Cloudifying the 3GPP IP Multimedia Subsystem: Why and How?

Roch Glitho

CIISE, Concordia University, Canada

Glitho@ciise.concordia.ca

***Abstract*—The 3GPP IP Multimedia Subsystem (IMS) has been specified as the service delivery platform of 3G networks. It subsequently became the de facto service delivery platform of 4G networks. Cloud computing is an emerging paradigm with inherent benefits such as scalability, elasticity and easy deployment of new applications and services. Scalability and elasticity are currently among the major roadblocks to the wide scale deployment of IMS. Cloudifying IMS can help in removing these roadblocks. It will also certainly bring many other advantages. However, this cloudification is no easy task and is still in its infancy. This paper motivates the cloudification of IMS, critically review the architectures proposed so far, sketch a vision and discusses the related research challenges.**

Keywords—IP Multimedia subsystem (IMS), 3G networks, 4G networks, cloud computing, virtualization, infrastructure as service (IaaS), platform as service (PaaS, software as service (SaaS), scalability, elasticity.

I. INTRODUCTION

The standard 3GPP IP multimedia subsystem (IMS) [1] is a key component of next generation networks (NGNs). It was specified as the service delivery platform of 3G networks, then became the de facto service delivery platform of 4G networks. It is an overlay control layer on top of an IP transport layer for the seamless and robust provision of IP multimedia services to end users.

IMS is made up of two layers: a service layer and a control layer. The service layer includes application servers (ASs) (e.g. signaling session initiation protocol (SIP) AS, presence server). The key functional entity of the control layer is the call state control function (CSCF). It uses SIP to control multimedia functions.

The Home Subscriber Server (HSS), a central data base which contains user related information, is another key component of the architecture. Several functional entities of IMS service and control layers interact with it. The diameter protocol is used for these interactions.

Figure 2 depicts a simplified IMS architecture. The serving CSCF (S-CSCF) is the main CSCF of IMS. There is usually several S-CSCFs in a same IMS although the figure shows just one. There also other types of CSCF (e.g. Interrogating CSCF – (I-CSCF)) that are not shown on the figure.

Prior to the interactions with HSS via the Cx interface, the S-CSCF interacts with the Subscriber Location Function (SLF) (not shown on the figure) to select the appropriate HSS, becasue there are usually several HSSs in the same IMS networks. The same applies to the application servers. The interactions with the HSS are done via the Sh interface which is based on the diameter protocol.

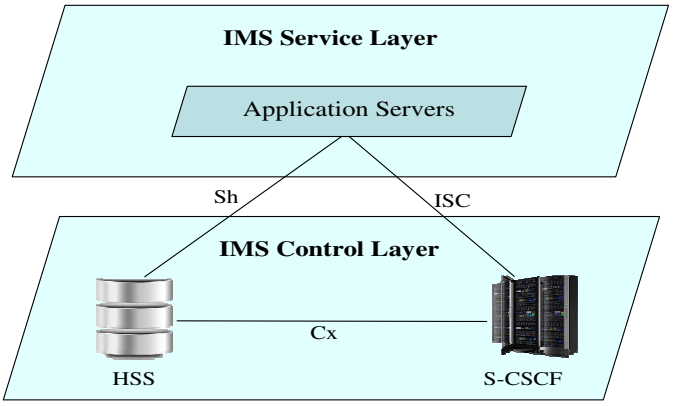

Figure 1 – A simplified IMS

Cloud computing [2] is a promising multi-facet paradigm with many inherent advantages, such as scalability and elasticity, efficient in resource usage, easy applications and services provisioning. It is often represented in the literature by the iconic diagram shown by figure 2. The diagram depicts the most critical facets the paradigm encompasses, meaning: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS).

Service providers use platforms (offered as PaaS) to provision applications and services that are offered as SaaS on a pay-per-use basis to end-users or other applications. The platforms ease the provisioning process by adding levels of abstraction to the infrastructure offered as IaaS. The infrastructure is the actual dynamic pool of resources used by the applications. Efficiency in the usage of these resources is achieved with virtualization.

Virtualization [3] is a broad computing concept that refers to the abstraction of computer resources and even network resources, thereby enabling efficiency through the sharing of these resources.

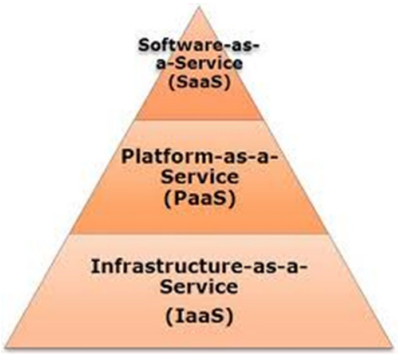

Figure 2 – Iconic representation of cloud computing

In this paper, we broadly define "cloudification of IMS" as the use of cloud computing concepts and principles for re-positioning or re-engineering IMS. The goal is generally to make IMS reap the benefits associated with cloud computing and also sometimes bring to the cloud world the advantages inherent to IMS. Products have been deployed, prototypes built, and research/general audience papers published.

The goal assigned to this paper is to answer to 2 following questions:

- Why cloudify IMS?
- How to cloudify IMS?

The next section focuses on the "why". We show that beyond the fad, cloud computing can potentially bring a viable solution to scalability and elasticity, two fundamental challenges faced by IMS. The third and fourth sections deal with the "how". In the third section, we review the solutions proposed so far and pinpoint their weaknesses. The fourth section presents our vision and discusses the related research challenges. We conclude in the last section.

We foresee paradigms such as software defined network (SDN) and network function virtualization (NFV) as powerful enablers of viable cloud architectures for IMS. However, they come with their own set of challenges.

II. CLOUDIFYNG THE IP MULTIMEDIA SUBSYSTEM: WHY?

Several papers discuss at a high level why IMS should be cloudified. However in our view, none of the reasons put forward is compelling. Reference [4] for instance claims that the use of cloud computing technology will lead to rapid developments of IMS value added services. It also claims that IMS provides the most significant opportunities for cloud computing (e.g. open standardized signaling protocol, differentiated QoS control). When it comes to reference [5] it talks of new value added services on demand.

Beyond the general discussions of the previous paragraph, we believe that scalability and elasticity are compelling reasons for cloudifying IMS. The scalability problem of IMS is well known. The proposed standard approach which consists of dynamically allocating pre-installed CSCFs and ASs to end-users has severe limitations. These limitations are now well documented in the literature [6, 7]. They are largely due to the fact that SIP is a text based protocol. Signaling delay might not be sustainable when several CSCFs and applications servers are deployed.

Elasticity of IMS has been less discussed in the literature. This is most probably due to the fact that classical Telco approach (i.e. overprovisioning) solves the problem although the approach is less and less viable. We show below that the traditional solutions proposed so far to tackle scalability in IMS leave wide open the elasticity issue. They enable an IMS which scales to some extent, but does not scale in an elastic way. We mean by traditional solutions, solutions that are not based on cloud computing.

The 3GPP user data convergence specifications [8,9] tackle to some extent the scalability issue in 3GPP systems at large, although it is not the primary objective. The primary objective is to tackle the data silo problem. They stipulate the separation of functional entities into control parts and data repositories. The control parts are called application front ends. The consistency of storage and data models is ensured through the separation.

An interesting requirement mentioned in the specifications is that control part and data storage part should scale independently. In the specific case of IMS this will certainly improve scalability because functional entities such as applications servers, CSCF and HSS can be separated into control part and data repositories as stipulated by the specifications. These application front ends and repositories will scale independently. However, this scalability will remain rather coarse grained because it will be at application front ends / data repositories level.

A few approaches have been proposed outside the standard bodies to address the same issues. Reference [10] provides an example. It proposes an early architecture in which any given node can host/ execute several IMS functional entities. This enables IMS functional entities and physical nodes to evolve dynamically based on network load, number of users, and system resources. This self- organised and adaptive architecture makes scalability possible. However, this scalability is too coarse grained because it remains at IMS functional entity level.

Reference [7] proposes another approach. It focuses on intra-domain scalability for ASs and uses the presence server as example. It provides load balancing and partitioning solutions to solve the signaling delay issues. However, it does not go beyond ASs and does tackle the control layer. Furthermore it does not address elasticity. Its granularity remains at AS level.

III. CLOUDIFYING THE IP MULTIMEDIA SUBSYSTEM: HOW? – PART I (STATE OF THE ART)

Several approaches have been proposed for cloudifying IMS. Some of them focus on specific IMS functional entities. Others deal with the entire IMS. There are also proposed solutions that consider next generation networks (NGNs) at large, including IMS. We provide a critical review below and show that they all fail in successfully tackling the two compelling issues, i.e. scalability and elasticity. They may enable some level of scalability but elasticity remains out of reach.

A. Approaches that focus on specific IMS entities

Reference [11] focuses on HSS. It proposes a distribution of HSS into a resource and a management layers. The resource layer is implemented in the cloud. Simulations are performed

to demonstrate performance gains. While the proposed solution enables an independent scaling of resource and management layers, elasticity will remain an issue because the granularity level will remain too coarse (resource/management layer level).

Research has been done in the database area (e.g. reference [12]) on scalable and elastic data bases for the cloud. However this research does not usually factor in the stringent Telco grade requirements. Furthermore it also does not address the integration of these scalable data bases with IMS functional entities. However, there is no doubt that the results of the research from the data base arena can be used as starting point in enabling elasticity in IMS data repositories. It might eventually be possible to integrate the Telco grade requirements.

Another functional entity that has attracted the attention of researchers is the presence service. Reference [13] proposes an early architecture for a virtualized presence service for future Internet. Scalability is ensured through the use of presence service substrates. However, the paper does not discuss the level of granularity of the substrates. It is therefore rather difficult to assess whether or not elasticity could be ensured.

Reference [14] also focuses on presence service. It describes a cloud -based implementation of presence service. The Eucalyptus cloud open source software [15] is used. The whole presence server is deployed on a virtual machine. Although the paper does not discuss scalability and elasticity, one can safely claim that the granularity will remain at presence server level, even if scalability is ensured.

Reference [16] deals with ASs at large and tackles elasticity to some extent. It proposes an architecture in which ASs are implemented in third party clouds and interact with the IMS core network via a broker. Each third party cloud provider assigns a group of virtual machines to given ASs.

The broker monitors resource usage and triggers specific actions when given thresholds are reached. It uses a down scaling algorithm and an up scaling algorithm which ensure some level of elasticity. New virtual machines for instance are allocated only when the average CPU utilization in the whole group is above the threshold. No new virtual machine is allocated when the average CPU utilization of a given virtual machine is above the threshold, but the average utilization in the group is still below the threshold.

Scalability is at VM level. Granularity is rather refined. However, there is no guarantee that the approach can be extended to the whole IMS. It will be rather difficult to extend it to complex IMS nodes such as the CSCF.

B. Approaches that deal with the whole IMS

Most approaches belonging to this category focus on the general problem of IMS and cloud integration. They usually do not show how scalability and elasticity could be achieved. Reference [5] discusses the overall scenarios for the integration of IMS with cloud. It assumes that IMS runs on top of the 4G Evolved Packet Core (EPC). Reference [17] provides tutorial level information on EPC.

In the first group of scenarios, IMS (with the EPC on which it runs) is re-engineered using cloud technologies. In the second group, IMS (with once again the EPC on which it runs) is used to access applications and services implemented in clouds. No architectural detail is provided. Furthermore, neither scalability nor elasticity is discussed.

Reference [4] proposes an IMS/Cloud integration scheme that falls in the second category if we use the categorization scheme of reference [5]. However it does not assume that IMS runs on EPC. The overall objective is to enable Android appliances to access high quality multimedia applications in the cloud via heterogeneous access networks (e.g. 3G, WiMAX).

The IMS QoS Policy and Charging Rules (PCRF) functional entity mediates between the heterogeneous networks and the cloud infrastructure to ensure QoS. It enables the selection of the access network that satisfies end users requirements. This is done by analyzing the traffic from the different access gateways. The focus is on QoS. The claimed benefit is "efficient and convenient communication". Here again neither scalability, nor elasticity is considered in the discussions.

The integration architecture proposed by reference [18] falls in the same second category and there is also no assumption that IMS runs on top of EPC. End-users access applications and services hosted in the cloud via an IMS CSCF. IMS scalability and elasticity are not discussed.

. *C. Approaches that deal with the whole IMS*

Reference [19] deals with 4G LTE cellular systems including access, core network and IMS. It defines the requirements a cloud computing platform should meet in order to be adequate for an implementation of 4G LTE cellular systems in the cloud.

These requirements include the support of mechanisms that enable the delivery of virtual resources as services for the execution of non-traditional instances such as eNodeB, EPC and IMS. They are used by the authors to evaluate commonly used cloud computing platforms in order to identify the missing features.

Three platforms are evaluated (i.e. Openstack [20], Eucalyptus [15] and OpenNebula [21]) and the missing features identified. This work could be used as input for selecting cloud computing platforms when prototyping cloud architectures for 4G LTE cellular systems including IMS. However, it does discuss at all the architectural issues including scalability and elasticity.

Reference [22] deals with the same 4G LTE systems. It introduces the concepts of Telco clouds and virtual Telco providers. Telco clouds are envisioned as standardized cloud environments that will enable the deployment of telecommunication software on pools of general purpose hardware deployed at key locations.

Virtual Telco will materialize the convergence between telephony services and network based computing infrastructure. The control plane will be turned into a distributed application that brings the benefits inherent to cloud (e.g. scalability, elasticity). The paper proposes a case

study for the mobility management entity (MME) of EPC. Unfortunately the applicability to IMS is not discussed at all.

IV. CLOUDIFYING THE IP MULTIMEDIA SUBSYSTEM: HOW? – PART II (VISION AND RELATED RESEARCH CHALLENGES)

We envision any architecture that brings to the IMS world all the benefits inherent to cloud computing as made up of fined grained substrates that can enable scalability. In our vision the 3GPP UDC specifications [8, 9] already discussed in this paper could be used as basis. However the two building blocks they proposed (i.e. application front ends and data repositories) will need to be further refined. The research challenges related to this vision are discussed below. The IaaS architectures and the PaaS architectures are successively considered.

A. *IaaS architectures*

Work from the database research arena such as the one already mentioned in this paper (i.e. Reference [12]) could be used as basis to enable the scalability and the elasticity of the data repositories. However, meeting the stringent telecommunication grade requirements will remain the major research challenge.

Software defined network (SDN) [23] is a paradigm that could enable scalability and elasticity in the application front ends. It decouples control and forwarding planes. The forwarding plane is remotely programmable via an open protocol.

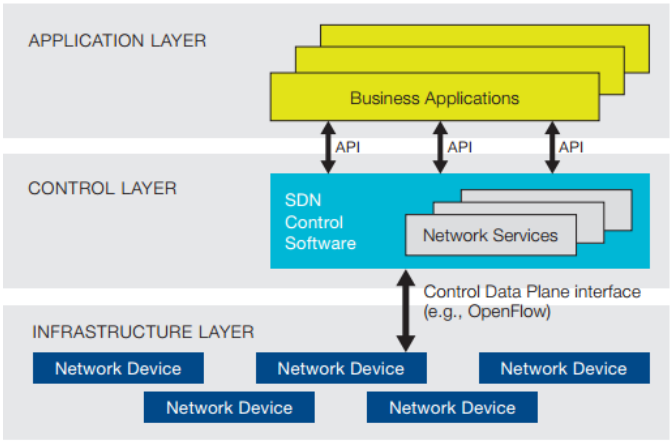

Figure 3 - SDN basic architecture (From reference [24])

Figure 3 shows the SDN reference architecture. It comprises three layers: infrastructure, control and application. The infrastructure layer contains programmable networking devices (e.g. switches) that perform packet forwarding and manipulation. The control layer includes the network OS/controller that programs the data plane and controls its resources via open protocol interfaces (e.g. Openflow).

The control enables as well network application development by exposing network capabilities to the application layer via open APIs. Both applications and devices could be virtualized to enable scalability and elasticity. One could well envision IMS application front ends further refined in application, control and infrastructure layer. However several research challenges need to be tackled before making such a vision a reality.

SDN as a paradigm is not yet very mature as explained in reference [23]. Many issues remain to be solved. The good news is that investigations have already started on the applicability of SDN to telecommunications networks, although to the best of our knowledge the specific case of IMS has not yet been considered. References [24,25] discuss some of the early results. These results could be used as starting point.

Yet another paradigm that could be considered is network function virtualization (NFV) [26]. NFV decouples network functions from the proprietary hardware on which they run. It then virtualizes them into software applications that run on general purpose hardware. The virtualized network functions can be decomposed into smaller functional blocks which can be assembled on the fly to perform the original function or even brand new functions. This will enable a high level of elasticity.

However, just like SDN, NFV is still in its infancy. Reference [27] is one of the very few papers that tackle the applicability of NFV to telecommunications networks. It focuses on the radio access network. The applicability to IMS remains a challenge and to the best of our knowledge no existing work has tackled it.

B. *PaaS architectures*

The PaaS architectures should raise the level of abstraction provided by the IaaS architectures. They should enable an easy development and management of different flavors of IMS functional entities and networks. Raising the level of abstraction of IaaSs based on paradigms such as SDN and NFV is no easy task. The reason is that approaches such as SDN and NFV usually mimic the underlying hardware and offer a very low level of abstraction.

Work has been done on how to raise the level of abstraction of SDN (e.g. reference [28]). However, this work remains embryonic. Furthermore, it does not factor in the specifics of neither telecommunications networks at large, nor IMS. A related challenge is how to make possible the use of these platforms by actors with different levels of expertise. While some of these actors might be seasoned programmers others might be less technically savvy.

A few solutions have been proposed for ordinary users in the telecommunications world (e.g. reference [29]). However to the best of our knowledge all the solutions proposed so far are applications development and management platforms that work in traditional environment (i.e. an environment which is not cloudified). It will be rather challenging to transpose them to a cloudifed environment where complex virtual resources need to be managed.

V. CONCLUSIONS

This paper has attempted to answer the why and how of IMS cloudification. It has shown that beyond the fad, it will be rather difficult to remove the scalability and elasticity roadblocks from the IMS wide scale deployment path without

using the cloud paradigm. We have provided a review of the architectures proposed so far for the cloudification of IMS and have shown that none tackles successfully the scalability and elasticity issues. A vision has also been proposed and the related research challenges have been discussed. In the vision the 3GPP UDC specifications are used as pillars. Paradigms such as SDN and NFV can be used to further refine the substrates proposed by the 3GPP UDC specification. This will enable an IMS that scales in an elastic way. However, there is still a long way to go because there are many inherent research challenges that need to be dealt with.

REFERENCES

[1] G. Camarillo and M.A. Garcia-Martin, The 3G IP Multimedia Subsystem (IMS), Third Edition, 2008

[2] L. M. Vaquero et al., "A Break in the Clouds: Towards a Cloud Definition", ACM SIGCOMM Computer Communication Review, Vol. 39, No1, January 2009

[3] A. Khan et al., "Network Virtualization: A Hypervisor for the Internet?," IEEE Communications Magazine., vol. 50, no. 1, Jan. 2012, pp. 136–43.

[4] W Zhang et al., Architecture and Key Issues of IMS-based Cloud Computing, 2013 IEEE Sixth International Conference on Cloud Computing, Santa Clara, July 2013

[5] F. Gouveia et al., Cloud Computing and EPC/IMS Integration: New Value Added Services on Demand, Mobimedia Conference, Sept. 2009, London, UK

[6] P. Agrawal et al, IP Multimedia Subsystems in 3GPP and 3GPP2: Overview and Scalability Issues, IEEE Communications Magazine, January 2008

[7] P. Bellavista et al, Enhancing Intra-domain Scalability of IMS-Based Services, IEEE Transactions on Parallel and Distributed Systems, Vol 24, No12. December 2013

[8] 3GPP Technical Report 22.985 V11.0.0, Services Requirements for User Data Convergence, http://www.3gpp.org/DynaReport/22985.htm

[9] 3GPP Technical Specification TS 123 335 v9.1.0, User Data Convergence, Technical Realisations and Information Flows, http://www.3gpp.org/DynaReport/23335.htm

[10] A. Dutta et al., Self-Organizing IP Multimedia Subsystem, International Conference on Internet Multimedia Services Architecture and Applications 2009 (IMSAA 2009), December 2009, Bangalore, India

[11] T. Yang et al., A New Architecture of HSS Based on Cloud Computing, 13th IEEE Conference o Communications Technology (ICCT), 2011, Sept. 2011, Jinan, China

[12] S. Das et al., ElassTraS: An Elastic, Scalable, and Self-Managing Transactional Database for the Cloud, ACM Transactions on Database Systems, Vol. 38, No.1, April 2013

[13] F. Belqasmi, N. Kara, R. Glitho, A Novel Virtualized Presence Service for Future Internet, IEEE International Conference on Communications (ICC) 2011, Workshop T2, Future Network, Kyoto, Japan, June 2011

[14] W. Quan et al., Research on Presence Service Testbed on Cloud Computing Environment, 2010 3th IEEE International Conference on Broadband Network and Multimedia Technology (IC-BNMT 2010), October 2010, Beijing, China

[15] Eucalyptus Open Source Cloud Software https://www.eucalyptus.com/

[16] P. Bellavista et al., QoS Aware Cloud Brokering for IMS Infrastructures, IEEE Symposium on Computer and Communications (ISCC 2012), July 2012, Cappadocia. Turkey

[17] M. Olsson et al., SAE and the Evolved Packet Core: Driving the Mobile Broadband Revolution, Elsevier, 2009

[18] J.-L Chen et al., IMS Cloud Computing Architecture for High Quality Multimedia Applications, 7th International Communications and Mobile Computing Conference (IWCMC 2011), August 2011, Istanbul, Turkey

[19] A.J. Staring and G. Karagiannis, Cloud Computing Models and Their Application in LTE based Cellular Systems, IEEE International Conference on Communications 2013 (ICC'13), 1st International Workshop on Mobile Cloud Networking, Budapest, Hungary, June 2013

[20] OpenStack https://www.openstack.org/

[21] OpenNebula http://opennebula.org/

[22] P. Bosch et al., Telco Clouds and Virtual Telco: Consolidation, Convergence, and Beyond, 6th IFP/IEEE International Workshop on Broadband Convergence Networks, May 2011, Dublin, Ireland

[23] S. Sezer et al., Are we Ready for SDN? Implementation Challenges for Software Defined Networks, IEEE Communications Magazine, Vol. 51, No7, July 2013

[24] J. Kempf, B. Johansson, S. Pettersson, H. Lüning, and T.Nilsson, "Moving the mobile Evolved Packet Core to thecloud," in *Wireless and Mobile Computing, Networking and Communications (WiMob), 2012 IEEE 8th International Conference on, 2012, pp. 784–791.*

[25] A. Basta et al., A Virtual SDN-enabled LTE EPC Architectecture: A Case Study for S-/P-Gateway Functions, IEEE SDN Forum, November 2013

[26] E. Patouni, Network Virtualization Trends: Virtually Anything Is Possible by Connecting the Unconnected, IEEE SDN forum for Future Networks and Services, 2013

[27] P. Demestichas et al., 5G on the Horizon: Key Challenges for the Radio Access Network, IEEE Vehicular Technology Magazine, Vol. 8, Issue: 3, September 2013

[28] N. Foster et al., Languages for Software Defined Networks, IEEE Communications Magazine, Vol. 51, Issue: 2, 2013

[29] N. Laga, E. Bertin, R. Glitho, N. Crespi, Widgets and Mechanisms for Service Creation by Ordinary Users, *IEEE Communications Magazine*, March 2012