Dynamic QoS/QoE assurance in realistic NFV-enabled 5G Access Networks

Conference Paper · July 2017

DOI: 10.1109/ICTON.2017.8025149

4 authors:

Jose-Juan Pedreno-Manresa
ADVA

Pouria Sayyad Khodashenas
Huawei Technologies

Muhammad Shuaib Siddiqui
i2CAT Internet and Digital Innovation in Catalonia

Pablo Pavon-Marino
Universidad Politécnica de Cartagena

Dynamic QoS/QoE Assurance in Realistic NFV-enabled 5G access Networks

Jose-Juan Pedreno-Manresa[1], Pouria Sayyad Khodashenas[2], Muhammad Shuaib Siddiqui[2] and Pablo Pavon-Marino[1]
[1] *Universidad Politécnica de Cartagena, Cuartel de Antiguones, Plaza del Hospital 1, 30202 Cartagena, Spain*
[2] *Fundació I2cat, C/ Gran Capità 2-4, Nexus I Building, 2nd floor, Office 203, Barcelona 08034, Spain*
e-mail: josej.pedreno@upct.es

ABSTRACT
Fifth generation mobile communication (5G) networks promise lower latency, higher traffic volume and data rates compared to what we see nowadays. Among all technologies, Network Function Virtualization (NFV) and Software Defined Networking (SDN) techniques over the cloud-enabled radio systems play an important role to enable resource pooling, scalability, layer interworking and spectral efficiency. Despite the potential benefits, there is always an additional challenge on assuring quality of service and user experience, especially on real life scenarios targeted by 5G. In this paper, we initially frame the dynamic QoS/QoE assurance problem; then with the help of simulation on a real life scenario, where ~100,000 persons get in and out of the Camp Nou stadium in Barcelona, we demonstrate the efficiency of our proposed QoS/QoE assurance algorithm.
Keywords: 5G, Quality of service, Quality of experience, Dynamic monitoring, NFV, Net2Plan

1. INTRODUCTION

From an operator's perspective, evolution of mobile networks implies a continuous attempt to find a trade-off between capacity increase (i.e. higher bitrates, more coverage, etc.) and total cost of ownership (i.e. CAPEX, OPEX). A viable solution points to the concept of flexible Radio Access Network (RAN) with simplified deployment and management. Although today's RAN is able to provide a high level of configurability (e.g. support for a variety of transport network and baseband configurations), there is still plenty of room to exploit the potential synergies with concepts like Network Function Virtualization (NFV) and Software Define Networking (SDN). Joint radio-cloud architecture is an effort to place intelligence at the network edge and to use virtualization technologies to build a cost-, spectrum-, and energy-efficient RAN ecosystem able to offer added value services and improved user experience [1].

The mixed radio-cloud environment poses extra challenges for service management and orchestration, especially on assuring the Quality of Service (QoS) on dynamic scenarios where the traffic volume and performance expectations changes over time. In principle, to handle the situation correctly, the logical cloud-enabled RAN manager/orchestrator needs to simultaneously take into account both, the radio status (i.e. volume of traffic, geographical distribution of traffic, etc.) and the cloud capabilities (i.e. available IT resources, VM to VM communication requirements, etc.) for all the action related to the service lifecycle management. To do so, forming a feedback QoS ensuring loop is inevitable. Such a loop consists of three main steps:

- Monitoring: a phase in which performance monitoring parameters are collected from the radio/cloud elements (e.g. SC physical network function, virtual machines, etc.) and handed over to the next step (decision-making). Depending on the nature of the resources, i.e. radio or cloud, QoS requirements, the available Service Level Agreement (SLA), etc. the monitoring parameters might vary from one use case to another.
- Decision-making: a phase in which performance metrics collected in the previous step are processed. Depending on the situation and available resources, a decision will be taken to ensure the level of QoS (with the help of a dedicated algorithm). In principle, the nature of such a decision-making algorithm can range from a greedy heuristic to a complex cognitive form.
- Reaction: upon making a decision, the management/orchestration system needs to coordinate the interaction with the other lower level modules such as Element Management System (EMS), Virtual Network Function Manager (VNFM) and Virtual Infrastructure Manager (VIM) to react appropriately.

This paper highlights the service management and orchestration challenges that arise when trying to guarantee the QoS over the C-RAN environment. Next, with the help of simulation we prove that how our proposed algorithm is able to guarantee the acceptable QoS/QoE level.

2. PERFORMANCE METRICS

In a joint cloud RAN system, two main categories of performance metrics needs to be collected. First, the radio performance measurements which are typically divided into a number of sub-groups [2], e.g., i- RRC connection related measurements, ii- E-RAB related measurements, iii- handover measurements, iv- cell level radio bearer QoS parameters, v- radio resource utilization related measurements, vi- UE-associated logical S1-connection measurements, vii- paging related measurements, viii- measurements related to equipment resources, ix- common LAs of overlapping RAT's coverage, x- RF Measurements. Parameters like these can be collected and processed by the cloud-enabled RAN management system to maintain a good overall level of QoS.

Second, the cloud monitoring parameters which are divided into three main sub-groups: i- Network Functions Virtualization Infrastructure (NFVI), ii- Network Functions Virtualisation (NFV) virtual machines preforming specific network functionalities and iii- SDN-based monitoring for virtual links. NFVI is the actual hardware plus the virtualization layer, i.e. hypervisor, used to host Network Services (NS) – a chain of VNF to provide an added-value service. Therefore, it is important to keep track of it. Key Performance Metrics (KPI) can be used for this purpose are: i- CPU utilization: CPU carries out the instructions of VNF, performing the arithmetical, logical, and input/output operations. Thus, it is important to monitor its utilization and possibility keep it as low as possible, ii- RAM utilization: RAM is used to load information required by VNFs for faster access thereby improving the overall performance. If a server runs out of RAM, a portion of the hard drive can be dedicated as the virtual memory. This process is called swapping, which will cause performance degradation since the hard drive is much slower than RAM (e.g. 1000 times slower), iii- Hard Disk Drive (HDD) utilization: the operation system kernel, hypervisor and agents on the server needs space on the HDD for normal operating processes including paging files and certain caches. The application running on the server, i.e. VNFs, also needs space to write temporary data to cache for efficient operation as well as permanent data that will be accessed by the user. Thus, low free space on a HDD might cause performance issues, iv- System hardware: server might include other devices such as HWA (e.g. GPU, FPGA, DSP, etc.), CPU fan, power supply, etc. that affect its overall performance. Health performance parameters of hardware as well as metrics such as temperature, air flow and humidity needs to be also monitored.

European Telecommunications Standards Institute (ETSI) in [3] divided the monitoring matrices of NFV into two main categories: i-VNF monitoring parameters, ii- NS performance metrics. From a conceptual perspective, a VNF is a virtual machine (VM) that runs a specific application inside which permits preforming the same functionality as a network middlebox (e.g. virtual packet gateway (vPGW) in comparison to the actual PGW hardware). ETSI in [3] suggested a way to specify different deployment flavours for the VNF in VNF descriptor (VNFD). These parameters can be either VM related information, e.g. CPU utilization, bandwidth consumption, etc. or VNF specific such as, calls per second, number of subscribers, number of rule flows per second, VNF downtime, etc. As mentioned previously, one or more of these parameters could be influential in triggering a reaction on the QoS loop. At NS level, monitoring parameters represent metrics that are tracked to check the level of NS compliance with the agreed SLAs (e.g. NS downtime). These parameters will be part of NS descriptor (NSD) as ETSI suggests in [3] and can be used for specifying different deployment flavours of an NS and/or to indicate different levels of NS availability. Examples for these parameters are: calls per second, number of subscribers, number of rules, flow per second, etc.

For the third monitoring category, i.e. the SDN-based, a deep-dive into the details of the SDN concept is necessary. SDN by definition provides control of the data plane flows and using southbound protocols, e.g. OpenFlow, it is able to count matching occasions. Counters can be per-table, per-flow, per-port and per queue. In principle, the counters may be implemented in software and maintained by polling hardware counters with more limited ranges. This input can be used by the cloud-enabled RAN management system as monitoring input to trigger actions on a QoS ensurance loop.

3. SIMULATION SCENARIO

In order to simulate a real like scenario with challenging QoS/QoE assurance and traffic volume dynamicity, we consider a radio access network topology in a densely populated area (Barcelona city centre) and stablish a time window around a popular football match. During football matches, several thousand people travel to the Camp Nou stadium, being this the perfect example of offered traffic and performance expectations changing rapidly.

We selected an area of 1.5km of radius with Camp Nou at its centre. Two different cell sites are considered: macro-cells are placed in the area using real location data [5] and we assume the presence of micro-cells equally distributed inside the stadium. All macro-cells are directly connected to a Core Central Office (CO), while micro-cells are directly connected to the macro-cell located at the stadium. We assume macro-cells, micro-cells and CO have hardware capabilities to instantiate VNFs.

At the beginning of the simulation, individual persons are situated at the edges of the scenario, around metro, bus stations, parking, important streets, and avenues. Then, each person walks toward the stadium at a random speed. Once everyone is inside the stadium, the match begins and all people remain stationary until it ends; at which moment each person commences to walk

Service	Chained VNFs *	Bandwidth req.
VoIP	NAT-FW-TM-FW-NAT	250 Kbps
Video Conference	NAT-FW-TM-VOC-IDPS	2 Mbps
Web Service	NAT-FW-TM-WOC-IDPS	4 Mbps

* IDPS: *Intrusion Detection Prevention*, FW: *Firewall*, NAT: *Network Address Translation*, TM: *Traffic Monitor*, VOC: *Video Optimization Controller*, WOC: *WAN Optimization Controller*

Table 1. Service chains considered for each service and bandwidth requirements.

again following the inverse travelled path until they reach their origin points. The simulation ends when everybody has reached their destination.

Throughout all the simulation, each individual has a probability (based on the total offered traffic, as an input parameter) to request a service: Web Service, VoIP or Videoconference. Table 1 shows the service requirements, each one chaining different VNFs in sequential order. The management logic assumes each services is handled by the nearest cell, following a known handover mechanism [6] and for the sake of simplicity, each cell has enough radio capabilities [7].

We developed an allocation algorithm to ensure that each service request can be allocated meeting both bandwidth and chained VNFs requirements. The algorithm pseudo-code is presented below:

```
SET servingSite = Nearest cell to the requester
FOR each ServiceRequest at time T
   IF Link to CO has NOT enough capacity to allocate the Offered Traffic
      Block Request (END)
   END IF
   FOR each VNF in Chain Associated to ServiceRequest
      IF existing instance of VNF with enough capacity to allocate a new session
          Allocate new session in existing VNF instance
         GOTO next VNF
      END IF
      IF enough hardware capacity to create new VNF instance
         Create new instance of VNF
         GOTO next VNF
      ELSE
         IF servingSite == Central Office
            Block Request (END)
          END IF
         SET servingSite = Central Office
      END IF
   END FOR
END FOR
```

3.1 Results

The simulation was performed using the online simulation tool from the open-source network planner Net2Plan [8][9] with the following input parameters: 99,354 persons (stadium maximum capacity), 56 macro-cells, 200 micro-cells, 10 Gbps bidirectional links between each macro-cell and CO and 1Gbps link between each micro-cell and the Camp Nou macro-cell. The Camp Nou macro-cell has a 50 Gbps link with the CO. For simplicity, we have considered a football match duration of 5 minutes. We assume each macro-cell site cabinet contains a server with 16 CPU cores, 64 GB of RAM and 10 TB of HDD capacity. Micro-cell sites have a server with 8 CPU cores, 32 GB of RAM and 7 TB of HDD capacity [10]. Hardware requirements and concurrent operation per VNF is shown in Table 2. The total offered traffic is 60 Gbps and each service request has a probability of 50% of being Web Service, 30% for VoIP and 20% Video Conference; all three services have a holding type of 100 seconds.

VNF	# of concurrent operations	Hardware req.
IDPS	2500	CPU: 2 cores, RAM: 2GB, HDD: 10GB
FW	2500	CPU: 2 cores, RAM: 3GB, HDD: 5GB
NAT	3000	CPU: 1 core, RAM: 1GB, HDD: 2GB
TM	2500	CPU: 1 core, RAM: 3GB, HDD: 2GB
VOC	1000	CPU: 2 cores, RAM: 2GB, HDD: 20GB
WOC	1500	CPU: 1 core, RAM: 2GB, HDD: 10GB

Table 2. VNFs hardware requirements and number of concurrent operations per instance.

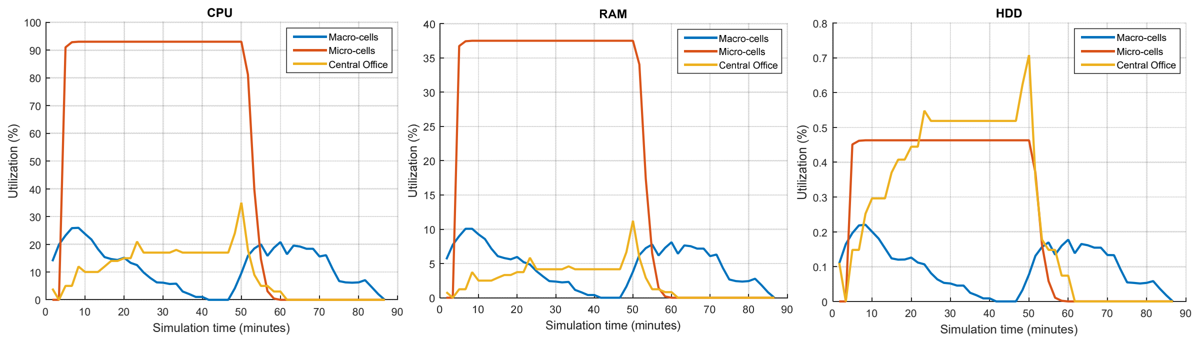

Figure 1. Hardware utilization (CPU, RAM and HDD) on macro-cell, micro-cell and Central Office sites.

Fig. 1 shows the aggregated utilization by all VNFs chains through the simulation. The first conclusion we can infer is that CPU is a bottleneck resource, especially in micro-cell sites. This fact in addition to the high density of users around the micro-cells during most of the simulation substantiate the high utilization of all three resources. We can also observe that at the beginning and the end of the simulation, while the utilization on the macro-cells is at its maximum, the opposite effect occurs in the Central Office. There is a peak on utilization around minute 50, because when the match ends and people exit the stadium all macro-cells around Camp Nou are at maximum utilization, having the Central Office to take a higher load in terms of VNFs instantiation.

Also, while the average blocking of the simulation is <10%, as can be seen in Fig. 2, there is a sudden increase with a maximum of 44% during the match. We can fairly assume that the other bottleneck is the links between the micro-cells and the macro-cell, are in fact one order of magnitude inferior in terms of bandwidth capacity.

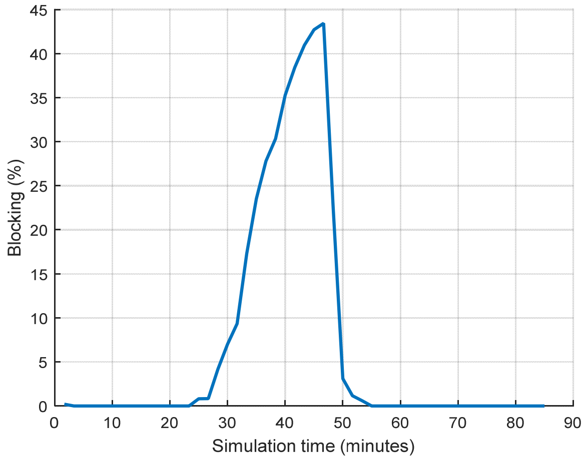

Figure 2. Instant blocking

4. CONCLUSIONS

With the coming of the 5G era, it is more important than ever to provide a good QoS/QoE. While facing new challenges like providing lower latency, higher data rate to emerging services or an increasing dynamicity in traffic volume. This work frames this problem proposing a dynamic simulation scenario, providing an algorithm for chain service allocation considering the aforementioned challenges. We demonstrate that with careful management and orchestration we can save essential resources from the central office while providing better latency to the users.

ACKNOWLEDGEMENTS

This work was supported by the European Union under the frameworks of H2020 SESAME (no. 671596), H2020 METRO-HAUL (no. 761727), the FPU fellowship program of the Spanish Ministry of Education, Culture and Sports (ref. no FPU14/04227), the Spanish project grants TEC2014-53071-C3-1-P (ONOFRE) and TEC2015-71932-REDT (ELASTIC).

REFERENCES

[1] B. Blanco, *et. al*, "Technology Pillars in the Architecture of Future 5G Mobile Networks: NFV, MEC and SDN," Elsevier Computer Standards & Interfaces, Available online 4 January 2017.

[2] 3GPP TS 32.425 "Performance Management (PM); Performance measurements Evolved Universal Terrestrial Radio Access Network (E-UTRAN)," V.10.6.

[3] "NFV Management and Orchestration - An Overview", GS NFV-MAN 001 V1.1.1, European Telecommunications Standards Institute (ETSI), 2014.

[4] S. Wang, S. Liu, C. Chou, "Design, implementation and performance evaluation of software OpenFlow flow counters in a bare metal commodity switch", *IEEE ISCC*, 2016.

[5] Spanish Ministry of Energy, Tourism and Digital Agenda – Information about Radioelectric Installations and Radiation Exposure Levels. [Online] Available: https://geoportal.minetur.gob.es/VCTEL/vcne.do [Last accessed: April 2017]

[6] C.-C. Lin, "Handover Mechanisms in 3GPP Long Term Evolution (LTE)", Ph.D. dissertation, University of Technology, Sydney 2013

[7] 3GPP TR25.814 version 7.1.0, "Physical Layer Aspects for Evolved Universal Terrestrial Radio Access (UTRA) (Release 7)," 3GPP, Tech. Rep. September 2006

[8] Net2Plan – The open-source network planner. [Online] Available: http://www.net2plan.com [Last accessed: April 2017]

[9] P. Pavon-Marino and J.-L. Izquierdo-Zaragoza, "Net2Plan: An Open Source Network Planning Tool for Bridging the Gap between Academia and Industry," IEEE Network, vol. 29, no. 5, pp. 90–96, Sep. 2015.

[10] DOLFIN FP7 Project Deliverables - D2.2 - DOLFIN requirements and system architecture. [Online] Available: http://www.dolfin-fp7.eu/?page_id=157 [Last Accessed: April 2017]