Understanding Quality Issues in the Cloud

Martín Varela

VTT Technical Research Centre of Finland

martin.varela@vtt.fi

Jouko Sankala

EXFO NetHawk

jouko.sankala@nethawk.exfo.com

November 2010

Executive Summary

With the enormous momentum that Cloud Services adoption has had over the past few years, it seems likely that many services which are critical to both enterprise and end users will be deployed in a Cloud context. The main driver for this migration is related to the economic advantages of Cloud based services compared to traditional models of hosting. However, as more critical services migrate to the Cloud, it becomes clear that the currently existing Quality of Service (QoS) guarantees being offered by Cloud providers are not sufficient. In this white paper we explore what the main QoS issues are today, their relation to the Quality of Experience (QoE) perceived by the users, how they can be improved, and what advantages exist for both providers and customers in the adoption of QoS / QoE monitoring and management technologies.

Contents

Introduction

As with many technological advances in the past, computing has steadily been commoditized over the years. The market penetration of computing devices, and more recently smart mobile devices has grown to the point of saturation in most developed countries.

Already in the late 60's, Leonard Kleinrock, one of the ``fathers'' of what is now the Internet, predicted that computing resources would one day become a commodity just like electricity. In the 80's and 90's several companies realized the importance of networks and their relation to computing services. For example, Sun Microsystems' motto of ``The network is the computer'' reflected this sentiment. In those early visions, the future ubiquitousness of networks, and the Internet in particular, would enable users to roam freely and access their computing environment from virtually anywhere.

Research continued in many areas that showed a clear tendency towards this vision of ubiquitous access to commodity computing resources. Clustering and virtualization technologies were intensely developed in the 90's, as well as the notion of Grid Computing which grew in popularity in the late 90's and first decade of the 21^{st} century, and become what is now known as Infrastructure as a Service (IaaS) and Platform as a Service (PaaS). Application Service Providers (ASPs) grew in popularity over the last decade[1], as well, giving way to Software as a Service (SaaS).

It now seems that the early vision of commodity computing is here, drawing an interesting parallel to the old days of Big Iron and thin clients, only now it is cheaper than could ever be imagined and hugely more flexible. Commodity

[1] A study from the early 2000s reports as many as 70% of respondents not being familiar with the ASP / SaaS concept, and only 6% were implementing it. In contrast, a 2010 Gartner study reports 95% of respondents as planning to maintain or increase their use of SaaS. Sources:
Ma et al. ``An exploratory study into factors of service quality for application service providers'', Information and Management, 2005.
Gartner, "Forecast Analysis: Software as a Service, Worldwide, 2009-2014", 2010

computing has come in the form of Cloud Services, and it is here to stay. With cheap network access available virtually everywhere, and even cheaper computing resources, there is an obvious economic benefit in moving the services which previously ran in users' PCs and corporate data centers to the Cloud. The result is that an ever-increasing number of services are being migrated to the Cloud, and users, both corporate and consumers, are ever more dependent on them.

This dependence on Cloud-based services creates new problems and opportunities for both users and providers. In particular, issues of service reliability and performance arise where before there might have been none, or very different ones. In the following sections we will discuss some of these issues, in particular those concerned with the quality of the Cloud Services, and its impact with respect to service adoption, Service Level Agreements (SLAs), performance monitoring solutions, technological choices, and the potential drawbacks and benefits, for both users and providers, arising from them.

Quality in the Cloud

What is Quality, Anyway?

The first obstacle to providing acceptable QoS for Cloud services is actually understanding what QoS means in this context. In general, the definition of what quality is tends to be a very subjective matter. Different people have different expectations of products and services, based on things such as previous experiences, socio-economical and cultural backgrounds, price paid for the product or service in question, etc.

When referring to communication networks and computing in general, there are technical metrics to define how well a networked service performs. In the most commonly used sense, QoS refers to how fast and reliably, as defined by these kinds of metrics, networks and services perform. For example, measuring the speed of file transfers, the percentage of time that a service is available to its

users, how often failures happen, are all ways of measuring QoS in the traditional sense.

While QoS metrics are well understood for many aspects of Cloud computing, they fail to convey useful information about the users' perception of the services' quality. Cutting edge research is now focusing on this perception of service quality from the users' point of view, which is called Quality of Experience (QoE). QoE can be loosely defined as the overall acceptability of a service, as perceived by the user. The key to understanding QoE is to understand that

- It is subjective, by definition
- It is application-dependent
- It is context-dependent

Understanding QoE requires therefore a multi-disciplinary view, covering technical, user-related, and business aspects, and how they relate for users of a given service. Figure 1 provides a conceptual overview of the factors involved in understanding QoE for Cloud services.

How does Quality Relate to Business in the Cloud?

The adoption of Cloud services has been strongly driven by the low cost option they represent. While there are several technical advantages to deploying these services, there is agreement in the literature on the fact that lower implementation, licensing and support costs are the key factor driving their adoption today. This has also been reflected in a number of recent interviews with Cloud business and technical experts in Finland and Sweden.

Lower cost, while a powerful motivator, is not always enough to attract and convert customers. Enterprise customers, in particular, usually have stringent needs in terms of service reliability, availability, possible migration paths, support, and many other service qualities that go beyond cost. Operators and service

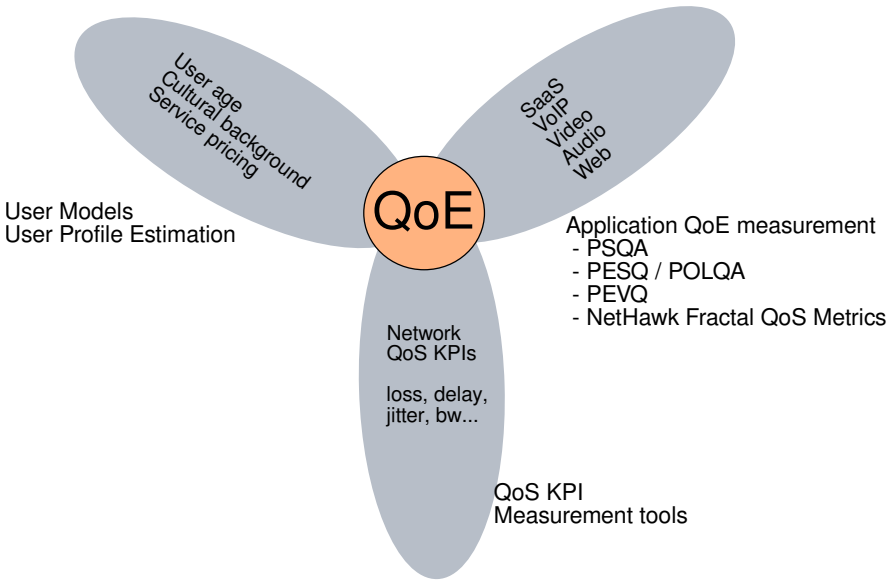

Figure 1: Cloud services QoE Landscape

providers can then differentiate their offering from other, often much cheaper, competitors by providing a higher standard of service.

Among the ways of providing added value to Cloud services, the typical enterprise-like features concerning customer support, service up-times and basic reliability metrics can be implemented with relative ease in a Cloud environment. This, however, might miss an important point concerning the users' experience. As more and more services migrate to the cloud, some services that used to be locally run applications on the users' desktops, or web-applications hosted in the company data center will reside in co-located, virtualized Cloud environments, competing for resources with other service instances running for other customers of the same provider. Ensuring that the *users' experience* remains at least as good as it was before will be one of the key factors in promoting the adoption of new Cloud services, and migrations of older ones to the Cloud.

It is here that understanding how the quality of Cloud services is perceived by users will play a major role in the success or failure of individual services. Given that low prices often make for low margins, service providers can stay above the

competition by providing a better experience for their users, while maintaining their own costs at a minimum.

Enhancing Cloud Offerings via Improved QoE

Service Levels for the Cloud

It is commonplace for service providers to establish Service Level Agreements (SLAs) with their customers, in order to formalize the (minimum) acceptable performance of the services they offer. Depending on what these services are, the SLAs will consider different metrics. For example, for network providers, SLAs typically consider link availability, QoS metrics such as bandwidth and delay, etc. In the case of hosted services, system up-time, guaranteed support times and similar metrics are used.

In a Cloud context, however, a combination of network, computing infrastructure, platform and SaaS-level SLAs need to be in place[2], if the customer wants to be sure that she gets what she is paying for. The problem is that the traditional metrics which are used in the definition of SLAs say next to nothing about how the service will perform from the actual users' point of view. Moreover, in many cases, several providers (SaaS, PaaS, IaaS, network) will be involved, and each might have completely different concepts of what service levels should be in their own domain. None of these, probably, will be close to what the user perceives as the service level when she is actually using the application / service.

The monitoring tools deployed today cannot offer a coherent, QoE-oriented view of service performance. Different tools, for monitoring hardware, networks and services need to be brought together in order to turn today's fragmented view of service quality into a simple, reliable view of the users' perception of the service quality.

[2]This applies mostly to enterprise-class services, but it may very well apply to end-user services, especially if they are subscription-based

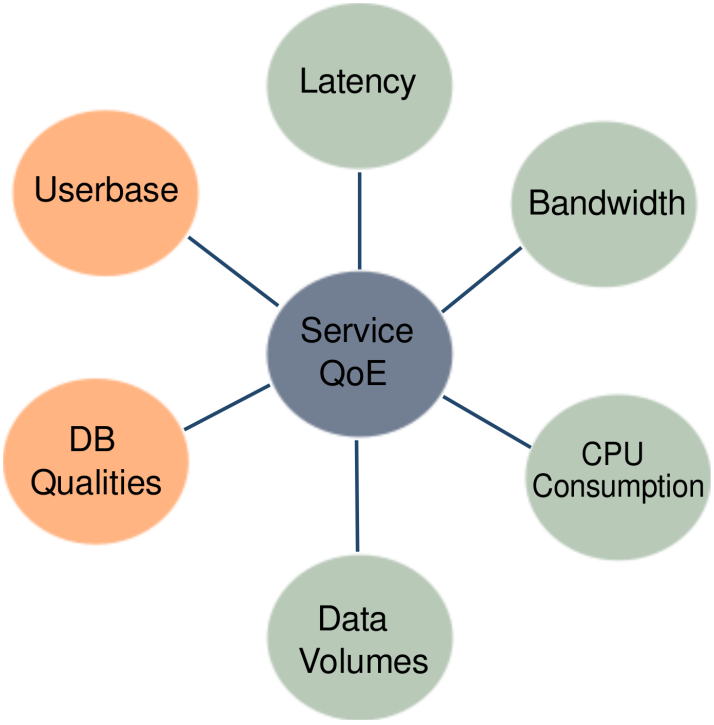

Figure 2: Cloud services can be classified along multiple dimensions, according to their intended use and performance requirements.

Active research is currently being carried out in order to make QoE of Cloud services possible. One of the paths taken to this end, is to provide means for classifying services along multiple dimensions, according to their requirements. Figure 2 provides a high-level view of the main high-level technical factors which influence QoE.

Some of these factors are well understood and enough to provide accurate QoE estimations for some services (for instance, voice and video). Others, however, need careful consideration and study in order to determine the best ways in which they can be measured and used to predict service performance. For example, understanding the impact of data complexity on the performance requirements of services requires good metrics for said complexity. Similar tools are needed to model user loads and behaviours, for example.

Other factors relating to the user experience (UX) should also be taken into account. UI design quality metrics, as well as other measures of usability probably have a rather strong influence on the perceived quality, and therefore should also be accounted for.

Once the technical and UX factors are well understood and proper user models are available, generic QoE metrics will become available for Cloud services. As of now, however, a case-by-case approach is needed to understand quality in a Cloud context.

QoE-based SLAs

With the understanding that users are willing to pay based on the perceived utility of the service (the proverbial ``bang for the buck''), providing them with an objective, verifiable measure of how good quality they will get from the service would set a service provider clearly above the competition. Such a perceptually-grounded metric of service performance would provide users with a SLA that is meaningful to them and to which they can easily relate. For service providers, having such metrics would allow to fine-tune and optimize the underlying infrastructure, in order to deliver acceptable quality while minimizing operating costs.

Experts interviewed in the course of the research leading to the present paper unanimously agreed that being able to provide QoE metrics and QoE-based SLAs would increase the value of the service for the customers, and provide strong differentiation for service providers. They also agreed that there are two caveats to this, namely that

- at present, nobody really knows how to define, let alone measure or control, what QoE is for Cloud services in general. and
- the benefits of having such SLAs in place could easily be offset by the implementation costs.

Both of these considerations are relevant, but fortunately, do not present insurmountable problems.

The Road to QoE in Cloud Services

Quality of Experience has been a hot research topic as of late. It started as an offshoot of QoS research[3], but it stands today as a research field of its own. Many advances, in particular concerning QoE of media services have been made, to the point where it is possible today to produce accurate assessments of the users' perception of the quality of media services in an automated, reliable, and cheap[4] fashion.

It has also been noted that some of the techniques and tools used for media QoE assessment can be, to some extent, adapted for other services. The adaptation is not straightforward, both due to technical issues, and to the fact that media services quality has been studied for a long time and is currently much better understood. Experts in the area, however, believe that it is possible to have working models of QoE for more generic services within 2 to 4 years. To this end, several European research initiatives are currently underway, with the short-term goal of producing functional tools and techniques for

- QoE models for different, generic Cloud services
- QoS to QoE mapping
- Improved monitoring of QoS and QoE in Cloud contexts
- Defining and (re)negotiating high-level, compositional SLAs

[3] And arguably also from the Telecommunications industry. For instance the ITU has defined dozens of recommendations concerning transmission quality for voice, audio and video, some of which are the canon for defining media QoE, and many of the laboratories dealing with QoE issues belong to, or are spin-offs of, telcos.

[4] Both in terms of cost and computational requirements.

The main obstacle to implementing this vision lies in the complex interactions between technical and user-level factors and the perception of service quality, both in understanding them and in providing a suitable, solid formalization to be used as a basis for reliable QoE.

From QoS to QoE

As noted previously, simply having QoS metrics for the components of Cloud services is not enough to understand (and thus predict and control) the QoE perceived by the users. In order to remedy this, multi-disciplinary teams, with experts from different areas such as hardware and service monitoring, network QoS, User Experience, Psychology and other cognitive sciences are working together to create user models capable of providing an understanding of QoE for different services. In particular, the goal is to come up with user models that are generic enough to adapt to virtually any service, and moreover, models which are parametric on several measurable lower level factors (such as network QoS, computational requirements for the services, user base sizes, etc.).

Understanding QoE in terms of the factors that define it is key to being able to make optimal use of computing and network resources, defining appropriate work flows and user interfaces, and providing an overall satisfactory experience to users, while being able to minimize expenses, both operative and capital.

Future Monitoring Solutions for the Cloud

There exists a plethora of monitoring tools for making sure different types of systems run smoothly and reliably. Network monitoring tools, hardware health monitors, service probes, and many other technologies are available and in use today. However, Cloud services require an integrated approach to monitoring, where the different components can be considered in unison, and the results understood in the context of the actual services under consideration. In particular, it should be noted that different performance factors will often affect different services in very disparate ways. As simple example, the network performance

requirements of a Web 2.0 style application and that of video streaming services are completely different. Making network monitoring tools aware of the requirements of the different services being monitored allows them to provide more meaningful results. These same principles apply to the other areas where monitoring is needed.

Raising the level at which current monitoring technologies operate, and thus bringing them closer to the QoE realm will greatly increase their practical value to Cloud service providers and users, and of course generate additional business for monitoring solutions manufacturers.

A New Vision for SLAs and Operations Support

Proper SLAs are critical to service providers and customers alike. For customers, they guarantee that the service they receive is indeed that for which they pay. For service providers, SLAs can be even more useful. The most obvious benefit is that of providing assurances to their customers, which is of course a good business practice, and can make a difference in a very competitive field such as Cloud services. Beyond this, however, SLAs can provide operators with clear performance targets to be met, and constraints on how their operations support systems need to perform.

In a Cloud environment, where there may be several providers involved in delivering a service to a customer, clear and meaningful SLAs can assure that each party is actually delivering what they should. In particular, it is interesting to consider the composition of SLAs in terms of the quality of the final service being delivered, which is itself composed of many different parts (computing infrastructure, software platform, network, etc.). This can greatly simplify reasoning about how to optimize the service's QoE by optimizing the performance of the underlying components.

This understanding, in turn, leads to better ways to improve the operational

efficiency of the service providers. One way to decrease operating costs, while at the same time improving service levels, is to add as much automation as possible to operating support systems. Automated detection and localization of failures, SLA renegotiation, automated ticketing are just examples of possible ways in which greater efficiency can be achieved. More complex approaches can also be taken, for example to distribute services on different hardware platforms automatically, depending on their performance needs, in order to minimize costs (e.g. power consumption) while still abiding by the SLAs and providing a good experience to users.

Conclusions

As more and more services migrate towards the Cloud, with ever decreasing prices, service quality will become an important differentiator between providers. Understanding the concepts and problems behind the Quality of Experience for Cloud services will place service providers and service consumers at an advantage.

Though a new field, Cloud QoE research is off to a very strong start, with several multi-disciplinary European and national-level projects attacking different issues, from QoE models, to SLAs, to the connections between traditional QoS and QoE. The business impact of QoE and its components is also being carefully studied. It is expected that within 2 to 4 years, workable generic QoE solutions for Cloud systems will start seeing wider industry adoption, as they become available. For concrete applications, specific solutions can already be implemented today.

Taking the Next Steps

Migrating applications to the Cloud often presents new problems concerning their performance, reliability, and overall perception of the service quality by the users.

In order to fully realize the possibilities that the Cloud offers, it is therefore imperative that the quality experienced by the actual users be the best possible, while constraining the application to be as resource-efficient as possible in order to minimize operational costs.

Understanding the quality of service in terms of the users' perception (i.e. QoE) is advantageous to both the consumers of said services, as well as the service providers themselves.

Customers of Cloud services will have better ways of ensuring that they are getting their money's worth for the services they contract. Service, platform and infrastructure providers will be able to provide better, more meaningful SLAs to their customers, differentiate their offerings based on quality, and be more cost-efficient in their operations.

VTT and EXFO-NetHawk are among the leaders of the ongoing European research efforts on Cloud services QoE, and are able to help infrastructure and service providers and consumers with their planning, monitoring and optimization needs.