On the Next Generations of Infrastructure-as-a-Services

Petcu, D.; Fazio, M.; Prodan, R.; Zhao, Z.; Rak, M.

DOI
10.5220/0005912503200326

Publication date
2016

Document Version
Final published version

Published in
CLOSER 2016: proceedings of the 6th International Conference on Cloud Computing and Services Science

License
CC BY-NC-ND

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)
On the Next Generations of Infrastructure-as-a-Services

Dana Petcu¹, Maria Fazio², Radu Prodan³, Zhiming Zhao⁴ and Massimiliano Rak⁵

¹Institute e-Austria and Computer Science Department, West University of Timișoara, 4 V. Pârvan, Timișoara, Romania
²Dicieama Department, University of Messina, 98166 Sant Agata, Messina, Italy
³Institute for Computer Science, University of Innsbruck, 21a Technikerstrasse, Innsbruck, Austria
⁴System and Network Engineering, University of Amsterdam, 904 Science Park, Amsterdam, Netherlands
⁵Department of Industrial Engineering and Information, Second University of Naples, 29 Sede di via Roma, Aversa, Italy

Keywords: Infrastructure-as-a-Service, Surveys, Previsions.

Abstract: Following the wide adoption by industry of the cloud computing technologies, we can talk about a second generation of cloud services and products that are currently under design phase. However, it is not yet clear how the third generation of cloud products and services of the next decade will look like, especially at the delivery level of Infrastructure-as-a-Service. In order to answer at least partially to such a challenging question, we initiated a literature overview and two surveys involving the members of a cluster of European research and innovation actions. The results are interpreted in this paper and a set of topics of interest for the third generation are identified.

1 INTRODUCTION

After a decade in which cloud computing has dominated the interest of the communities involved in distributed systems, we ask ourselves what can bring more this paradigm in the next decade.

The first generation of cloud computing products, following the stages of technology triggers, peaks of inflated expectations, as well as through of disillusionment, is based on the concept of centralized pool of services, the expectation that one-size-fits-all, and intensive use of virtualization technologies. The development focus was put on building the proper tools to offer the provider-side service management at costs affordable by their clients.

Triggered by various factors, like user’s special requirements, integration with other technologies, concerns like vendor-lock-in, interoperability or security, we assist nowadays to the raise of a second-generation of more mature cloud services, technologies and products. This generation is characterized by a go-to model for enterprises, born-in-the-cloud applications, decentralization tendencies, integration with Internet of Things, Big Data or mobile technologies, user-driven, software-defined or energy-efficient service management and orchestration, and even the shrinking or dissolution of the virtual machines with the rise of containers or unikernels. We look forward to see what the third generation of Cloud products will bring out-of-the-box, as well when the user adoption and acceptance is world-wide spread so that the cloud can be declared ubiquitous. However, the research activities that are supporting the new generation should start today and challenges in front should be identified as soon as possible. This paper is aligned to this desire. It is trying to provide a peek into a possible version of the future, by following the opinions of the specialists working currently in research activities related to Infrastructure-as-a-Service (IaaS). The result is a list of topics of interest for future research.

The paper is organized as follows. The next section describes the methodology that was used. Section 3 discusses several topics that are currently on the research desks. Section 4 issues some hypothesis about the topics of interest for the third generation. Conclusions can be found in the last section.

2 METHODOLOGY

The sources of information and the methodology used in this paper to elaborate on future research directions are discussed on what follows.

Sources of Information and Stages. In a first stage of the study we have have appealed to the fol-
lowing sources: (1) research papers indexed in ScienceDirect and IEEE Xplore catalogs from the period 2014-2016 which are identifying challenges and open issues related to the evolution of IaaS; (2) cloud-related blogs and news provided by cloud service owners or industry-related journals; (3) the text of the most recent call for research and innovation actions in the field of cloud computing (European Commission, 2015), in the frame of H2020-ICT, as result of a long process of specialists and decision-makers consultations; plus, occurrences of its keywords in SCOPUS catalog of scientific articles and in the searches done using Google Trends. The result is a list of topics related to IaaS supposed to reflect the present, near future and far future activities in the field.

In a second stage, the opinion of the specialists involved in 20 actions\(^1\) partially funded be the European Commission in the frame of FP7-ICT, CIP and H2020-ICT programmes was requested. Their research subjects are varying from heterogeneous clouds or DevOps for multiple clouds, virtualization technologies or security in clouds to IoT-cloud integration, cloud-based Big Data, or mobile cloud. Two surveys were pursued respectively in October and November 2015, to identify the followings: (1) the on-going research activities that are expected to support the second generation services (with 16 contributors, 62% from academia and 38% from enterprises); (2) the foreseen research challenges in front of the third generation services (with 42 contributors, from which 20% developers, 24% software architects, 81% researchers, 27% managers). The results are presented in two on-line documents available on the web site (InfraCluster, 2015). In this paper, we are commenting, linking and interpreting further the results.

**Complementarity with other Studies.** A complementary set of topics for future research was identified by (Jeferry et al., 2015) under the title of technological challenges and goals for the evolution of cloud services and environments. The considered approach is different from ours as it starts from a number of application scenarios for which the requirements cannot be met by current cloud technologies, but can drive the evolution of future cloud technologies. The identified scenarios are referring to: joint collaborative business intelligence platforms with multiple data sources; born-digital data interoperability through semantic technologies; awareness of the application characteristics enabling the infrastructure to provide resources that accommodate the requirements that are evolving during runtime; model-driven integration in cloud IDEs; joint application and infrastructure controllability; quality-of-service/experience-critical applications on the cloud; adaptable parametric applications; generic application templates.

The industry decision makers are trying also to figure out where cloud computing is going in the near future. Articles like (Yousif et al., 2014) are reflecting very well the main current concerns that should be remove in the near future (again a different approach from that of the current paper). The present is driven by the requirements like: single catalog or service-management interfaces allowing smooth transitions between clouds; systems of engagements for agile type of application and application development; short development cycles and continuous delivery model (DevOps); compliance, security, and control measures; spanning from customer premises to the public cloud (including hybrid cloud). The line between the infrastructure and the application is expected to blur, as the line between development and quality assurance and operations, as well as the line between compute, networking, and storage.

We will try in what follows to not touch the subjects already tackled by the above two cited studies.

3 **THE SECOND GENERATION:**

**NEAR FUTURE RESEARCH**

We focus in this section on the expectations of the second generation in terms ranging from challenges to be solved to technologies under use.

**Integration Challenges.** Examining the recent research reports, we observed that a large number of activities are focusing on the integration of cloud computing with other technologies. We are pointing here towards several surveys on challenges.

**Vertical Tailored for Specific Industry.** The paper (James and Chung, 2015) is revealing a new direction currently followed by cloud service providers to tailor their services specifically for their customer and their industrial needs. The main challenges are related to the provision of tailored solution to various levels, from infrastructure services, growth patterns, or software functionality, to privacy and security, as well as to follow the requirement to interoperate with third party community tools.
CloudIoT Paradigm. Cloud and IoT are merged together enabling a large number of application scenarios (Botta et al., 2016). Several new services are added to the infrastructure level: Things, Sensing, Sensing and Actuation, Sensor Event, Sensor, Video Surveillance -as-a-Service. The main challenges are related to privacy, legal and social aspects, large scale, security, reliability, performance, heterogeneity. The problems to be solved are referring to identification, addressing, naming, environmental context-based services, large scale support for multi-networking, flexible mechanisms for logically isolated network partitions, convergence to a common open service platform environment to develop applications.

Cloud Providing Capabilities for Big Data. Several recent surveys are analyzing the approaches, environments, and technologies on areas that are key to Big Data analytics capabilities and discuss how they help building analytics solutions on clouds. In particular (Assuno et al., 2015) focuses on technical issues of enabling cloud analytics and indicates a set of gaps and recommendations on future research directions on cloud-supported Big Data.

Mobile Cloud Computing. Migrating an application processing to the cloud data centers enables the execution of resource-intensive applications on the mobile devices. However, the resource-intensive migration approaches and the limitations of the wireless medium hinders the applications from attaining optimal performance in the cloud. Executing the application with low cost, minimal overhead, and non-obtrusive migration is a challenge. The paper (Ahmed et al., 2015) presents the state-of-the-art of mobile application execution frameworks and discuss the optimization strategies facilitating effective design, efficient deployment, or application migration with optimal performance. The identified challenges are referring to: optimal application and execution framework design, efficient deployment and user-transparent execution, real time optimized management of heterogeneous environments, automated service provisioning, scalability, availability of services.

Secure the Cloud. The latest surveys, e.g. (Ali et al., 2015) or (Tari et al., 2015), presents recent solutions reported in the literature to overcome the security issues. In particular, the latter explores the challenges of cloud security, with a special focus on data utilization management aspects, including data storage, data analytics, and access control.

Cloud Networking. The functional problems associated with the evolution of cloud computing from the network perspective are revealed in (Moura and Hutchison, 2016). If the application could inform the providers how its traffic should be treated, an application performance improvement would be possible.

User-driven Cloud Service Management. The paper (Rehman et al., 2015) presents a comprehensive discussion on the existing approaches to cloud service management, and evaluates them against the factors required for the user to manage the cloud service. While more automation for load management and moving workloads around are achieved, understanding and controlling what is going-on gains in importance (Yousif et al., 2014). Quality metrics are re-defined from an user perspective; for example (Nabi et al., 2016) surveys the current mechanisms and metrics for availability and found that most solutions do not provide protection against application failures.

Decentralized Models for Clouds. The paper (Khan et al., 2015) studies alternative models of cloud computing, like the community clouds at the edge, volunteer-, edge-, social- or mobile-cloud computing. The requirements for a community cloud system are related to autonomy, security, self-management, ease of use, social and economic mechanisms, support for heterogeneity, standard middleware interfaces.

Software-defined Environment. It enables the abstraction from the infrastructure, making the infrastructure programmable. It is the main driver for cloud automation. OpenStack is a central element to this concept, according (Yousif et al., 2014).


Research Topics of Interest for the Near Future. Taking into account that the key-words/phrases of the next call for actions exposed by in the workprogramme H2020-ICT (European Commission, 2015) are the results of several rounds of involved communities’ consultancy, we consider them in this paper as buzzwords for the research activities to be performed in the near future, and, consequently, representative for the second generation of products.

We estimate the volume of the activities around these buzzwords as follows: (a) to reflect the current interest, the volume of searches reported by Google Trends using the keywords; (b) to reflect the solutions already available, the volume of result reports registered in SCOPUS and related to the keywords; (c) to reflect the incoming solutions, the on-going activities related to the keywords in the cluster of projects mentioned earlier. The results are normalized relative to the highest number of hits/values, in order to express a relative importance provided to certain topics.

Figure 1 provides a graphical representation of the result. Perfect match between the three percents associated with a key-word/phrase is not encountered, as the subjects for the on-going research activities.
Interest in the current topics following the keywords: in SCOPUS (work already done by researchers); in Google Trends (interest for the future of a large community); in the cluster of projects (on-going work).

Figure 2: Percents of respondents who are positioning the selected topics in the present (green and magenta). are evolving compared with the previous ones and do not match exactly the hot topics depicted by peaks in search volumes (subjects for the next generation).

Current technologies. Following the second survey, Figure 2 is pointing towards several subjects of current research interest that are foreseen to lead to second generation products in near future.

Containers. Are used for application distribution, cost effectiveness, security and portability. It is expected to generate a new focus on automation tools, including the interest to provide automated component portability based on pre-defined criteria. The main expectation is to reduce complexity by using container abstractions.

Continuous Development of Cloud-ready Applications. While the rapid development, agile or DevOps are now standard procedures, continuous application development is an emerging trend. It involves the constantly move of new releases into production, while each iteration of an application is staged for deployment once testing is completed. The steps in the process are usually small and the iterations move through the development pipeline quickly. Therefore the model accelerates the overall development process and improves efficiency.

Web-powered Applications. Developing cloud-based applications compatible with multiple platforms is becoming a must. With new initiatives for Internet protocols and scripting, the Internet becomes the main platform for these applications.

Fat Cloud Server. Infrastructure services are reshaped to support servers allowing larger instances or more virtual machines per server, in-memory database instances or remote direct access memory.

BYOD. The flexibility demanded by the mobile workforce is the main reason for the integration between the personal cloud services for the companies employees in a Bring your own device (BYOD) environment. End users are currently using mobile devices to put their data into personal cloud services for storage, streaming or syncing.

Tools and Standards in Use. Following the first survey, Table 1 provides an insight on the percent of initiatives that are using today tools and standards for the development of the second generation products.

Table 1: Most used tools and standards.

<table>
<thead>
<tr>
<th>Tools</th>
<th>%</th>
<th>Standards</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>OpenStack</td>
<td>92.86</td>
<td>CloudML</td>
<td>28.57</td>
</tr>
<tr>
<td>KVM</td>
<td>50.00</td>
<td>TOSCA</td>
<td>21.43</td>
</tr>
<tr>
<td>Chef</td>
<td>42.86</td>
<td>Openflow</td>
<td>21.43</td>
</tr>
<tr>
<td>Docker</td>
<td>35.71</td>
<td>WS-Agreement</td>
<td>21.43</td>
</tr>
<tr>
<td>VMWare</td>
<td>21.43</td>
<td>OCCI</td>
<td>14.29</td>
</tr>
<tr>
<td>Eucalyptus</td>
<td>21.43</td>
<td>OVF</td>
<td>14.29</td>
</tr>
<tr>
<td>Puppet</td>
<td>21.43</td>
<td>ETSI GAL</td>
<td>14.29</td>
</tr>
</tbody>
</table>

4 THE THIRD GENERATION: CURRENT EXPECTATIONS

Following the second survey, the main topics selected as being of interest for the research and innovation to create the third generation are depicted in Figure 3. Note that the complex topics related to security, software engineering, multiple clouds were excluded.

Support for a New Software Generation. We estimate that the IaaS will evolve towards its omnipresence, freeing the cloud service consuming software from the cloud services, by ensuring that the new generation software, or existing modular or event-reactive ones, will be able to be described to an abstract level that is service agnostic, will be able to form automatic and transparent combinations of hardware and software resources according to its needs, while resource provisioning, deployments, runtime migrations, multi-tenancy with cost-effectiveness and data protection, or the recovery from minute-to-

323
minute failures will be managed automatically. More details are provided bellow.

**Social-media Like Evolution.** The applications are expected to form automatic associations of hardware and software resources according to their needs, e.g., following a social-media style in which a database can preferentially select a trusted server or storage array (this approach requires that the resources of a datacentre will mould itself around the task required, rather than following the current opposite way). Application developers will no more worry about provisioning servers, storage, or communication as the provisioning process will happen automatically.

**Event-driven Architecture.** This trend is reflected in the recent Lambda architecture, in which code is executed in response to an event, ensuring an important step in moving away from server-centric design. The cloud becomes a generic compute engine, and the developers do not need to organize the resources as they simple just run the code.

**Failure Proof Systems.** New software systems need to be developed to deal with the likely minute-to-minute failures of the consumed resources.

**Thin Cloud Connection.** Applications will no more belong to only one cloud, but will be able to reside in various clouds and to use various on-premise other applications. Different parts of applications will be able to float around in and out of current cloud, e.g. following a policy-based SLA management plan.

**Modular Cloud Applications.** The software development process is expected to place an emphasis on modular software, allowing their components to be modified without shutting down the full.

**Software Moving Away from Hardware.** The software is expected to float away from the hardware and software resources: the software will described in abstracted space and will go through several adapters before it interacts with hardware (being therefore practically hardware agnostic).

**Multi-tenancy in Software.** The production of multi-tenant software is challenging. When multiple users access the software relying on one cloud, the software developer needs to make sure that data is kept separate and the associated costs are identified.

**IaaS Model Evolution.** The IaaS model will evolve towards the integration with machine-to-machine computing, solving the challenges of separate technology stacks and dealing with limited memory, storage and computation capacity of the edge devices, speed of deployment, resource distribution, cost-effective scalability, resilience, easiness of management or security.

**Integration with Operational Technologies.** A challenging area is the merge of information technology with operational technology. The information technology is established around the open-source, Internet protocols, international standards. The operational technology is characterized by proprietary standards and non-cross functional equipment. This includes the hardware and the software that is used to control and monitor the manufacturing equipment and processes, and most communication is accomplished between machines. The two groups have separate technology stacks, network architectures, protocols, standards, governance models, and organizations. By
converging them in one solution, organizations can create better products, achieve cost and risk reductions, improve performance, flexibility and efficiency.

**Fog Nodes Integration.** The fog computing promises to bring small computation and storage capabilities, supporting execution of applications that require low latency interactions with sensors, actuators, end-users. Fog nodes are implemented using embedded systems, in industrial control boards or home routers. Their limited memory, storage and computation is a challenge in their integration into the cloud architecture to enable the execution of application logic. Their integration will leverage containers as main technology for application delivery and execution.

**Smart Gateways.** The success of fog computing depends on the resilience of the smart gateways directing tasks on an Internet teeming with IoT devices. This smartening will rely on features such as out-of-band access, automatic detection and recovery from outages, cellular connectivity, or high-level security. Moving the processing of data to the edge raises the challenge to maintaining the availability of these smart gateways and their communication path to the cloud. Resilience is needed for business continuity, with redundancy, security, monitoring of power and cooling and failover solutions in place to ensure maximum uptime. Speed of deployment, cost-effective scalability, and ease of management with limited resources are also main concerns.

**Inverse Cloud Models,** like machine-to-machine computing (M2M) or geo-distributed cloud, have been proposed as an alternative to the bandwidth-intensive cloud approach. Company branded versions, like fog computing, edge computing, in-network distributed cloud, or Cloud 2.0 were proposed in this context. M2M provides compute, storage, data or application services to client endpoints, like the cloud computing model. Unlike it, M2M distributes data transmission among endpoints and routers, instead sending them to a server. However, it has its challenges, including limited resources and network capacity, security, or resource distribution.

**IaaS Market Orientation.** We estimate that the IaaS will evolve towards diversification, ensuring special features like those sustaining user mobility, user as service provider, service composition, personal data service configurability, or speciality cloud services. More details are provided bellow.

**Personal Cloud – Configurable Services.** The personal cloud providers are expected to focus on business users along with offering superior value added services for users, supplemented by the integration of wearable technologies. The users will pay to easily deploy configurations and services they are acquainted to use. The exploitation of containers in embedded and resource constrained devices gives the opportunity to encapsulate user preferences, applications and services in virtual environments and to deploy them through the cloud in personal devices.

**Users as Service Providers.** Social tools will bring increased collaboration to the cloud. As people become more accustomed to these tools, customers use them to communicate with fellows. The face of cloud is expected to change into a solution which will be delivered as a service with end users behaving more like service providers as a result.

**Specialized Clouds.** Clouds will raise through their ability to avoid specialized hardware acquisition costs and outdated equipment. E.g. video editing will create expanding markets for large instances and specialized GPU clouds. Running high-end graphics applications is requiring today substantial hardware infrastructure investment; newly cloud-based graphics technologies are showing that end users can run high-end graphic design applications with a web browser.

**Personal Clouds – Data Services.** Data generation activities among users and the need to access data from anywhere using any device increases the demand for personal cloud for real-time data access and its sharing. Additionally, the growing need to create a backup of critical data, along with provisions for data recovery and planning for digital contingency, is stimulating the personal cloud adoption.

**Connecting Cloud Services.** By connecting cloud services, the companies should be able to effectively manage their assets. However, a close look to which providers offer cost-efficient or reliable services is needed. The ability to build agile hybrid clouds will influence the design of the middleware.

**Datacentre Services.** We expect to see an evolution of the software-defined datacentres as ecosystems, in which services are abstracted from infrastructure, changes and updating are done automatically based on intelligent orchestration or new database tools, security is software-defined; such ecosystem should enable warehouse-scale computing using purpose-designed chips, new services like supercomputing on demand or massively federated, scalable software architecture with orchestration through network awareness. More details are provided bellow.

**Warehouse-scale Computing.** Low-power processors will be able to treat many workloads to support massively federated and scalable software architecture. A new generation of warehouse-scale computing is coming and custom chips for cloud partners and purpose-designed chips as well as energy-aware management are expected to be supplied.

**Ecosystems.** The datacentres will be like biolo-
gical organisms: having different states, growing and shrinking according to workloads, automatically corrected and changed. An over-arching system will rule equipment via software.

Software-defined. The software-defined datacentres with all infrastructure virtualised and delivered as a service is changing the way organisations view the value of the underlying physical infrastructure. Within the datacentre the services will be abstracted from the infrastructure. Software-defined security will become part of the software-defined datacentre.

Orchestration through Network Awareness. When large-scale applications with thousands components are deployed across multiple datacenters the SLAs need to be enforced and software usage and faults need to be monitored and managed. Cloud management and orchestration must be aligned with various products and services. Consequently, everything from orchestration to database tools will evolve. Datacentre operators will add value to cloud orchestration through network awareness and integration of cloud orchestration with their management platforms.

Private Cluster of Demand. Technologies currently limited to supercomputing will make it into the mainstream. The web-server-sized instances will be replaced by on-demand private clusters. New cloud services are expected to emerge, like supercomputers on demand, high-performance storage, or new ways of storing and processing data.

5 CONCLUSIONS

The paper touched the subject of predicting the unpredictable, the far future of cloud, based on a limited analysis of the trends in on-going research activities in the European area and a small set of visionary papers. Based on them, we concluded that the evolution of the IaaS model and market in the future will follow four main directions: significant changes in the support of data application development, datacenters reorganisation, model evolution towards smooth integration with other models, market orientation towards specialization and personalization. Complex topics like software engineering, security, multiple clouds were intentionally excluded (future studies).

ACKNOWLEDGEMENTS

This research is partially supported by the grants RO-PCE-0260-AMICAS and EC-H2020-ICT 643946-CloudLightning, 644048-BEACON, 644179-ENTICE, 643963-SWITCH, 644429-MUSA.

REFERENCES


