Mobile Edge Computing – an important ingredient of 5G Networks

Lav Gupta and Raj Jain, Washington University in St. Louis
H. Anthony Chan, Huawei Technologies, USA

1. Networks of the future

5G and its timeline

5G is a collective name for technologies and methods that would go into the future networks to meet the extreme capacity and performance demands. The phrase 'no latency, gigabit experience' summarizes the user expectations that the industry is aspiring to meet. Both of the major standardization bodies, International Telecommunications Union (ITU) and European Telecommunications Standards Institute (ETSI) have initiated activities relating to 5G [ITU15][ITU13 IMT2020][ETSI15] with commercial deployments expected in 2020.

What can users expect?

Some of the key performance parameters targeted to be achieved in 5G networks are: per device data rates up to 20 Gbps, less than 1ms latency contribution of the radio part, mobility at 500 km/hour and terminal localization within 1 meter. It will aim for service continuity in trains, sparse and dense areas, support for connecting 20 million user devices and more than a trillion Internet of Things (IoT)/Machine to Machine (M2M) devices with high reliability.

Technologies relevant for 5G

The design of 5G networks would revolve around virtualization and programmability of networks and services. It is envisioned that transition to 5G will be facilitated by today's emerging technologies such as Software Defined Networking (SDN), Network Functions Virtualization (NFV), Mobile Edge Computing (MEC) and Fog Computing (FC) [YI15]. SDN and NFV provide new tools that enhance flexibility in designing networks. These complementary technologies enable programmability of control and network functions and eventual migration of these key constituents of the network to the cloud. In the next section we focus on MEC, the central theme of this article.

2. Mobile edge computing

What is MEC?

Relevance of cloud computing to mobile networks is on an upward spiral. Social network services like Facebook and Twitter, the content from YouTube and Netflix, and navigation tools from Google Maps are all on clouds. Besides, users' increasing reliance on mobile devices to carry out compute and storage intensive operations, whether personal or business related, require offloading to the clouds for achieving better performance extending battery life. These objectives would be difficult and expensive to realize without bringing the cloud closer to the edge of the network and to the users. In response to this requirement the mobile operators are working on Mobile Edge Computing (MEC) in which the computing, storage and networking resources are integrated with the base station. Compute intensive and latency sensitive applications like augmented reality and image processing can be hosted at the edge of the network. Fig. 1 shows this concept.

eprint
IEEE Softwarization Newsletter, March 2016

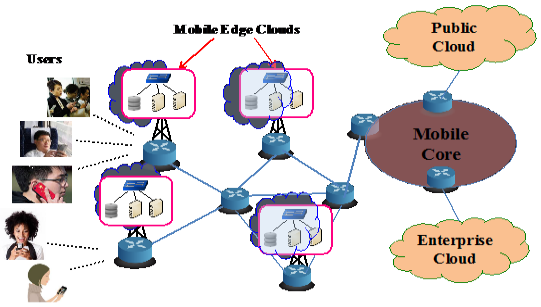

Fig. 1 Mobile Edge Clouds

MEC and application splitting

Mobile-edge computing provides a highly distributed computing environment that can be used to deploy applications and services as well as to store and process content in close proximity to mobile users. This would enable applications to be split into small tasks with some of the tasks performed at the local or regional clouds as long as the latency and accuracy are preserved. A number of challenging issues arise in distributing sub-tasks of an application among edge and other clouds. When splitting of an application does happen, the mobile edge cloud takes care of the low latency, high bandwidth and locally relevant jobs.

The MEC Server platform

The key element of MEC is its Commercial-Off-The-Shelf (COTS) application server, which is integrated with the base station. The MEC server provides computing resources, storage capacity and connectivity as traditional cloud infrastructure would. Additionally, it provides access to user traffic and radio network information that can be used by application providers to tailor their applications and services for enhanced user experience. It hosts software for real-time analytics and machine-intelligence. It can serve queries from devices that require response times of below 100 ms. Offline or batch processing, data intensive and high latency tasks are relegated to larger clouds. Important constituents of the MEC server are shown in Fig. 2 [adapted from ETSI14].

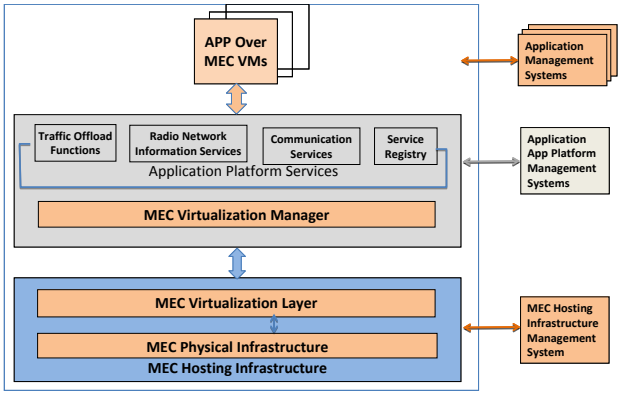

Fig. 2 MEC Server Platform

MEC for monetizing the 5G RAN

Traditionally, the radio access network (RAN) has been the ‘dumb pipe’ for voice calls and data. In the 5G network, the operators would be able to make these pipes ‘intelligent’ by overlaying distributed edge cloud computing onto the RAN. With virtualization at the edge of the RAN, the mobile network operators can allow multiple third party tenants at the base

station. The application providers have two main incentives to host their applications, or a suitable sub-division of it, on the edge – first, they get ultra-low latency (which is colloquially referred to as zero latency) and high-bandwidth. Secondly, the Radio Network Information Service (RNIS) module of the MEC server gives them real time network information about the cell load, subscriber specific bandwidth, and subscriber location. This way the mobile operator can take load off the core network, reduce congestion and make more money out of the edge network.

Managing the edge clouds

From the point of view of application service providers, deploying and managing distributed applications across multiple clouds is a difficult proposition. It becomes very difficult for the providers to co-ordinate with individual clouds service providers each with their own interfaces and inter-cloud network providers to manage their application. They need a versatile application deployment and management platform to be able to optimize use of resources, ensure performance and contain cost. We are working on an open source management platform called MCAD (Multi-cloud Application Delivery) that would allow application and 5G service providers to specify multi-cloud virtual resource deployment policies, create virtual resources, deploy services in the most appropriate cloud(s) and manage them while in operation. Previously named AppFabric [PAU14], the platform will communicate with various cloud/network management systems and find the optimal locations for virtual resources (virtual machines, storage and virtual network functions) based on the required cost and performance criteria of an application.

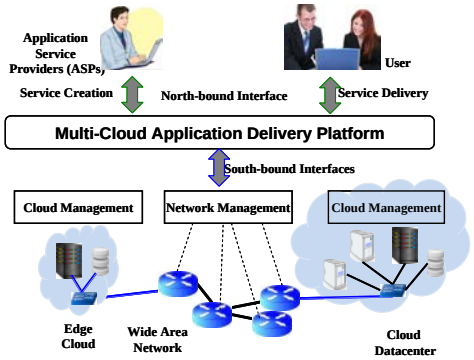

Fig. 3 The MCAD Based Multi-cloud Management Architecture

As shown in Fig. 3, MCAD design features a hybrid control system with a central global controller and per cloud/network local controllers. MCAD allows the cloud and networking resource owners to exercise complete control over their resources while tenants host their applications and program their policies on virtual resources anywhere on the participating clouds.

3. Summary

In conjunction with SDN and NFV, MEC will be crucial in effecting low latency, high bandwidth and agility that will be able to connect trillions of devices. To keep pace, multi-cloud application delivery platforms will be able to efficiently amalgamate resources from edge and large clouds and take real time network conditions into account for maximal user experience.

References

1. [ITU15] Work plan, timeline, process and deliverables for the future development of IMT, ITU 2015
2. [ITU13-IMT2020] -High level 5G architecture, Network softwarization, gaps; http://www.itu.int/en/ITU-T/focusgroups/imt-2020/Pages/default.aspx
3. [ETSI15] 5G- The 5G Infrastructure Public Private Partnership: the next generation of communication networks and services, ETSI, 2015
4. [ETSI14] Mobile-Edge Computing – Introductory technical whitepaper, ETSI, 2014
5. [YI15] S. Yi, C. Li, Q. Li, “A Survey of Fog Computing: Concepts, Applications and Issues,” ACM Mobidata’15, 2015
6. [PAU14] Subharthi Paul, Raj Jain, Mohammed Samaka, Jianli Pan, "Application Delivery in Multi-Cloud Environments using Software Defined Networking," Computer Networks Special Issue on cloud networking and communications, Volume 68, 5 August 2014, Pages 166–186

Lav Gupta is a senior member of IEEE. He received BS degree from Indian Institute of Technology, Roorkee, India in 1978 and MS degree from Indian Institute of Technology, Kanpur, India in 1980. He is currently pursuing PhD degree in Computer Science & Engineering at Washington University in St Louis, Missouri, USA.

He has worked for about 15 years in the area of telecommunications planning, deployment and regulation. With the sector regulatory authority he worked on technology and regulation of next generation networks. He was recipient of best software award from Computer Society of India in 1982 and best faculty award at Etisalat Academy, UAE in 1998.

Raj Jain is a Fellow of IEEE, a Fellow of ACM, a Fellow of AAAS, a winner of 2015 A.A. Michelson Award, 2006 ACM SIGCOMM Test of Time award and ranks among the top 90 in Citeseer's list of Most Cited Authors in Computer Science. Dr. Jain is currently a Professor of Computer Science and Engineering at Washington University in St. Louis. Previously, he was one of the Co-founders of Nayna Networks, Inc - a next generation telecommunications systems company in San Jose, CA. He is the author of ``Art of Computer Systems Performance Analysis," which won the 1991 ``Best-Advanced How-to Book, Systems" award from Computer Press Association.

H. Anthony Chan received his Ph.D. in physics at Univ. of Maryland, College Park in 1982 and then continued post-doctorate research there in basic science. After joining the former AT&T Bell Labs in 1986, his work moved to electronic packaging and reliability, and then moved again to network architecture and standards. He was professor at University of Cape Town during 2004-2007. His current research in Huawei Technologies is in emerging broadband wireless network technologies. Dr. Chan is a Fellow of IEEE and is a distinguished lecturer of IEEE ComSoc and of IEEE Reliability Society