Towards market-oriented Clouds

**Maurice Gagnaire • Pascale Vicat-Blanc Primet •
Dimitra Simeonidou**

© Institut Télécom and Springer-Verlag 2010

In the literature, Cloud computing is seen as an evolution of Grid computing, which itself is an evolution of distributed computing and supercomputing. The distinction between the Grid and Cloud concepts is not clearly defined, both of them referring to an externalization of the private data centres. Two main advantages are expected from these two approaches. Instead of investing in the installation and management of its own computing infrastructure, an enterprise may benefit from shared computing resources managed on a remote site by a third party. This third party is known as the Grid/Cloud service provider (GSP or CSP, respectively). This means that both capital expenditures and operational expenditures of the enterprise may be considerably reduced against a moderate increase in service expenditure. In the years 1990s, Grid computing was mainly developed in the perspective to provide affordable high-performance computing to small and medium enterprises. Circa 2000, Cloud Computing has been introduced by the service sector with the objective to extend the externalization of hardware resources to software resources. This distinction added to the fact that in 20 years, not only professionals but also residential users benefit from high-

speed Internet access are at the origin of the Cloud computing concept. The economical models of Grid computing and of Cloud computing are quite different. Cloud computing enables the externalization of software resources at a very large scale for residential users; whereas; Grid computing mainly refers to the provisioning of externalized large computing facilities for professionals. Today, both types of services are accessible via simple Web portals. This evolution has motivated the emergence of numerous start-ups and companies specialized in service provisioning with an expected market size by 2011 around \$95 billion in business and productivity applications (email, office, CRM, etc.) according to Merrill Lynch analysts. The fields of application of Cloud computing seems almost unlimited, all the sectors of our economy being concerned. Applications are ranging from distant medical diagnostic, collaborative image processing, scientific computation, financial operations, and industrial processes to radio astronomy already financially benefiting from advantages of Clouds or Grids. Two other criteria are also to be considered to distinguish between Grid and Cloud services. The consumers of Cloud services being in majority residential users, the mode of payment of the consumers is carried out on-the-fly with a duration-based charging. The consumers of Grid services being in majority enterprises, their first requirement is rather the availability of the computing resources than the price of the service itself. Resources are in that case reserved in advance for a longer period of time than in the case of Cloud services. A last distinction may be done between Grids and Clouds. In the case of Grids, heterogeneous computing resources managed by different entities may have to interoperate to provide the required service. This is not the case in the context of Clouds where comparatively, a smaller computing power is requested for a given service. Homogeneous hardware

M. Gagnaire (✉)
Institut Telecom/Telecom ParisTech,
Paris, France
e-mail: gagnaire@telecom-paristech.fr

P. V.-B. Primet
INRIA/Ecole Normale Supérieure de Lyon,
Lyon, France
e-mail: pascale.primet@inria.fr

D. Simeonidou
University of Essex,
Colchester, UK
e-mail: dsimeo@essex.ac.uk

resources managed by a single entity can be associated for this type of application. In the remaining of this editorial, we use by simplification systematically the terms “Cloud computing” to introduce each of the papers of this special issue of the Annals of Telecommunications. It has to be noticed that authors of some of the papers included in this special issue use in fact explicitly the terms “Grid computing”.

The National Institute of Standards and Technology of the USA specified three Cloud services delivery models known as Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). The on-demand provisioning of distributed computing facilities relies on SaaS and PaaS delivery models. IaaS consists in the possibility for the CSP to borrow networking facilities belonging to independent carriers in order to satisfy the computing requests generated by client entities. Four deployment models of Cloud computing are today considered: private Cloud, public Cloud, community Cloud and hybrid Cloud. Private Cloud corresponds to the case where the Cloud infrastructure is reserved for a single client entity (for instance a unique enterprise). In the context of private Cloud, confidentiality of the client’s data is obtained by means of traffic isolation techniques and authentication procedures. However, all the clients of a private Cloud share a common set of resources for instance in the context of a collaborative work. Traffic isolation and authentication procedures are mainly used to protect this set of clients from those that do not collaborate to the same work. At the opposite, public Cloud corresponds to the case where the cloud infrastructure can be used to the benefit of a large number of client entities, this number being not limited a priori. In public Cloud, the service requirements may strongly differ from a client to another one. Public Cloud also known as utility computing may concern any user connected to the Internet. In public Cloud, the data of each client entity is protected thanks to a the management of access control mechanisms by the CSPs. Examples of utility computing platforms include Amazon Web Services, Google AppEngine, and Microsoft Azure. Community Cloud is a variant of private cloud in the sense that the same Cloud infrastructure may be used by several enterprises or administrative entities imposing the same type of constraints for the satisfaction of their computing requests. For instance, several enterprises working in the field of biotechnology require in general for their numerical simulations the same type of constraints in terms of algorithmic complexity, storing capacity, and computation delay. For economical purposes, such enterprises have an interest in sharing the same Cloud infrastructure. Hybrid Cloud refers to the case where the Cloud infrastructure is used as a mix of private and community deployment models. In France, the Grid’5000 Cloud infrastructure made of nine access

nodes around the country is dedicated to the research community. It can be viewed as a hybrid Cloud. Cloud middleware consists in a specific software enabling the sharing of heterogeneous resources and the dynamic establishment of virtual organizations. It is installed and integrated into the existing infrastructure of the Cloud hosting companies. Major Cloud middleware available today are Globus toolkit (GTK), gLite and UNICORE. The GTK toolkit has emerged as the *de facto* standard for several important connectivity and resource allocation protocols. It addresses security, information discovery, resource management, data management, communication, fault detection, and portability issues. It is also in charge of service orchestration.

Numerous aspects of Cloud computing still remain open problems. In parallel to the constant increase of Internet traffic, Cloud services gain in popularity year after year. This rapid increase of the Cloud market necessitates scalable Cloud infrastructures and middleware. New business models have to be proposed to take into account the multi-tenant nature of Cloud services. The lack of flexible market-oriented models for Cloud services explains why today the effective increase of commercial Cloud is not as rapid as expected. Interoperability between Cloud platforms will be rapidly a necessity. This interoperability imposes a convergence towards standards accepted by the international community. In this context, multi-domain Cloud infrastructures and services still have to be developed. This poses the problem of the disparity that may exist in terms of technologies used by the multiple tenants. Thus, it is not possible to exploit today Cloud services requiring computing, storage, software and network resources at an international scale for instance because of the disparity of the equipment exploited by the different carriers at the network level. Inter-domain path computation necessitates indeed the specification of a uniform description language of the networking resources. Many enterprises are worried about security aspects in Cloud environment. The specification of Service Level Agreements (SLA) is necessary to describe explicitly the requirements of the end-users. The level of trust guaranteed by the CSP must be one of these parameters. Such SLAs are mandatory to manage charging and billing of Cloud services. Multiple investigations are carried out in this matter. The multi-tenant nature of the Cloud market necessitates the development of suited optimization algorithms trying to maximize the income of the CSP with the agreement of network providers and the owners of the computing and storage devices while satisfying the largest number of end-users. Such a multi-criteria optimization problem is complex and referred as the maximization of the social welfare by the specialists. In market-oriented Clouds, the end-users typically negotiate with a CSP the on-demand provisioning of computing,

storage and software resources. Like in the context of financial market, the terms “bids” and “asks” are introduced to refer to the equilibrium that must be found between the buyers and the sellers, respectively

This special issue entitled “Towards market-oriented clouds” aims to provide an up-to-date overview of the most recent advances in the field of research initiatives, applications and standardization of Cloud computing. Ten papers from Europe, North-America, and Asia have been selected after a peer review process. This special issue is organized in three sections. Section 1 includes four contributions dealing with market-oriented Clouds. Section 2 is focused on resource virtualization and Cloud infrastructure. These two first sections present the results of advanced investigations. Section 3 is made of three brief reports providing an up to date overview of the state of advancement and of the recent services obtained by three major European experimental projects in the domain of Cloud computing. Most of the open issues we have mentioned in the previous paragraph are covered by these ten papers.

The four first papers of Section 1 are directly related to market-oriented Clouds. The first contribution due to Jörn Altmann, Costas Courcoubetis, and Marcel Risch is entitled “A marketplace and its market mechanism for trading commoditized computing resources”. This work has been carried out in the context of a collaboration between the College of Engineering of Seoul National University in South-Korea and the Athens University of Economics and Business in Greece. In their introduction, the authors justify the slow uptake of Cloud computing market observed these recent years by the lack of sustainable business models for Cloud resource provisioning. In this context, they develop an original service-oriented platform called GridEcon enabling to emulate market scenarios for Cloud environment. Unlike other existing comparable platforms which are limited to resource allocation services, the GridEcon platform flexibility enables to consider a large variety of value-added services provisioning. This flexibility takes into account the possibility of a multi-tenant environment for resource provisioning. The paper describes into details the quantitative parameters on which relies the economical model of the platform such as the unit-of-trade, the bids, and the asks. A CSP decides to accept or to reject a job request on the basis of these various parameters. Thus, the GridEcon platform can be used to test and validate new business models for market-oriented Clouds. In that sense, it can be exploited to investigate new market-oriented approaches for Cloud services provisioning that are much less rigid than the comparable approaches adopted currently by the major companies of the sector. A set of numerical results outline the capacity of the GridEcon platform to manage efficiently bids and asks under a large variety of market environments. The second contribution due to

Mohammad Mehedi Hassan, Biao Song, and Eui-Nam Huh from Kyung Hee University in South-Korea is entitled “A market-oriented dynamic collaborative Cloud services platform”. This paper deals with the same topic as the previous one in providing an original focus on the interoperability and scalability aspects. These two properties are addressed by means of specification of a dynamic collaborative platform implying multiple cloud providers. This approach is particularly interesting in the context of collaborative or portable Cloud services that could be developed by different CSPs. The basic idea that has driven to the design of this collaborative platform is to minimize conflicts between Cloud providers. For that purpose, an innovative combinatorial auction-based Cloud market model is proposed. Conflict minimization is obtained by means of a partner selection process among the Cloud providers, each group of Cloud providers proposing a single bid in the auction process. Defining the optimal groups of Cloud providers in order to minimize conflicts is nevertheless a hardly tractable optimization problem. The authors propose then as an alternative to this limitation by proposing a multi-objective genetic algorithm (MOGA). This optimization model uses individual information such as price and quality of information proper to each service and to each provider. It also exploits the memory of past collaborative relationships between the various providers. The originality of this last aspect consists in memorizing the number of auctions won or lost by each provider during the lifetime of the system. Two variants of the MOGA approach are considered. One is based on the non-dominated sorted algorithm, the other one being based on the strength Pareto evolutionary genetic algorithm. The efficiency of these two approaches is evaluated and compared via multiple simulations scenarios. The third contribution due to Lucile Denoeud-Belgacem from Future-Master Company, France, and Eric Gourdin, Ruby Krishnaswamy, and Adam Ouorou from Orange Labs also from France is entitled “Combinatorial auctions for exchanging resources over a Grid network”. This paper, investigates the context where a large number of small organizations collaborate to a same project in a Cloud environment. Each of these organizations may be viewed as a resource consumer and also as a resource provider for the other partners. Numerous market-oriented approaches for resource allocation have been proposed in the literature for this type of context. Combinatorial auctions in which prices are adjusted dynamically in time are known to be a well suited approach to find a solution optimizing the social welfare of all the participants. In this perspective, both an exact and an approximate solution to the winner determination problem are considered. A dynamic adaptation of prices on a time-slot base enables end-users to regulate the generation of their service demands according to their more or less

stringent expectations. On the basis of this information, each end-user may then choose the most suited time-slot during which it decides to be served. Like in the paper of Altmann et al., value-added services provisioning are facilitated, but in this paper thanks to the proposal of an original bidding language. The fourth contribution due to Quan Liang and Yuanzhuo Wang from Fujian University of Technology and the Chinese Academy of Sciences, respectively, both in China, is entitled “The representation and computation of QoS preference with its applications in Grid computing environments”. This paper proposes a technique inspired from the Analytic Hierarchy Process used in multi-criterion decision theory to select service requests generated by a large number of users in a Cloud environment. Service selection is based on the Quality of Service (QoS) requirements of the users. For that purpose, a three-step approach is considered first to specify user’s QoS expectations, then to convert these specifications into a computable form suited to service selection, and finally to determine if selected services benefit effectively of the expected QoS. The efficiency of the analytical formulation of this multi-criterion decision process is then applied to an experimental platform. Inspired from this theoretical analysis, a Grid activity monitoring tool called Grid Vision and Analysis System has been developed and implemented on the ChinaGrid experimental platform.

Section 2 begins with the fifth contribution of this special issue. This paper dealing with resource virtualization is due to Fabio Baroncelli, Barbara Martini, and Piero Castoldi. The two first authors are from CNIT, the inter-university national consortium for telecommunications in Italy. The third author is from the Santa Anna teaching and research Institute also in Italy. Their paper entitled “Network virtualization for Cloud computing” proposes to consider data network provisioning as a commodity that can be offered indirectly by the carriers to the end-users via the mediation of a CSP. In this perspective, the authors introduce the concept of Network as a Service (NaaS). A Network Virtualization Platform (NVP) enabling service mediation between NaaS and the other Cloud services such as IaaS, PaaS and SaaS is proposed. Judiciously, the authors insist on the specificity of NaaS compared to the functionalities of existing network control planes. Indeed, if a control plane such as GMPLS used in current wavelength division multiplexing (WDM) optical networks facilitates on-demand connectivity to the end-users via User Network Interfaces, it is not suited to the inherent abstraction of the infrastructure required in Cloud environment. In addition, existing control planes are not designed to deal with advanced network resources reservations and do not include negotiation functionalities. In this context, the authors make the distinction between two possible approaches known as directed signaling and undirected

signaling that have been proposed recently for Cloud-oriented NaaS. They explain why undirected signaling is better suited for a multi-domain and multi-vendor environment. A very interesting overview of some of the major research projects dealing with these aspects (Mupbed, Phosphorus, Frederica, G-Lambda Carriocas, etc.), including also hybrid approaches (Phosphorus, Dragon) is provided. Two use cases are considered as proof of concept of the proposed NVP. The sixth paper also belonging to Section 2 is, like the previous one, dedicated to resource virtualization at the level of the network infrastructure with a particular focus on WDM optical networks multi-domain environment. This contribution is entitled “A new framework for GLIF inter-domain resource reservation architecture (GIRRA)”. It is authored by Gigi Karmous, Silvana Greco, Admela Jukan and George Rouskas. The first three authors are from Braunschweig University in Germany. George Rouskas is from North Carolina State University, USA. In comparison to the paper of Baroncelli *et al.*, the specificity of this paper consists in focusing on scientific applications requiring very large data transfers of the order of the petabyte at an international scale. Such data transfers can only be satisfied by means of long-haul WDM lightpaths provided on-the-fly to the end-users. The field of application of this study is the Global Lambda Integrated Facility also known as GLIF that provides high-speed optical connections to the scientific community at the intercontinental scale. Typically, the GLIF infrastructure relies on Carrier Grade Ethernet circuits established onto a Synchronous Optical Network WDM architecture. The fact the GLIF consortium now considers to open its resources to commercial applications reinforce the interest of this contribution. The authors begin to underline the fact that inter-working between existing Network Resource Managers proper to each network domain (for instance each country) remains today a very open issue. An inter-domain path establishment procedure called GIRRA is proposed by the authors. GIRRA is currently under discussion within standardization bodies such as the Open Grid Forum (OGF). The fact GIRRA also includes accounting and billing aspects serves as a link between Sections 2 and 1 of this editorial. The seventh paper is entitled “A funding and governing model for achieving sustainable growth of computing e-infrastructures”. It is authored partially by the same team as the first paper of this special issue, that is Ashraf Bany Mohammed and Jörn Altmann from the College of Engineering of the Seoul National University, South-Korea. This contribution can also be viewed as another link between Sections 1 and 2 of this editorial. In spite of the deployment of large distributed computing and networking infrastructures, the slow uptake of Cloud computing market is underlined. The arguments mentioned in this paper are more macro-economic oriented than those

adopted in the first paper of Section 1. Indeed, the authors deplore the fact distributed e-infrastructures are still supported by short-term public initiatives and the governance of these computing and network resources still remains exclusively under the control of the research community. From this observation, it is proposed a simple market-oriented funding model of e-infrastructures susceptible to serve as a booster for the real take-off of Cloud economy. Basically, the authors depict a new funding and governing analytical model aiming to motivate safer investments in Grid infrastructures.

Section 3, the last one of this special issue, includes three brief reports. The first paper of this section (the eighth contribution of this issue) is entitled “UNICORE-6: recent and future advancements”. It is authored by Achim Streit and numerous co-authors, all involved in the UNICORE project. Achim Streit, the editing author, is from Forschungszentrum Jülich GmbH Company in Germany. The UNICORE project initiated 10 years ago in Germany was focused initially on supercomputing applications. Today, this project covers the broader spectrum of Cloud services. The practical achievement of the UNICORE project is to propose an open source software tool enabling to exploit dynamically distributed resources. In its current version, UNICORE-6 corresponds to an operational three layered architecture including client, service and system. It uses multiple standards from the Web services and the Cloud domains. In comparison to the papers presented in the two previous sections that are more or less theoretical, the authors adopt a very pragmatic approach to describe the UNICORE-6 system architecture. The information provided in this paper maybe of practical interest for the implementers. The impact of the obtained results towards the standardization bodies is discussed into details. The

ninth paper is due to Uwe Schwiegelshohn from the Technical University of Dortmund, Germany. This contribution is entitled “D-Grid: a national Grid infrastructure in Germany”. Unlike UNICORE-6 that was initiated before the emergence of Cloud computing, this paper provides a state of advancement of the D-Grid project that was launched in 2005. The D-Grid project consists on the setting up of a national Cloud infrastructure in Germany open to both academics and commercial service providers. Multiple projects are actually exploited by different teams over the D-Grid infrastructure. Without going into the technical details, the author depicts the three chronological phases that characterize the evolution of the usage of the D-Grid infrastructure. The tenth and last paper of this issue is entitled “The next generation ARC middleware”. The editing author is Farid Ould-Saada from the University of Oslo, Norway. This paper describes the main evolutions of the Advanced Resource Connector (ARC) middleware for Cloud environment that has been developed in Norway since the year 2002. In its original version, the ARC middleware was strongly inspired from the GTK open source toolkit. The European Union has recently funded an extension of the ARC project under the name of KnowARC. This extension consists in implementing the service-oriented architecture concept with the adjunction of standard-compliant Web Service interfaces. KnowARC implements most of the recommendations of the OGF standard. The efficiency of ARC for Cloud environment is today confirmed by its wide deployment in multiple countries in Europe for a large variety of applications, from medical diagnostics to high-energy physics.

The guest editors express their thanks to the authors and to the reviewers for all their efforts in the preparation of this special issue.