often characterized by well-specified service levels (i.e., reliability and QoS constraints such as latency constraints) that have to be respected. Furthermore, end-to-end service quality is increasing in importance and is paramount for real-time and/or interactive services but especially for carrier grade telecommunication services such as for instance IMS (IP Multimedia Subsystem)[1].

An end-user requesting a service does not really care or need to know if the service requested and consumed is Cloud based or a traditionally hosted one. An end user mainly cares about the price for a service and the expected and received service quality – the *end-to-end service quality.*

This includes several issues, such as End-to-End service availability, End-to-End service performance (e.g. latency, jitter, throughput), End-to-End service reliability, End-to-End service accessibility and End-to-End service retainability. More details about the above issues can be found in [5].

In a Cloud deployment case, the end-to-end service scenario can get quickly very complex in terms of number of actors and providers involved in the end-to-end service delivery chain and hence all the boundaries between, i.e., the horizontal chain including User Equipment, Access Network, Core Network, Data Center and the top-down chain across the various cloud layers from Software-as-a-Service (SaaS) to Infrastructure-as-a-Service (IaaS). The scenario can easily get more complex in case of services spanning across multiple data centers or for instance 3rd party infrastructures involved in the DC (see Section 3 below).

Hence, in order to enable more telecom like applications and services to be run in a distributed cloud environment, networked systems need to become more intelligent and able to support end-to-end QoS by joint optimization across networking, computing and storage resources.

[1] More information is available at: `http://www.3gpp.org/Technologies/Keywords-Acronyms/article/ims`

In order to provide the required end-to-end service quality for cloud based services, a Service Level Agreement (SLA) framework is required to express the required level of service quality and related Key Quality Indicators (KQIs), to measure, monitor, correct or police, repair and finally to guarantee the required level of service quality, when coupled with proper service engineering practices. A chain of multiple SLAs is required covering the end-to-end scenarios. This results in a complex system of multiple SLAs covering all the boundaries between actors and providers.

Additionally, those SLAs have different levels of technical content as an SLA between an end user and an application service provider might be quite different from an SLA between a Cloud Service Provider (CSP) and a Network Service Provider (NSP).

1.1 Proposition

This paper aims to highlight the importance of end-to-end service quality for cloud services especially for the case of telecom carrier grade services. We will mainly focus on the multi-tenancy aspects (as this enhanced level of resource sharing raises some issues in terms of stability and reliability of cloud services) as well as the area of Service Level Agreements for end-to-end scenarios.

Technology-wise, we have today some basic building blocks that may enable cloud infrastructures to exhibit stable and predictable performance to customers. Indeed, on the side of network provisioning, standards exist enabling the possibility to provide connectivity with end-to-end QoS guarantees, such as IntServ [12] and DiffServ [11].

Similarly, on the side of computing technologies, platforms for real-time and predictable computing are becoming increasingly accessible, not restricting to the traditional area of real-time and embedded systems, but recently spreading also over the area of predictable cloud computing [15] [29].

However, one of the major obstacles that keeps hindering the potential for a worldwide deployment of these technologies and especially for telecom services, is the fact that, in many distributed and cloud computing scenarios, there is not merely a single business entity responsible for the service delivery. Instead, we may have multiple different, unrelated business entities with contrasting and competing requirements, interacting for the provisioning of end-to-end cloud services to customers and finally end users. For example, multiple cloud, storage and network service providers may be involved for the delivery of a distributed cloud service to a community of end users.

In this context, setting up proper SLAs among the involved players for delivering strong QoS guarantees to customers, may become overly challenging. However, the main problems arising in such interactions may be mitigated by proper SLA engineering techniques trying to fragment the overall problem into simpler ones to be tackled separately, when possible, and a thoughtful intervention of standardization.

The next section will present some of the related work existing in those areas followed by some scenarios to explain the potential complexity of actors involved at the present. Finally the paper identifies blank spots of required research and standardization work in this area.

2 Related Work

This section shortly reviews existing standards and research efforts addressing end-to-end Cloud/Network service delivery with QoS considerations. Due to space constraints, not each individual activity in this area can be mentioned.

2.1 Standards

ETSI. ETSI is currently involved in several activities related to the above mentioned issues. Of major importance towards the scope of end-to-end cloud service quality provisioning is the work [3] started by the ETSI NFV Reliability & Availability sub group. A first report of that group is expected for late 2013. ETSI NFV detected the importance of end-to-end considerations and kicked off a Specification document in April 2013 on “NFV End to End Architecture Reference” (Work Item DGS/NFV-0010). A publication of a first version is planned for autumn 2013. However, the issue of service quality for virtualized network functions will be a key issue to work on inside the ETSI NFV activity and will be probably touched on by several working and expert groups of the ETSI NFV Group, such as, e.g., in the “Reliability and Availability WG”. The work of this ETSI NFV consists of providing a pre-standardization study before considering later a broader standards proposal in a new or existing standardization group.

The second related ETSI activity is the Technical Committee (TC) CLOUD which aims to address issues associated with the convergence between IT (Information Technology) and Telecommunications. The focus is on scenarios where connectivity goes beyond the local network. TC CLOUD will also address interoperability aspects of end-to-end applications and develop formal test specifications to support them[2]. The recent related Technical Report from TC CLOUD is TR103125, V1.1.1, “Cloud, SLAs for Cloud Services” aiming to review previous work on SLAs including ETSI guides from TC USER and contributions from EuroCIO members and to derive potential requirements for cloud specific SLA standards. Connected to TC CLOUD is the third ETSI hosted and related activity, the Cloud Standards Coordination (CSC) task[3]. ETSI has been requested by the EC through the European Cloud Strategy [19] to coordinate with stakeholders in the cloud standards ecosystems and devise standard road-maps in support of EU policy in critical areas, such as security, interoperability, data portability, reversibility and SLAs. Especially the subgroup dealing with Cloud SLAs might produce a highly interesting output document in regard to existing SLA standards when looking on use cases demanding end-to-end Cloud service quality.

[2] More information at: `http://portal.etsi.org/portal/server.pt/community/CLOUD/310`

[3] More information at: `http://csc.etsi.org/website/private_home.aspx`

The final report towards the European Commission is expected for autumn 2013.

NIST. The Cloud Computing Group of the National Institute of Standards and Technology (NIST) has published and is currently working on a series of reports being of value to the topic of end-to-end cloud service quality[4].

The NIST Cloud Computing Reference Architecture [20] contains a reference architecture widely used by industry, also introducing actors such as the "Cloud Broker", which might play a major role in the end-to-end cloud service delivery chain.

NIST Special Publication [21], in "Requirement 3: Technical Specifications for High-Quality Service-Level Agreements", highlights already the importance of how to define reliability and how to measure it. This is amplified by "Requirement 10: Defined & implemented Cloud Service Metrics" on the industry need for standardized Cloud Service Metrics.

NIST took this already to the next level and is especially addressing those two requirements in the NIST Cloud Computing Reference Architecture and Taxonomy Working Group (RATax WG) [22], in addition to other works on SLA taxonomy and Cloud Metrics [23].

Finally, in an updated Version 2 of Special Publication 500-291 "NIST Cloud Computing Standards Roadmap", which is currently undergoing internal review and approval process, NIST is also investigating on cloud Standards for Service Agreements. However, regarding end-to-end service quality, the document refers to considerations done recently by the TM Forum – more details on that in the next paragraph.

TMF. The Tele Management Forum (TM Forum) has started recently some effort on Multi Cloud Management which is potentially of high importance for the end-to-end cloud service quality topic.

TM Forum has created a set of business and developer tools to help service providers and all players in the multi-cloud value chain implement and manage services that span across multiple partners. Organized as "packs", these initial tools focus on managing SLAs between partners [24].

Document TR178 [30] is a good starting point into that topic as this technical report takes a wider view considering also related existing work at e.g. DMTF, OGF, NIST, ITU-T, OASIS and other TMF related activities.

The report recommends a set of business considerations and architecture design principles that are required to support end-to-end Cloud SLA Management with the aim to facilitate discussion regarding SLA consistency across Cloud Deployment Models and Services Models. TMF is currently planning the work on a version 2 of that document until late 2013 in order to add especially a section related to Cloud Metrics and Measurements. Furthermore, TM Forum started to work on several Multi-Cloud Service Management Reports (TR194-TR197) which are yet not finalized and published. Looking at the work started it appears that this work is essential to follow and potentially extend when reasoning about

[4] More information can be found at: `http://www.nist.gov/itl/cloud/index.cfm`

end-to-end cloud service quality matters. Some of the highlighted points will be also reflected in Section 4.

OGF. The Open Grid Forum (OGF) developed two Web Services (WS) Agreement Specifications. First, the GFD-R.192 WS Agreement Specification [45], a protocol for establishing agreement between two WS parties, such as between a service provider and consumer. And second, the GFD-R-P.193 WS-Agreement Negotiation specification [46], a protocol for multi-round negotiation of an agreement between two parties, such as between a service provider and consumer which works on top of WS-Agreement.

Furthermore, OGF started the Open Cloud Computing Interface (OCCI) working group[5], aiming to realize a set of open specifications, protocols and APIs [40][39][43] for enhancing interoperability across various implementations related to the management of cloud infrastructures and services. Projects aiming to provide an implementation of the OCCI specifications include the well-known OpenStack[6] and OpenNebula[7]. The currently available specifications are GFD.183 OCCI Core [40], GFD.184 OCCI Infrastructure [39] and GFD.185 OCCI RESTful HTTP Rendering [43].

2.2 Research

IRMOS. The IRMOS European Project[8] has investigated on how to enhance execution of real-time multimedia applications in distributed Service Oriented Infrastructures and virtualized Cloud infrastructures. One of the core components developed in IRMOS is the Intelligent Service-Oriented Networking Infrastructure (ISONI) [8] [9]. It acts as a Cloud Computing IaaS provider for the IRMOS framework, managing (and virtualizing) a set of physical computing, networking and storage resources available within a provider domain. One of the key innovations introduced by ISONI is its capability to ensure guaranteed levels of resource allocation for individual hosted applications. In ISONI, each distributed application is specified by a Virtual Service Network (VSN), a model describing the resource requirements, as well as the overall end-to-end performance constraints. A VSN is a graph whose vertexes represent Application Service Components (ASCs), deployed as VMs, and whose edges represent communications among them. In order for the system represented by a VSN to comply with real-time constraints as a whole, QoS needs to be supported for all the involved resources, particularly for network links, CPUs and storage resources. To this purpose, VSN elements are associated with precise resource requirements, e.g., in terms of the required computing power for each node and the required networking performance (i.e., bandwidth, latency, jitter) for each link. These requirements are fulfilled thanks to the allocation and admission control

[5] More information is available at: `http://www.opennebula.org/`

[6] More information is available at: `http://www.openstack.org/`

[7] More information is available at: `http://www.opennebula.org/`

[8] More information is available at: `http://www.irmosproject.eu`

logic pursued by ISONI for VM instantiation, and to the low-level mechanisms shortly described in what follows (a comprehensive ISONI overview is out of the scope of this paper and can be found in [8] [9] [4].

Isolation of Computing

In order to provide scheduling guarantees to individual VMs scheduled on the same system, processor and core, IRMOS incorporates a deadline-based real-time scheduler [31] [15] [18] for the Linux kernel. It provides temporal isolation among multiple possibly complex software components, such as entire VMs (with the KVM hypervisor, a VM runs as a Linux process). It uses a variation of the Constant Bandwidth Server (CBS) algorithm [10], based on Earliest Deadline First (EDF), for ensuring that each group of processes/threads is scheduled on the available CPUs for a specified time every VM-specific period.

Isolation of Networking

Isolation of the traffic of independent VMs within ISONI is achieved by a VSN individual virtual address space and by policing the network traffic of each deployed VSN. The two-layer address approach avoids unwanted cross-talk between services sharing physical network links. Mapping individual virtual links onto diverging network paths allows for a higher utilization of the network infrastructure by mixing only compatible traffic classes under similar predictability constraints and by allowing selection of more than just the shortest path. Traffic policing avoids that the network traffic going through the same network elements causes any overload leading to an uncontrolled growth of loss rate, delay and jitter for the network connections of other VSNs. Therefore, bandwidth policing is an essential building block to ensure QoS for the individual virtual links. It is important to highlight that ISONI allows for the specification of the networking requirements in terms of common and technology-neutral traffic characterization parameters, such as the needed guaranteed average and peak bandwidth, latency and jitter. An ISONI transport network adaptation layer abstracts from technology-specific QoS mechanisms of the networks, like Differentiated Services [11], Integrated Services [12] [13] and MPLS [14]. The specified VSN networking requirements are met by choosing the most appropriate transport network, among the available ones. Other interesting results from the research carried out in IRMOS include: algorithms for the optimum placement of distributed virtualized applications with probabilistic end-to-end latency requirements [16]; the use of neural networks for estimating the performance of Virtual Machines execution under different scheduling configurations [18]; techniques for reduced down-time in live-migration of VMs with time-sensitive workloads [37]; and others. The effectiveness of IRMOS/ISONI has been demonstrated for example through an e-Learning demonstrator [15].

SLA

Within IRMOS, an SLA management framework spanning across the three main cloud service models (SaaS, PaaS, IaaS) has been developed, through a combined approach of SLAs with real-time attributes (and QoS attributes in general) according to the needs of the service to be deployed and executed. A set of tools has been developed which support the tasks of the different actors (from application modeling down to resource virtualization) and an SLA life cycle between them. In IRMOS the SLA life cycle is structured in three phases:

- Publication phase
- Negotiation phase
- Execution phase

More details can be found in [25]. This paper also describes in detail the different types of dynamic SLAs among the different actors:

- Application SLA: agreement established between the Client as a business customer and the Application Provider; this SLA contains the high-level QoS parameters of the application required and defined by the Client.
- Technical SLA: agreement negotiated between the PaaS Provider and the IaaS Provider. This agreement contains low-level QoS parameters associated with the infrastructure.

Within the IRMOS project an extensive SLA state of the art analysis has been performed [26] [27] [28] also covering several other EC funded research projects such as RESERVOIR and SLA@SOI.

SLA@SOI The SLA@SOI EU Project[9] developed an open-source framework addressing [41] negotiation, provisioning, monitoring and adaptation of SLAs through the entire cloud service life-cycle. The framework included [42] both functional and non-functional characteristics of services, such as QoS constraints, which can be formalized through an XML-based syntax.

OPTIMIS. The OPTIMIS EU Project [10] investigates on orchestration of cloud services [1] specifically addressing how to deploy intelligently legacy applications based on their preferences and constraints regarding trust, risk, eco-efficiency and cost factors. For example, in [17], a model for optimum allocation of cloud services is presented that considers a mix of trust, risk, eco-efficiency and cost factors in the overall optimization goal. OPTIMIS also investigates on how to properly leverage both private, hybrid, federated and multi cloud environments for services development and deployment.

ETICS. The ETICS (Economics and Technologies for Inter-Carrier Services) European Project investigated on the criticalities for the creation of a new ecosystem of innovative QoS-enabled interconnection models between Network Service Providers (NSPs) impacting all of the actors involved in the end-to-end service delivery value-chain. ETICS investigated on novel network control, management

[9] More information is available at: `http://sla-at-soi.eu/`

[10] More information is available at: `http://www.optimis-project.eu`

and service plane technologies for the automated end-to-end QoS-enabled service delivery across heterogeneous carrier networks.

The business models analysis [6] and the overall architecture [7] results from ETICS constitute fundamental building blocks allowing for the construction of management of network Inter-Carrier Service Level Agreements.

EC – Expert Group. In July 2013 an Expert Group on Cloud SLA's of the European Commission published a report on "Cloud Computing Service Level Agreements - Exploitation of Research Results" which provides a very detailed insight and analysis on research results achieved by European and National funded research projects [47].

3 Deployment Scenarios

Provisioning of cloud computing applications and services to end-users requires complex interactions among a number of players and business entities. There exist a nearly unlimited amount of scenarios with increasing number and type of actors, the figure below shows the potential complexity:

The scenario includes for instance:

- One or more Cloud Service Providers (CSPs), including potentially Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS) and Infrastructure-as-a-Service (IaaS) providers.
- One or more Network Service Providers (NSP), including heterogeneous networks such as the Access Network and Core Network NSPs.
- One or more Application Service Providers (ASPs)
- The Cloud Customer and End User, who may be the same or different entities, depending on the context.
- A multitude of heterogeneous user equipment, requiring potentially different access network technologies such as DSL, Wifi, LTE, ...
- And finally a Broker serving as contact point and contractual partner for the customer.

In early cloud deployments, NSPs played merely the role of providing connectivity among data centers and end-users through their communication networks, in a way that is service- and mostly also cloud-agnostic. As a consequence, delivering cloud based applications and services to end-users needs at least interactions among Access Network NSP(s), Core Network NSP(s) and Cloud Service Provider CSP(s). However, traditional data centers heavily centralized within a few geographical locations fall short when constraints on response-times become tight (e.g., real-time applications). Indeed, ensuring predictable and stable QoS levels in such conditions becomes overly challenging and requires carefully thought interactions among all these business entities.

Though, over the last years, such a picture has been undergoing quite a change. On one hand, CSPs have been expanding their presence on the territory by adding more and more data centers across the planet. Even though some of

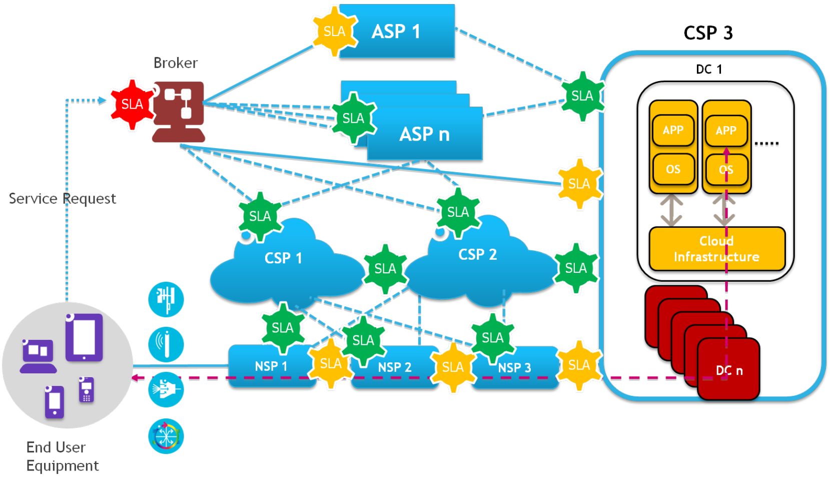

Figure 1. end-to-end scenario(s)

the most successful providers (e.g., Amazon EC2[11]) have still nowadays barely one or two data centers per continent, there are other efforts towards creating way more distributed data center architectures for provisioning of cloud services [33] [34] [32], such as those leveraging "containerized" and modular data center solutions [35] [36].

On the other hand, Telecom operators have been deploying all over the planet their ICT infrastructure in a completely distributed fashion (e.g., think of Access Network NSPs). In a networking world that is heavily shifting from the use of custom hardware boxes towards virtualized network functions realized in software [2], the infrastructures of NSPs is evolving towards more and more general-purpose hardware hosting virtualized software-based solutions, with the need of addressing vertical and horizontal scalability of said solutions which are typical of cloud-based solutions. As a consequence, NSPs are in the unique position of needing to build internally scalable and heavily distributed infrastructures for hosting virtualized network functions, while at the same time being potentially able to reuse such infrastructure for the provisioning of general-purpose cloud services but with a novel, heavily distributed, close-to-the-edge and unprecedented low-latency infrastructure.

Generally speaking, distribution of cloud services so as to get closer to the edge and the end users is a must, while low latency becomes more and more important for users, whose requirements evolve at an amazing speed from needing a mostly storage-only cloud to needing full fledged remote desktop-like solutions.

Moving cloud services closer to the edge mitigates partially the problems for delivering cloud services with stable end-to-end QoS levels. Indeed, when interacting users are geographically close, the variability in the network response

[11] More information is available at: `http://aws.amazon.com/ec2`

is highly reduced, mostly due to the reduction in the number of network segments and NSPs to traverse for closing a single round-trip interaction with the cloud. However, for users distributed across geographically distant locations, and for many cloud applications that already exist nowadays in which the interactions among users spread across an unimaginable number of data items spread all around the globe (e.g., think of collaborative tools such as video-conferencing, shared boards, interactive real-time editing of office documents or mastering of media contents), it is crucial that end-to-end QoS is still guaranteed through appropriate set-up of a properly interacting end-to-end cloud service supply/delivery chain, especially for those services that are to be delivered in a professional way. This requires proper interfaces and standards to allow, for example, the network management infrastructure (e.g., the OSS/BSS) to tie together with cloud management systems (e.g., Cloud Orchestrator), and possibly the existence of Cloud Brokering agents that, analogously to aggregator websites nowadays, are capable of interacting with all these systems to find suitable solutions for customers, matching their needs.

Consider again Figure 1 which clearly shows the potential complexity and especially the large amount of SLAs involved among all the actors. The customer wants to have a single point of contract, meaning one SLA about the service with all characteristics and clearly defined quality metrics. In this set-up the Cloud Broker facilitates meeting of such customer requirement. The Broker then based on customer requirements as expressed in the SLA selects the right ASP as well as CSPs and NSPs in order to fulfill those requirements. This could be done all by the broker or in a more cascaded way. At the end, this whole process results in a large number of SLAs in order to clearly define the accountability between the actors when delivering the contractual defined and required QoS.

End-to-end QoS for cloud services can only be achieved through a careful negotiation of service levels among all the providers, both in the network and in the IT space. Furthermore it is required to have clearly defined quality metrics to monitor and report and finally to trigger countermeasures in case of SLA violation always with the overall target to keep the end-to-end service quality as required.

4 Conclusion and outlook

End-to-end service quality for cloud services is heavily depending on SLA handling in a multi-provider and multi-vendor setup, coupled with proper resource management strategies in a challenging environment with heterogeneous and potentially widely distributed resources. A major challenge for the management of end-to-end Cloud SLAs is the aggregation of individual SLAs across the vertical and horizontal end-to-end path with all their related metrics and KPIs (main metric of interest for the Service Provider)/KQIs (main metric of interest for the customer). TMF, as indicated above, started some work within the Multi-Cloud Service Management Activity which required further work especially regarding the integration/stacking of multiple SLAs. Furthermore additional research and/or standardization effort is required, e.g., to:

- Define clear and meaningful metrics for all the different types of resources as well as reporting schemes and APIs between the multitude of vendors and providers. Work has been started on this at QuEST EB9 Group, TMF and NIST at least. SLA metrics require appropriate definition and categorization to align with expressed SLA objectives as well as to detect, react and specify consequences when those are not met. There will be no real SLA management and hence no deployment for mission critical or interactive real-time services without crystal clear defined metrics and the definition of how to measure, report and manage them.
- Get a more automated SLA management, as required to develop machine readable SLAs in order to achieve faster provider discovery, comparison and monitoring of service quality (see also related recommendations in [44], page 60, Section 6 Federation).
- Further the very complex end-to-end view across all the horizontal and vertical layers and actors, in order to ensure not just service quality but also issues like security and accountability for cloud based services (see also related recommendations in [44], page 61, Section 7 Programmability & Usability and page 63, Section 9 Security).
- Design and engineer proper resource management and scheduling frameworks for cloud computing infrastructures, enabling the possibility to ensure proper levels of temporal isolation among VMs deployed by independent customers (see also related recommendations in [44], page 60, Section 5 Multiple Tenants).
- With the expected quick increase in number of available cloud data center locations across the planet, it will become more and more challenging to properly/optimally place but especially to dynamically relocate applications, VMs, data, across one or more cloud infrastructures, in order to achieve desired and desirable trade-offs among efficiency in management of the infrastructure and users' quality of experience and expectations; more research on scalable, adaptive resource management policies, coupled with agile software infrastructures, is needed for handling the cloud computing scenarios of tomorrow (see also related recommendations in [44], page 60, Section 6. Federation and page 61, Section 7 Programmability & Usability).
- Deal with energy efficiency, a critical issue that needs to be addressed at all levels of computing, from industrial deployments to research, and from hardware to software; designing SLAs containing QoS constraints, but at the same time capable of leaving a degree of flexibility to the CSP or other involved entities enabling more energy-efficient management of resources, need to be further investigated (see also related recommendations in [44], page 61, Section 7 Programmability & Usability).
- Tomorrow cloud applications will make more and more use of massive amounts of data, and normal users of cloud applications will expect/pretend that they can query amazingly huge data sets in one instant; resource management and scheduling for meeting QoS constraints and providing temporal

isolation in presence of “big-data” types of workloads presents a set of novel challenges that have to be urgently addressed by research in the domain of cloud computing and virtualized infrastructures (see also related recommendations in [44], page 56, Section 1 Data Management).

As a final concluding remark, we highlighted in this paper some of the most important efforts in research and industry to tackle end-to-end service quality, but there is still significant work ahead in order to be able to deploy mission critical or interactive real-time services with high demands on service quality, reliability and predictability on cloud platforms.

References

[1] C. Zsigri, A. J. Ferrer, O. Barreto, R. Sirvent, J. Guitart, S. Nair, C. Sheridan, K. Djemame, E. Elmroth, J. Tordsson: Why Use OPTIMIS? Build and Run Services in the Most Suitable Cloud Venues. OPTIMIS Whitepaper. October 2012.

[2] M. Chiosi et al.: Network Functions Virtualisation – Introductory White Paper, SDN and OpenFlow World Congress. October 22–24 2012, Darmstadt, Germany. Available at: `http://portal.etsi.org/NFV/NFV_White_Paper.pdf`

[3] ETSI NFV Portal: `http://portal.etsi.org/portal/server.pt/community/NFV/367`

[4] T. Voith, M. Stein, K. Oberle: Quality of service provisioning for distributed data center inter-connectivity enabled by network virtualization. Elsevier Journal on Future Generation Computer Systems 28 (2012) 554-562.

[5] E. Bauer and R. Adams: Service Quality of Cloud Based Applications. Wiley-IEEE Press, December 2013 (tentative publication date), ISBN 9781118763292

[6] M. Dramitinos and C. Kalogiros: Final business models analysis. January 2013.

[7] P. Zwickl, H. Weisgrab: Final ETICS architecture and functional entities high level design. February 2013.

[8] T. Voith, M. Kessler, K. Oberle, D. Lamp, A. Cuevas, P. Mandic, A. Reifert: ISONI Whitepaper v2.0. July 2009.

[9] K. Oberle, M. Kessler, M. Stein, T. Voith, D. Lamp, S. Berger: Network virtualization: The missing piece. In Proceedings of the 13^{th} International Conference on Intelligence in Next Generation Networks, pp. 1–6. October 2009.

[10] L. Abeni , G. Buttazzo: Integrating Multimedia Applications in Hard Real-Time Systems. In Proceedings of the 19^{th} IEEE Real-Time Systems Symposium, Madrid, Spain, December 1998.

[11] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss: RFC2475 – An Architecture for Differentiated Service. IETF, December 1998.

[12] J. Wroclawski: RFC 2210 – The Use of RSVP with IETF Integrated Services. IETF, September 1997.

[13] J. Wroclawski: RFC2211 – Specification of the Controlled Load Quality of Service. IETF, September 1997.

[14] E. Rosen, A. Viswanathan, R. Callon: RFC3031 – Multi-protocol Label Switching Architecture. IETF, January 2001.

[15] T. Cucinotta, F. Checconi, G. Kousiouris, D. Kyriazis, T. Varvarigou, A. Mazzetti, Z. Zlatev, J. Papay, M. Boniface, S. Berger, D. Lamp, T. Voith, M. Stein: Virtualised e-Learning with Real-Time Guarantees on the IRMOS Platform. In Proceedings of the IEEE International Conference on Service-Oriented Computing and Applications (SOCA 2010), Perth, Australia, December 2010.

[16] K. Konstanteli, T. Cucinotta, T. Varvarigou: Optimum Allocation of Distributed Service Workflows with Probabilistic Real-Time Guarantees. Springer Service Oriented Computing and Applications, Vol. 4, No. 4, December 2010.

[17] K. Konstanteli, T. Cucinotta, K. Psychas, T. Varvarigou: Admission control for elastic cloud services. In Proceedings of the IEEE 5^{th} International Conference on Cloud Computing (CLOUD 2012), pp. 41–48, Honolulu, Hawaii, USA, June 2012.

[18] G. Kousiouris, T. Cucinotta, T. Varvarigou: The Effects of Scheduling, Workload Type and Consolidation Scenarios on Virtual Machine Performance and their Prediction through Optimized Artificial Neural Networks. Elsevier Journal of Systems & Software (JSS).

[19] European Commission: Unleashing the Potential of Cloud Computing in Europe. COM (2012) 529 final, Brussels, September 2012. http://ec.europa.eu/information_society/activities/cloudcomputing/docs/com/com_cloud.pdf

[20] NIST Cloud Computing Reference Architecture. Special Publication 500-292, September 2011. http://www.nist.gov/customcf/get_pdf.cfm?pub_id=909505

[21] US Government Cloud Computing Technology Roadmap Volume I, Special Publication 500-293, Volume I, November 2011. http://www.nist.gov/itl/cloud/upload/SP_500_293_volumeI-2.pdf

[22] NIST Reference Architecture and Taxonomy Working Group (RATax WG), Wiki. http://collaborate.nist.gov/twiki-cloud-computing/bin/view/CloudComputing/ReferenceArchitectureTaxonomy

[23] Cloud Metrics Sub Group, Wiki. http://collaborate.nist.gov/twiki-cloud-computing/bin/view/CloudComputing/RATax_CloudMetrics

[24] Multi Cloud Management. http://www.tmforum.org/MultiCloudManagement/13928/home.html

[25] G. Gallizo, R. Kübert, G. Katsaros, K. Oberle, K. Satzke, G. Gogouvitis, E. Oliveros: A Service Level Agreement Management Framework for Real-time Applications in Cloud Computing Environments. CloudComp 2010, Barcelona

[26] R. Kübert, G. Gallizo, T. Polychniatis, T. Varvarigou, E. Oliveros, S. C Phillips, K. Oberle: Service Level Agreements for real-time Service Oriented Infrastructures. Achieving Real-Time in Distributed Computing: From Grids to Clouds. IGI Global, May 2011.

[27] G. Katsaros, T. Cucinotta: Programming Interfaces for Realtime and Cloud-based Computing. Achieving Real-Time in Distributed Computing: From Grids to Clouds. IGI Global, July 2011.

[28] E. Oliveros, T. Cucinotta, S. C. Phillips, X. Yang, T. Voith, S. Middleton: Monitoring and Metering in the Cloud. Achieving Real-Time in Distributed Computing: From Grids to Clouds. IGI Global, July 2011.

[29] S. Xi, J. Wilson, C. Lu, C.D. Gill: RT-Xen: Towards Real-time Hypervisor Scheduling in Xen. ACM International Conference on Embedded Software (EMSOFT), October 2011.

[30] TR178: Enabling End-to-End Cloud SLA Management. TM Forum.

[31] F. Checconi, T. Cucinotta, D. Faggioli, G. Lipari: Hierarchical Multiprocessor CPU Reservations for the Linux Kernel. In Proceedings of the 5^{th} International Workshop on Operating Systems Platforms for Embedded Real-Time Applications (OSPERT 2009), Dublin, Ireland, June 2009.

[32] M. Alicherry, T. V. Lakshman: Network aware resource allocation in distributed clouds. Proceedings of INFOCOM 2012, pp. 963-971, March 2012, Orlando, FL.

[33] V. Valancius, N. Laoutaris, L. Massouli, C. Diot, P. Rodriguez: Greening the Internet with Nano Data Centers. In Proceedings of the 5^{th} international conference

on Emerging Networking Experiments and Technologies (CoNEXT'09), 2009, pp.37–48. ACM, New York, NY, USA.

[34] K. Church, A. Greenberg, J. Hamilton: On Delivering Embarrassingly Distributed Cloud Services. Hotnets, October 2008, Calgary, CA.

[35] IBM Global Technology Services – Case Study. A containerized IT solution increases Port of Fos-sur-Mer efficiency. March 2012.

[36] IBM Global Technology Services. Columbia County builds a scalable modular data center to improve availability, doubling IT capacity while leaving the same energy footprint. April 2010.

[37] F. Checconi, T. Cucinotta, M. Stein: Real-Time Issues in Live Migration of Virtual Machines. In Proceedings of the 4^{th} Workshop on Virtualization and High-Performance Cloud Computing (VHPC 2009), Delft, The Netherlands, August 2009.

[38] Thijs Metsch: Open Cloud Computing Interface - Use cases and requirements for a Cloud API. Open Grid Forum, 2009.

[39] T. Metsch, A. Edmonds: GFD-P-R.184 Open Cloud Computing Interface - Infrastructure. Open Grid Forum, June 2011.

[40] R. Nyren, A. Edmonds, A. Papasyrou, T. Metsch: GFD-P-R.183 Open Cloud Computing Interface – Core. Open Grid Forum, June 2011.

[41] P. Wieder, J.M. Butler, W. Theilmann, R. Yahyapour: Service Level Agreements for Cloud Computing. Springer, 2011.

[42] W. Theilmann, J. Lambea, F. Brosch, S. Guinea, P. Chronz, F. Torelli, J. Kennedy, M. Nolan, G. Zacco, G. Spanoudakis, M. Stopar, G. Armellin: SLA@SOI Final Report. September 2011.

[43] T. Metsch, A. Edmonds: GFD-P-R.185 Open Cloud Computing Interface – RESTful HTTP Rendering. Open Grid Forum, June 2011.

[44] L. Schubert, K. Jeffery: EC Cloud Expert Group Report. Advances in Clouds – Research in Future Cloud Computing. `http://cordis.europa.eu/fp7/ict/ssai/docs/future-cc-2may-finalreport-experts.pdf`

[45] A. Andrieux, K. Czajkowski, A. Dan, K. Keahey, H. Ludwig, T. Nakata, J. Pruyne, J. Rofrano, S. Tuecke, M. Xu: GFD-R.192 Web Services Agreement Specification (WS–Agreement). Open Grid Forum, October 2011.

[46] O. Waeldrich, D. Battré, F. Brazier, K. Clark, M. Oey, A. Papaspyrou, P. Wieder, W. Ziegler: GFD-R-P.193 WS-Agreement Negotiation. Open Grid Forum, October 2011.

[47] D. Kyriazis: European Commission Directorate General Communications Networks, Content and Technology Unit E2 - Software and Services, CLOUD. Cloud Computing Service Level Agreements: Exploitation of Research Results. `https://ec.europa.eu/digital-agenda/en/news/cloud-computing-service-level-agreements-exploitation-research-results` June 2013.