



Contents lists available at ScienceDirect

Journal of Systems Architecture

journal homepage: www.elsevier.com/locate/sysarc

Challenges in real-time virtualization and predictable cloud computing

Marisol García-Valls^{a,*}, Tommaso Cucinotta^b, Chenyang Lu^c^a Distributed Real-Time Systems Laboratory, Department of Telematics Engineering, Universidad Carlos III de Madrid, Av. de la universidad 30, 28911 Leganés, Madrid, Spain^b Bell Laboratories, Alcatel-Lucent, Blanchardstown Business and Technology Park, Snugborough Road, Dublin, Ireland^c Cyber-Physical Systems Laboratory, Department of Computer Science and Engineering, Washington University in St. Louis, 1 Brookings Dr., St. Louis, MO 63130, USA

ARTICLE INFO

Article history:

Received 10 May 2013

Received in revised form 20 July 2014

Accepted 30 July 2014

Available online 9 August 2014

Keywords:

Cloud computing

Soft real-time systems

Virtualization

Resource management

Quality of service

SLA

ABSTRACT

Cloud computing and virtualization technology have revolutionized general-purpose computing applications in the past decade. The cloud paradigm offers advantages through reduction of operation costs, server consolidation, flexible system configuration and elastic resource provisioning. However, despite the success of cloud computing for general-purpose computing, existing cloud computing and virtualization technology face tremendous challenges in supporting emerging soft real-time applications such as online video streaming, cloud-based gaming, and telecommunication management. These applications demand real-time performance in open, shared and virtualized computing environments. This paper identifies the technical challenges in supporting real-time applications in the cloud, surveys recent advancement in real-time virtualization and cloud computing technology, and offers research directions to enable cloud-based real-time applications in the future.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

The widespread availability for the masses of high-speed Internet connections at affordable rates, by means of DSL and more recently optical technologies, paired with an unprecedented connectivity through cellular and wireless technologies, is enabling an inescapable shift towards distributed computing models. Applications relying merely on physical resources and data available in the local *personal computer* (PC) are slowly but consistently becoming part of the history, as the PC declines leaving the way to a new era of distributed computing. This is subsumed into the recently expanding paradigm of Cloud Computing [87], in which resources are rented in an on-demand and pay-per-use fashion from cloud providers. Just as a huge hardware machine, cloud computing data centres deliver an infrastructure, platform, and software applications as services that are available to consumers. Such services are referred to as IaaS (*infrastructure as a service*), PaaS (*platform as a service*), and SaaS (*software as a service*), respectively [88]. Cloud applications are capable of running and spreading out their computations and data on as many nodes as needed, and they can access huge amounts of data directly available within the premises of cloud data centers. Shortly, cloud computing is enabling the next generation of computing services, heavily geared towards

massively distributed and on-line computing, as well as enabling a new model of on-demand *high performance computing* (HPC) accessible to anyone from anywhere, whenever needed.

As new application domains enter the cloud world progressively, real-time systems are also expected to move in this direction due to the tremendous possibilities and augmented utility that this paradigm could bring about. Examples are both on the hard and soft sides of real-time systems, such as military distributed control systems applied to remote surveillance, early response and warning systems, unmanned vehicles with augmented intelligence from the sensor cloud, or cloud gaming, among others.

Cloud computing, and particularly the use of public clouds, brings advantages on the technical, environmental and business sides, allowing multiple under-utilized systems to be consolidated within fewer physical servers hosting them. A cloud provider can manage physical resources in a very efficient way by scaling on the several hundreds and thousands of customers (a.k.a., *tenants*) with dynamically changing workload requirements, by re-optimizing the infrastructure in a completely automated (or semi-automated) fashion whenever needed, providing high levels of availability and reliability. One of the most important technologies that enabled this paradigm shift in computing is *virtualization*, and particularly *machine virtualization*.

Machine virtualization (also referred to as processor virtualization) allows a single physical machine to emulate the behavior of multiple machines, with the possibility to host multiple and

* Corresponding author.

E-mail addresses: mvals@it.uc3m.es (M. García-Valls), tommaso.cucinotta@alcatel-lucent.com (T. Cucinotta), lu@cse.wustl.edu (C. Lu).

heterogeneous *operating systems* (called guest operating systems or guest OSs) on the same hardware. A *virtual machine monitor* (VMM), or *hypervisor*, is the software infrastructure running on (and having full control of) the physical host and which is capable of running such emulation.

Virtualization allows for server consolidation in data centers, where multiple operating systems that would leave their underlying hosts under-utilized can be moved to the same physical resources. This enables the achievement of a reduction of the number of required physical hosts, and their improved exploitation at higher saturation levels, thus saving costs and energy [86].

The multi-tenant nature of cloud computing has a great influence on the increasingly rich and challenging user requirements on cloud infrastructures. Users demand not only access to on-line storage, but also to real-time and interactive applications and services. This is also witnessed by visionary products already on the market, such as lightweight computers which are almost incapable of doing anything locally, unless they are connected to the “Cloud”.

In a *high-performance cloud computing* (HPCC) environment, applications have much stronger temporal requirements; as such, the characteristic of performance, including resource guarantees and timely provisioning of results, becomes critical. It is actually an open research area to match such requirements with virtualized environments due to I/O overhead and jitter of the required duration of executed instructions. Moreover, activities or jobs are mostly allocated to a specific core, but they often have synchronization dependencies with respect to other activities.

Merging cloud computing with real-time is a complex problem that requires to also focus on (among others) the efficient access to the physical platform. Although real-time hypervisors typically may allow applications to access to the physical machine, in virtualized environments for cloud computing, it is clear that the hardware is typically not directly accessible by the user-level application software layers. With the current available technology, it could be possible to improve service response times from a cloud platform using high performance techniques. The main problems lie on the multi-tenancy of the cloud computing platforms that execute on heavy loaded servers the requests of several independent users. Currently, there are a number of commercial real-time hypervisors (some providing hierarchical scheduling) for safety critical systems of different origins such as WindRiver, Acontis Technology, SysGO, OpenSynergy, LynuxWorks, or Real Time Systems GmbH. However, it is not likely to see them among the mainstream cloud technology in the immediate future due to performance levels and compatibility problems at the low level execution layer. Mainstream and real-time ones were created with different objectives. For example, real-time hypervisors were not invented for maximizing throughput of user requests and providing statistical guarantees on service contracts, but to preserve temporal isolation and determinism.

Cloud computing technology is, in origin, not targeted at hard real-time applications which typically run in closed environments. This paper targets at the significant challenges in applying cloud computing technologies to *soft real-time applications*. Examples of soft real-time domains are, for instance, online video streaming (e.g. Netflix on Amazon EC2), cloud-based gaming, and telecommunication management. Such applications can benefit significantly from cloud computing due to their highly dynamic workloads that desire elastic allocation of resources. Cloud-based gaming is gaining momentum in the market. Concrete instances such as Netflix running in Amazon EC2, and both Xbox and Playstation are planning to offer cloud-based gaming. For example, Microsofts Xbox One game console allows computation of environmental elements to be offloaded to the cloud; Sony recently acquired Gaikai, a major open cloud gaming platform. As more latency-sensitive games and players move toward the cloud, it is becoming increasingly

Table 1
Delay tolerance in traditional gaming [4].

| Game type | Delay threshold (ms) |
|----------------------|----------------------|
| First person shooter | 100 |
| Role playing game | 500 |
| Real-time strategy | 1000 |

important to meet varying latency requirements in the cloud computing environment. User studies [3], for example, show that networked games require short response delay, even as low as 100 ms, e.g., for first-person shooter games [2]. Table 1 provides delay tolerance for on-line gaming applications.

In the telecommunication industry, there is a major shift from hardware-based provisioning of network functions to a software-based provisioning paradigm where virtualized network functions [94] are deployed in private or hybrid clouds of network operators [89]. For example, IP Multimedia Subsystem (IMS) components are traditionally designed and calibrated to run on specific hardware platforms with precise real-time and reliability requirements, given a target maximum workload specification, such as maximum number of supported subscribers or call attempts per second. Matching the same requirements in a virtualized context where a multitude of virtual machines (VMs) share the same physical hardware for providing a plethora of services with highly heterogeneous performance requirements to independent customers/end-users brings many challenges, some of which can be tackled as summarized in this paper.

For soft real-time applications, bypassing system software and directly exposing the hardware to applications is neither needed nor it may be the most productive approach. However, the integration of real-time scheduling policies within virtualization platforms produces a direct benefit, and the system can deliver real-time performance to the application in a hierarchical manner. In this paper, we describe the problems arising from mixing the requirements of soft real-time workloads when deployed in the context of distributed and virtualized physical infrastructures, such as in cloud computing. We describe the service level agreement notion and the challenges to support real-time attributes in them. The paper provides a survey that focuses on soft real-time applications that demand certain degrees of service level agreements in terms of real-time performance, but does not require hard real-time performance guarantees. We describe some of the available approaches to integrate the real-time model in a virtualized application model. We provide some approaches to HPCC, and the challenges introduced by the network communication as it requires I/O access incurring in extra delays; some solutions for improving the efficiency of the network are presented.

This survey is complementary to a recent review on real-time virtualization for embedded systems [98]. While [98] focused on hard real-time embedded systems, we address predictable and real-time performance issues in not only embedded systems, but also cloud computing systems, including approaches to soft real-time performance and QoS issues. The paper is structured as follows. Section 2 offers an overview of the virtualization technology for cloud computing focusing on the real-time issues that appear therein. Also, the type of virtualization approaches and their performance levels are provided to give an idea of the suitability for real-time domains. Section 3 presents the challenges for merging real-time and cloud computing: we provide a mapping of terminology between the cloud computing and the real-time worlds; we present specific issues concerning the access to the platform resources faced by virtual machine monitors and hypervisors; we overview different approaches proposed in the literature for scheduling virtual machines; and we explain the network challenges for achieving real-time support and some HPC techniques to tackle

متن کامل مقاله

دریافت فوری ←

ISIArticles

مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات