Minimizing Technical Complexities in Emerging Cloud Computing Platforms

Andreas Menychtas, George Kousiouris,
Dimosthenis Kyriazis, and Theodora Varvarigou

National Technical University of Athens,
Zografou Campus, 9 Iroon Polytechniou Str. 15773, Athens, Greece
{ameny,gkousiou,dimos}@mail.ntua.gr, dora@telecom.ntua.gr

Abstract. Cloud Computing is considered nowadays as the future of ICT systems leveraging new methodologies for developing, providing and consuming services. Even though many people believe that "Cloud" is just another buzzword for utility computing, this new computing paradigm is not only changing the design of modern computing platforms in technical level, but it also impels, from the market perspective, the creation of new value chains and business models. However, many technical complexities still remain, which disallow the wide adoption of Clouds to eventually address the new business trends and requirements of end-users. In this paper we identified and analyzed the key challenges for the emerging cloud platforms in order to minimize these technical complexities while the innovative approaches emerging from European research activities are presented.

1 Introduction

Although cloud computing [1] as another distributed computing paradigm is not something new, nowadays seems that the number of people and organizations exploiting the cloud computing capabilities is increasing and the research interest in cloud technologies is expanding. The main IT players such as Google and Microsoft have already developed platforms [2,3] to offer cloud services hosted in their datacenters and at the same time hundreds of new companies worldwide are involved in the service delivery value chain either by using their owned infrastructures or by providing added value services utilizing the infrastructures of the main players.

The new cloud ecosystems are changing the way the computing, storage and networking resources are purchased and consumed creating new business models and value chains. In contrast with the proprietary software where the license schemas are rather simple, the cloud based services -exploiting the advantages of the cloud features for scalability, multi-tenancy and reliability- are strongly related with the business aspects of the application and platform influencing all process of the service lifecycle. Currently, this is getting even more complex since the IT services are not independent each other but are often federation

M.R. Guarracino et al. (Eds.): Euro-Par 2010 Workshops, LNCS 6586, pp. 603–610, 2011.

of other services, aggregating data and information from various sources. However, a number of technical complexities in the new computing environments deter the placement of composite applications and services. As cloud computing passed the "Peak of Inflated Expectations" and is moving towards the "Plateau of Productivity" according to the Hype Cycle of Gartner Research [4], issues like interoperability, data lock-in and QoS degradation are considered of major importance for the wide adoption of such systems. Therefore minimizing the technical complexities allows the involvement of more players in the elastic services market offering cost efficient services with high QoS and security guarantees without large investments on infrastructures.

The technical challenges for the emerging cloud systems span all layers of the established cloud model (SaaS, PaaS and IaaS) [5] with most of them affecting the functionality and the performance of system components (both in the same and cross layer). In the following figure we summarize the most important of them, which are also expected to draw the main research interest for the next few years. In addition, as the tight coupling of system components is of high importance for the future cloud platforms in order to provide efficient management and operation capabilities, we also present and analyze the main architectural design and cross-layer challenges.

Fig. 1. The challenges for the future cloud computing platforms

The rest of the paper is structured as follows. In Chapter 2 we identify the main technical challenges which will be addressed the next years and illustrate approaches on how to minimize their complexities. Cloud architectural issues are presented in Chapter 3, along a series of critical cross-layer issues that need to be addressed in the forthcoming period. Chapter 4 contains the conclusions of this work.

2 Technical Challenges of Future Cloud Platforms

In order to enable the wide adoption of Clouds and the involvement of SMEs, independent users and developers, the future cloud infrastructures have to be

attractive as technical and business solutions. This implies providing advanced capabilities for all infrastructure layers and mechanisms to support the individual business and market requirements of each application. In next sections we have identified and analyzed the technical challenges from this perspective.

2.1 Performance Analysis

One of the most critical issues in modern cloud platforms is the performance analysis of the application running on a distributed infrastructure. This problem has been thoroughly investigated the previous years in the context of grid technologies (like in the works of [6,7,8]. However, in the current cloud business model, the different roles of SaaS, PaaS and IaaS limit the flow of information from one layer to the other (like source code or hardware capabilities knowledge). In this context, the task of analyzing performance characteristics of an unknown application running on unknown resources becomes almost impossible. In order to minimize the complexity inserted, the IRMOS project [9] follows a multilevel approach that meets most of the PaaS responsibilities.

First of all, the application and its components are described in an XML format through the Papyrus tool. This way, the platform has a complete description of the application, its structure and a number of behavioral and functional characteristics of each individual component. Afterwards, each component is benchmarked through a process analysed in [10]. This aids in modeling the application behavior with regard to changing resources assigned and the effect on the QoS output. From the IaaS part, further analysis is conducted on whether co-scheduling of VMs in the same host influences the performance of each individual VM, thus reducing the effective resource allocation performed by the cloud provider.

2.2 Interoperability

One of the main challenges of future cloud platforms is the interoperability issue. To this direction, the emerging REST protocol [11] is expected to have increased uptake. This is due to the fact that through the standardized interfaces that are required from the former interoperability at least in terms of interfaces is achieved.

This alleviates from the need to have advanced mechanisms for service composition. However the need for semantic bridging between the different providers still remains. Having the same interface is only the first step. The choice of what type of service to use and what type of resources is needed is critical. Research up to now, like in the FUSION [12] project, has progressed to some extent in this area, through the usage of an intermediate, bridging semantic description to which each provider adapts. However, if we are to meet the full expectations of a global and diversified IT market, this process must be performed on the fly and automatically, without the need for intermediate adjusting mechanisms that usually include manual intervention at some level.

2.3 Cloud Federation

Like in the previous case regarding interoperability, the realization of cloud federation in projects like RESERVOIR [13] is based on a predefined schema that is followed by both providers that wish to federate. However this implies human intervention and it limits the amount of dynamicity. In order to have a full scale autonomous platform that is able to federate on the fly with other IaaS providers, automated semantic bridging between e.g. the ontologies used by both is compelling.

A number of issues arise, from the usage of distributed IT infrastructures, which are not technical from a first glance. This mainly has to do with legal issues (e.g. data location) regarding the operational aspects of cloud platforms. The new project OPTIMIS [Optimis] aims to investigate, among others, the aforementioned critical parameters. Having as a starting point a legal analysis of requirements posed by a number of involved parties like legislation dictations or specific user constraints, OPTIMIS data services will be called to implement inter-Cloud data transfer mechanisms that will cover both the functional and the performance-driven point of view. Furthermore policies enforcement mechanisms for data that are transported to federated Clouds and for selecting the optimal data for federation with regard to their nature and characteristics will be investigated.

In order to meet these goals, aspects such as QoS requirements, functional requirements (e.g. how data are accessed from an external network across multiple domains), energy efficiency, performance constraints, data locality and integrity must be taken under consideration. For this purpose, modeling of the data mechanisms will be pursued in order to aid in the management of data sets during operational deployment of the latter in federated (or not) cloud platforms. What is more, a decision needs to be made regarding which parts of existing or newly deployed data will be federated in order to save resources. This decision must weigh critical factors such as what is the nature or usage of the data sets contained at the moment in the infrastructure. For this purpose, profiling mechanisms must be in place in order to assist in this process.

2.4 Data Management

Given that a major limitation of existing distributed and vitalized environments is the insufficient support for data-intensive services, the data management features of the cloud platforms are determinant for delivering cost effective applications and services to the ICT players and end-users. Therefore a great challenge for the success of the future cloud platforms is the integration, both in technical and business levels, of the computational, storage and network resources in an efficient manner to facilitate the delivery of data intensive services with QoS and security guarantees. This is one of the challenges that will be addressed by the VISION project.

VISION Cloud will include several innovative technical and technological approaches in data management. First of all it will raise the abstraction level of

storage, encapsulating the data into objects with user-define and system defined attributes. Metadata will be used for effective access, management and manipulation of the storage enabling scalability and simplification of all storage and data functions. In addition, the problem of data interoperability and data lock-in will be addressed with the implementation of advanced data management functionality for migration and federation of data across geographically distributed administrative domains. Certainly data resources are not independent from the computational ones. To this direction, solutions providing secure execution of computational tasks near their data will be architected. The access to storage will be also highly simplified and efficient with mechanisms to define domain-specific optimizations which will make the content visible to users instead of its underlying storage container. The aforementioned advancements in data management are expected to achieve significant and quantifiable improvements in service delivery productivity, quality, availability, reliability and cost.

2.5 Application and Service Marketplaces

The notions of low-entry cost, scalability and dynamic total ownership cost for using the cloud technologies are fundamental for the Cloud adoption and its economic success. However, in the existing cloud paradigms this comes with limitations regarding the involvement of players with competitive applications in the cloud ecosystem because of the various, often complex, business and technical requirements. In addition, third parties are difficult to deploy their applications in the cloud infrastructures, create new business models and establish synergies since these cannot be fulfilled from a single provider.

In the mobile phones paradigm there are several approaches addressing this problem with the most known and successful the iPhone App Store. The developers and providers join these marketplaces selling their applications and services using various business and revenue models. These solutions leaded many developers and providers to be involved extending their businesses in the mobile market while end users are able to discover hundreds of services and applications to satisfy their needs. In cloud computing paradigm, the marketplace concept is still immature and with many technical complexities.

4CaaSt project [14] targets to minimize these technical complexities designing a cloud marketplace that supports all phases of the service lifecycle (knowledge, intentions, contract and settlement). The marketplace will offer to the providers the ability to publish services and applications in a managed environment, which controls the business terms and conditions (price, revenue sharing, promotion, etc) and also includes integrated rating and billing capabilities, unlike most existing marketplace environments. The 4CaaSt infrastructure will be designed to allow a hosting of compositions allowing the definition of combined models and end-to-end SLAs. While existing marketplaces focus on the trading with standalone services, 4CaaSt service compositions can be published in the marketplace supporting various business terms and conditions. It allows defining business policies taking into account the price models of a service, handling revenue sharing among multiple partners, and executing composed SLAs.

2.6 APIs

The interoperability, federation and marketplace capabilities of future cloud environments need to be supported by advanced, but also efficient and dynamic, APIs. Developing programming interfaces for deploying applications on the Cloud as well as blueprints for describing these applications is a complicated process because of the need to support tailored applications which may be in addition compositions of existing or new applications. Besides, the variety of business characteristics of the applications should be reflected on the design of the APIs. The above introduce additional complexity and overhead in the process of developing and adapting applications for clouds. To eliminate these complexities for all the involved entities -users, developers, providers- the APIs should allow automated or guided human-interactive facilitation of applications and compositions without reducing though the cloud capabilities for interoperability, scalability and QoS provisioning. IRMOS project [9] follows an approach to this direction for applications with real-time requirements. Modeling tools not only enable the deployment of applications on the Cloud but also allow the description of their rich set of high level operational and business requirements in a language that can be interpreted by the platform to a set of low level performance parameters. Furthermore application wrappers can be configured for providing high level monitoring data to the platform for evaluation and comparison with data from the infrastructure to guarantee, through automated corrective decisions during runtime such as resource renegotiation and migration, the smooth operation of the application.

3 Cloud Architectures

Clouds of the future will not be able only to manage and virtualize several types of resources (network, storage, computational) but also to communicate with legacy systems and internet enabled "things" such as wifi locators. The challenge for the system architects is to design a system tha includes services that interact dynamically and continuously, spanning between different domains, and ranging from the application level and down to the level of network resources management and the execution environment. This inlcude a careful synchronization of this rich set of services so as to efficiently operate, manage and reconfigure all the resources under real-time conditions, providing to the end-users the required Quality of Service, agreed in the SLAs. IRMOS project followed an architectural approach that included services to support application developers in engineering their applications, while other services support, in real-time, the application execution.

A major challenge for SaaS providers wanting to exploit the benefits of cloud computing is to manage QoS commitments to customers throughout the lifecycle of a service. The PaaS offers SaaS providers services and tools for estimating resource needs in advance of execution, negotiating QoS with service providers, provisioning virtualized resources. Furthermore, assessment tools for the technical and economic outcomes of provisioning policies and management actions are

provided in case either the application or resources do not perform as expected or need to be adjusted. The IRMOS approach considers analysis and decision support to determine which actions are triggered. In addition, the performance of the monitoring and control between cloud layers is as essential factor in ensuring that QoS guarantees are maintained.

An essential element of cloud computing is the ability to deliver on-demand services with minimal manual configuration. All subsystems need to be self-managed and reconfigured in order to achieve management efficiencies, to react to QoS failures (such as an SLA violation or network link failure) in a timely way and avoid the escalation of such problems. Cloud utilization involves several processes that span in different cloud layers and stakeholders. Therefore, the cloud platforms of the future must not only provide a set of services but also cross layer workflows that consider the control channels and information exchanges which are required to support management of applications and application compositions throughout the full lifecycle.

3.1 Cross Layer Issues

The current business model that dominates the service oriented computing paradigm dictates the 3-tier approach. While very adaptive and flexible from a business point of view, this separation of roles between software, platform and infrastructure providers creates another series of challenges.

First of all, the issue of hardware description exists. Up to now, there is no accurate and widely accepted hardware metric in order to describe a computational resource. The unit that is widely used refers to the processor clock speed. However this is far from sufficient. The PaaS provider is not aware of the scheduling policies of the IaaS provider. Therefore, when a virtual machine (VM) is requested based only on processor speed, the effect of co-scheduling other VMs on the same host is not taken into account, despite the fact that the latter influences significantly in some cases the performance of the application. Furthermore, hardware failures may affect the application execution. The identification of the responsible in this case is critical given that this layer should be held accountable for breaching the SLA contract. Third party presence may be necessary in order to ensure that the allocation of resources in IaaS layer are the requested. However, the existence of third party software internally to the cloud provider is not expected to be something the latter would easily permit. If these points are addressed, then the responsibility for not meeting QoS levels falls on the estimation from the PaaS layer.

Another issue, this time between the SaaS and PaaS roles is the confidentiality regarding the source code of the various application components. While the most promising performance estimation techniques require some knowledge of the source code for accurately depicting the dependencies from various performance characteristics, this is not available in the context of current distributed computing infrastructures due to the lack of willingness to disclose application internal characteristics. This feature leaves the PaaS provider with the only option of 'black box' approaches for the prediction of the application behavior.

4 Conclusions

While cloud infrastructures have up to now fulfilled part of their promises and have emerged as sound technological solutions for end users and providers, a number of issues still exist that hinder the harvest of the potential benefits of this paradigm. These issues, coming both form the technical and business constraints of the current cloud implementations are close related each other and span all the layers of cloud model, obstructing the wide adoption of Clouds and the involvement of SMEs and individuals. Clouds have the power to extend the technological barriers for providing distributed services in global scale and to create new value chains and networks for applications. However, many challenges still remain and to this direction, a number of European research projects are significantly contributing so as to minimize the technical complexities and leverage the cloud platforms to the higher levels of innovation and automation.

References

1. Buyya, R., Yeo, C.S., Venugopal, S., Broberg, J., Br, I.: Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility
2. Google app engine - google code, `http://code.google.com/appengine`
3. Windows azure platform, `http://www.microsoft.com/windowsazure`
4. Hype cycle definition i& overview, gartner research, `http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp`
5. Lenk, A., Klems, M., Nimis, J., Tai, S., Sandholm, T.: What's inside the cloud? an architectural map of the cloud landscape. In: CLOUD 2009: Proceedings of the 2009 ICSE Workshop on Software Engineering Challenges of Cloud Computing, pp. 23–31. IEEE Computer Society, Washington, DC, USA (2009)
6. Chen, Y., Iyer, S., Liu, X., Milojicic, D., Sahai, A.: Sla decomposition: Translating service level objectives to system level thresholds, p. 3 (June 2007)
7. Lee, J.W., Asanovic, K.: Meterg: Measurement-based end-to-end performance estimation technique in qos-capable multiprocessors. In: Proc. of the 12th IEEE Real-Time and Embedded Technology and Applications Symp., pp. 135–147 (2006)
8. Stube, A.O., Rexachs, D., Luque, E.: Software probes: Towards a quick method for machine characterization and application performance prediction. In: International Symposium on Parallel and Distributed Computing, vol. 0, pp. 23–30 (2008)
9. Irmos project, `http://www.irmosproject.eu`
10. Kousiouris, G., Checconi, F., Mazzetti, A., Zlatev, Z., Papay, J., Voith, T., Kyriazis, D.: Distributed interactive real-time multimedia applications: A sampling and analysis framework. In: Proceedings of the 1st International Workshop on Analysis Tools and Methodologies for Embedded and Real-time Systems (WATERS) (2010)
11. Fielding, R.T.: Architectural styles and the design of network-based software architectures. PhD thesis (2000); Chair-Taylor, Richard N.
12. Alexakis, S., Bauer, M., Pace, A., Schumacher, A., Friesen, A., Bouras, A., Kourtesis, D.: Application of the fusion approach for assisted composition of web services. In: Establishing The Foundation of Collaborative Networks. IFIP International Federation for Information Processing, pp. 531–538. Springer, Boston (2007)
13. Reservoir project, `http://www.reservoir-fp7.eu`
14. 4caast project, `http://4caast.morfeo-project.org`