Optimization of Multi-server Video Content Streaming in 5G Environment

Eugen Borcoci, Tudor Ambarus, Joachim Bruneau-Queyreix, Daniel Negru, Jordi Mongay Batalla

Optimization of Multi-server Video Content Streaming in 5G Environment

Eugen Borcoci, Tudor Ambarus
University POLITEHNICA of Bucharest
Bucharest, Romania
emails: eugen.borcoci@elcom.pub.ro,
tudorambarus@yahoo.com

Joachim Bruneau-Queyreix, Daniel Negru,
LaBRI Lab, University of Bordeaux, Bordeaux, France
emails: joachim.bruneau-queyreix@labri.fr
daniel.negru@labri.fr
Jordi Mongay Batalla, National Institute of
Telecommunications Warsaw, Poland
email: jordim@interfree.it

***Abstract* — This paper presents a preliminary work, proposing an architectural control plane solution for optimization of multiple-server video content streaming in 5G wireless environment. The starting point was an existing video streaming delivery system, having a light, over-the-top (OTT) architecture, which performs for each client request of a content object, an initial selection among multiple servers and paths, then followed by in-session dynamic media adaptation. This work extends the above system concepts to a different environment, i.e., heterogeneous Cloud Radio Access Network and cooperation with Mobile Edge Computing. The proposed solution can support recently developed multi-server and multi-path dynamic adaptive streaming over HTTP.**

Keywords — Content delivery; 5G; Server selection; Path selection; C-RAN; DASH.

I. INTRODUCTION

Content, media and especially video traffic have become significant part of Internet and integrated networks traffic, including mobile one and will still grow in the next years. Estimations show [1][2], that in 5G networks, the data rate required for a mobile user equipment (MUE) will have to increase to 10 Mbps or more for high-definition (HD) video service, and 100 Mbps for ultra-high-definition TV (UHDTV), in various mobility scenarios. Other applications (e.g., 3D video conferences) might require even higher transmission rates up to 10 Gbps. Some forecast [3] show that video traffic (e.g., TV, video on demand, Internet video streaming, peer to peer) is estimated to become between 80 and 90 percent among overall consumer traffic.

On the other side, among the strong requirements to be addressed by 5G [1], some are related to very low end-to-end (E2E) latency (few milliseconds) especially for critical communications. For media video streaming, this requirement could be met by applying content delivery networks (CDN) - like techniques [4], i.e., placing in an intelligent way content servers and replica servers, in the proximity of communities of end users. The content objects are cached in several servers, based on criteria as content popularity, time-life, CDN provider policies, etc. One challenge to be solved in 5G is to decide the locations where to locate the original and caching servers. The solution can be also determined by the 5G architecture adopted for the Radio Access Network (RAN) and also for the core network, which aggregates and performs control of several heterogeneous RANs [2][5].

This paper proposes a control plane architectural solution for video content delivery optimization, applicable in 4G and or 5G networks environment, if several (multiple) content servers (and/or caching) and paths are available, working to serve a given user. Note that the algorithms and procedures to place the servers and then to place/store/replace the media objects and also the dynamic control of the time-life of the media objects in these servers do not constitute the target of this work.

The starting point of this work has been a previously designed light architecture system [6-8], for efficient video streaming and delivery, acting in Over-the-Top (OTT) style, i.e., controlling only a Content Server and User/Client functionalities and working on top of the current multiple-domain Internet. The system operation is based on collaboration between several entities: a Service Provider (SP), several Content Servers (CS) and the End User (EU). An assumption is valid: the geographical locations of servers and mapping of different media objects to servers are known by the SP management entity. When a user request for a media content object arrives to SP entity, the system performs an *initial selection among multiple servers and paths* pairs. Then, during the video streaming session, two methods have been used to preserve/enhance the Quality of Experience (QoE) perceived by the user: *media flow adaptation* (adaptive streaming protocols) and/or *server switching*. For the video session phase, the Dynamic Adaptive Streaming over Hypertext Transfer Protocol-HTTP (DASH) [10][11], has been selected and implemented.

The novel contribution of this paper consists in the following aspects. *First* it extends the initial system concepts (shortly described above, and detailed in [6-8]), to novel network environment like 5G having a Cloud Radio Access Network (C-RAN) – based architecture, and possibly including Mobile Edge Computing (MEC) capabilities. *Second,* for server selection phase it is considered not only an OTT approach but an extension; the network status and channel information, existent at RAN level is used as additional input in the overall optimization algorithm. *Third*, the system proposed here supposes not only a single-server-at-a-time selection, but multiple servers, allowing a single

client (see Multi-description DASH in [9]) to receive streams in parallel from several servers.

The paper structure is the following. Section II is a short overview of related work. Section III outlines the overall 5G environment architecture based on C-RAN and MEC concepts. Section IV discusses some multi-server content delivery problems in 4G or 5G environment and introduces the architecture proposed for C-RAN and MEC contexts. Section V is focused on multiple server selection based on multi-criteria algorithms. Section VI contains conclusions and future work outline.

II. Related Work

Media/content delivery systems over the current public Internet frequently use light OTT architectures. They are more simple and cheap, in comparison with complex solutions involving network resources management and control, like - CDNs [4] or Content Oriented Networking [12].

The work presented in [6-8] has proposed and developed an OTT-style content streaming system (named DISEDAN), having as business actors the SP, (owning several Content Servers - CS) and EUs, which consume the content. The SP basically delivers content in OTT style (however, an SP might own and manage a transportation network). The solution consists in: (1) *two-step server selection* (first at SP side and then at EU side) based on multi-criteria optimization algorithms that consider context- and content-awareness and (2) *in-session,* so-called *dual adaptation*, consisting of media adaptation and/or content source adaptation (i.e., streaming server switching) when the quality observed at EU suffers degradation.

For in-session adaptation, the DASH technology has been selected. It is attractive because it uses conventional HTTP Web servers [10][11]. The DASH minimizes server processing power and is video codec agnostic. A DASH client continuously selects (on-the-fly) segments having the highest possible video representation quality that ensures smooth play-out, in the current downloading conditions.

The basic variant of the system presented above (i.e., pure OTT style and standard DASH) has limitations. First, in its basic version ignores some possible information on network status; the server selection is optimized only by using SP knowledge (static and/or dynamic) about CSs status and then some client/user information (available locally or learned by the client by probing several CSs). Also, during in-session adaptation, each client (using DASH and/or server switching) tries to maximize, in a selfish way, its own QoE. Therefore no overall optimization is performed – from the network resources usage point of view. This work proposes to solve such limitations, in the context of 4G and 5G.

The single server-single client DASH performance can be improved as in [9], by using multiple-server DASH (MD-DASH), with better features w.r.t. bandwidth, link diversity and reliability. In [9], an innovative lightweight streaming solution is introduced, by taking advantage of bandwidth aggregation over multiple paths using simultaneously multiple content sources. This evolving approach outperforms the QoE delivered by current DASH-based or P2P-based solutions. Results in [9] show advantages in terms of quality delivered at the End-User's side and buffer occupancy. In addition, splitting content into multiple independent sub-streams provides the opportunity to implement easy-to-design content- and server-adaptation mechanisms. The MD-DASH is adopted in the system proposed in this paper.

A related problem, in multiple-server systems, is servers' location. The work [5] analyses the performance of several caching solutions for 4G, 5G networks. Fig. 1 shows (based on [5]) four possible levels of caching (i.e., a hierarchy) in a generic cellular network: in Internet, in Mobile Operator Network (MON) core, in Base stations (BS) of the RAN, or even in user terminals. The last case is advantageous if advanced Device–to-Device (D2D) direct communications are available.

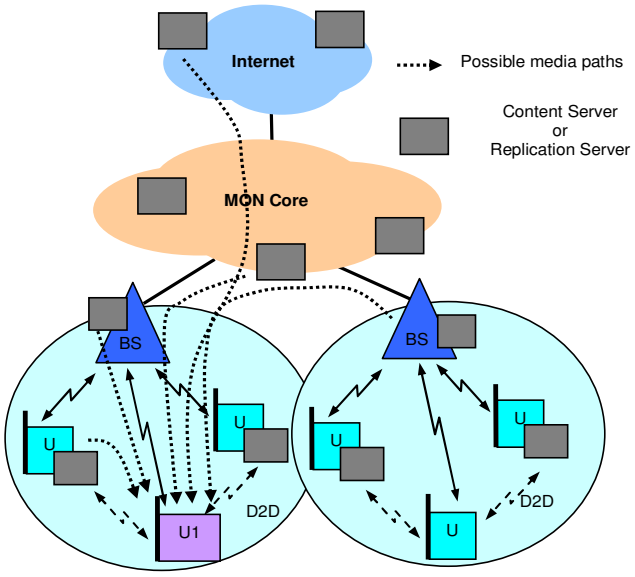

Figure 1. Hierarchical caching levels - possible in a mobile cellular network

MON – Mobile Operator Network; BS- Base Station; D2D – Device to Device communication; U- generic Mobile User Terminal/Equipment; U1- Consumer User instance.

Note that placing caching servers in proximity of potential users (i.e., in RAN or even user terminals) can be very valuable in 5G environments, in order to meet the very low E2E latency requirement (order of miliseconds) [1][2].

The article [13] optimizes HTTP-based multimedia delivery in multi-user mobile networks by combining the client-driven dynamic adaptation scheme DASH-3GPP with network-assisted adaptation capabilities. The adaptive HTTP streaming with multi-layer encoding (scalable video coding – SVC) allows efficient media delivery in multi-user scenarios. Additionally, the proposal takes benefit from mobile edge computing (MEC) deployed in RANs, close to

the users, in order to provide network assistance in the optimization process. A novel element- mobile edge-DASH adaptation function (ME-DAF) is introduced, which combines SVC-DASH-MEC to support efficient media delivery in mobile multi-user scenarios. The ME-DAF is inserted in the Data Plane managing effectively the DASH requests and media flows for multiple users. Our approach is different, in the sense that we also use MEC capabilities to provide network assistance, but he DASH sessions for multiple users are not concentrated in a single element, thus we avoid some scalability problems.

III. THE CLOUD RAN AND MOBILE EDGE COMPUTING

The emergent 5G will bring novel network and service capabilities [1][2]. It will ensure user experience continuity in various contexts like high mobility (e.g., in trains), dense or sparsely populated areas, or heterogeneous technologies. The target application range is broad: manufacturing, automotive, energy, food and agriculture, education, city management, government, healthcare, public transportation, and so forth.

The 5G has very ambitious goals and raises challenges [1][2], in terms of data volume, number of connected devices, latency, energy consumption, flexibility, etc. The 5G will be fully driven by software: a unified operating system is needed, in a number of points of presence, especially at the network edge. To achieve the required performance, scalability and agility the 5G can rely on technologies like Software Defined Networking, (SDN) Network Function Virtualization (NFV), Mobile Edge Computing (MEC) and Fog Computing (FC).

Recently, C-RAN architecture has been proposed [14-18], applicable both in 4G or 5G, able to provide among others, high spectral and energy efficiency. In C-RAN, the traditional base station (BS) is split into two parts: baseband units (BBUs) clustered as a BBU pool in a centralized location and several distributed remote radio heads (RRHs) plus antennas, which are located at the remote site. A high bandwidth low-latency optical or microwave transport network connect the RRHs and BBU pool (the connection is realized in hub-style from several RRUs to a single BBU). The RRHs perform radio frequency functions and support high capacity in hot spots. The BBU pool is virtualized and performs several functions as large-scale collaborative processing (LSCP), cooperative radio resource allocation (CRRA), and intelligent networking. The BBU pool communicates with RRHs via common public radio interface (CPRI) protocol, which supports a constant bit rate and bidirectional digitized in-phase and quadrature (I/Q) transmission, and includes specifications for control plane and data plane.

Several functional splits between BBU and RRH in 4G and 5G C-RANs are possible [19]. Shifting more functionality to the RRH can decrease the capacity requirement and increase delay requirement on the fronthaul links, but complicate and increase the cost of RRHs. If we consider the functional stack layers defined already in 4G, as Radio Frequency (RF), Physical Processing (PHY), Medium Access Control (MAC), Radio Link Control (RLC), Packet Data Convergence Protocol (PDCP), one might have for RRH functions: 4G - RF and 5G: RF,PHY, or RF,PHY,MAC, or RF,PHY,MAC,RLC. In general, the fronthaul network (between BBU and RRH) constraints have high impact on worsening C-RAN performance; the scale size of RRHs accessing the same BBU pool is limited and could not be too large due to the implementation complexity.

Each variant of the C-RAN architecture has some advantages and limitations. A „highly centralized" C-RAN, is easily upgradable and allows network capacity expansion; it can support multi-standard operation, maximum resource sharing and multi-cell collaborative signal processing. However, it has high bandwidth requirement between the BBU and RRHs. A „partial centralized" C-RAN requires much lower transmission bandwidth between BBU and RRH, by integrating some baseband processing into RRH.

C-RAN allows to operators to save costs and use green and efficient infrastructures. Open interfaces offers support for algorithms customization. The RAN virtualization solution can allow: HW/SW decoupling, multivendor I/O, flexible deployments, etc.

However, 5G new strong requirements and services (especially in terms of latency, energy efficiency, etc.) are difficult to be met by C-RAN only. Here, Mobile Edge Computing (MEC) can help.

MEC is a recent network architecture developed by the European Telecommunications Standards Institute (ETSI), [20] enabling distributed cloud computing capabilities and an IT service environment at network edge. By running applications and performing related processing tasks closer to the customers, network congestion is reduced and applications perform better. MEC can be implemented at the BSs, and enables flexible and rapid deployment of new applications and services (middleware, infrastructure) for customers. So, the operators can open their RAN to authorized third-parties, such as application developers and content providers. Location services, Internet-of-Things (IoT), video analytics, augmented reality, local content distribution, and data caching are some of the use cases identified by MEC.

The main element is the MEC application server (it can be integrated in RAN), which provides computing resources, storage capacity, connectivity and, if necessary, access to RAN information. It supports a multi-tenancy run-time and hosting environment for applications. The applications can be constructed as virtual appliances and packaged as operating system virtual machine (VM) images. They can be provided by equipment vendors, service providers and third-parties. The MEC application server can be deployed at the macro base station eNodeB LTE/4G or at the Radio Network Controller (RNC) in 3G networks. It can also collect data about storage, network bandwidth, CPU utilization, etc., for each application or service deployed by a third party. Therefore application developers and content providers can take advantage of close proximity to cellular subscribers and real-time RAN information.

The MEC "edge" approach can cooperate with C-RAN architecture; MEC will add flexible decentralization and proper dynamic instantiation and orchestration of virtual machines serving for network management in close proximity to terminals. In a heterogeneous C-RAN environment the MEC server can be deployed either at BBU pool or in eNodeBs.

IV. VIDEO CONTENT DELIVERY SOLUTIONS IN HETROGENEOUS C-RAN

C-RAN technology can efficiently support video content delivery, especially when intelligent cooperative caching is applied [5][19]. The powerful C-RAN BBU can control all radio access technologies (RAT), and possibly facilitate the video encoding and transmission towards user over different RATs. Hierarchical cooperative caching framework in C-RAN is proposed in [19] with contents jointly cached at the BBU and at the RRHs. However, the fronthaul C-RAN constraints have high impact on lowering CRAN performance and the scale size of RRHs; accessing the same BBU pool is limited and could not be too large in terms of RRHs number, due to the implementation complexity. On the other side heterogeneity is a frequent characteristic to be considered in integrating today various RATs.

The Heterogenous CRANs (H-CRAN), [22] takes into account the heterogeneous networks (HetNets). The RAN components are *Low Power Nodes (LPN)* (e.g., pico BS, femto BS, small BS, etc.) aiming to increase capacity in dense areas with high traffic demand and *High Power Nodes* (HPN - e.g., macro or micro BS) that can be combined with LPN to form a HetNet.

The H-CRAN architecture can include a central entity, which is the extended (eBBU pool), containing baseband processing units (the architectural layers are L1-baseband, MAC and Network). The BBU pool is linked via Gateway to the external Internet. Several peripheral "islands" realized with different technologies are linked to the BBU pool in hub – style, via two types of links: backhaul (BBU – HPNs), or fronthaul links (BBU pool – LPN). Several configurations can exist like: 2G/3G/LTE islands containing Base station Controllers (for 2G/3G), Macro Base Stations (MBS) seen as HPNs and LPNs, i.e., RRHs (the latter can be linked directly to the BBU pool via fronthaul links); 5G MBSs (as HPNs) and RRHs; WiMAX BS (HPN) and RRHs; IEEE 802.11 HPN Access Point (AP) and RRHs. Each peripheral island can be seen as an alternative path connected to Internet via Gateways.

The H-CRAN can support efficiently video and media delivery services [22]. Recall that in conventional delivery solutions the video packet encoding and scheduling is done at head-end station (HS). Data will flow on predetermined paths (via assigned RATs) to mobile user equipments (MUE). However, the path from HS to MUE has a long delay for the feedback represented by the Network State Information (NSI); so, only certain quasi-static info is accessible to the HS and this determines low performance for adaptive flow control and video encoding techniques. Therefore bringing content sources closer to end user by caching could significantly improve the performance of adaptive systems.

Three techniques are proposed by this paper to be combined, to improve the video content delivery in H-CRAN: (a)distributed caching, (b)multi-server DASH-based delivery and (c)MEC approach, to achieve optimization of RAN resource allocation and QoE improvement at end user level.

Caching variants can be used in H-CRAN. When no eBBU Pool caching is applied, then the eBBU pool is directly connected to the RATs. However, it still can improve the delivery because it can easily obtain their online NSI and may utilize it in the packet scheduling (multi-RAT scheduler). The priorities of different video packets (e.g., those generated by Scalable Video Coding - SVC) or QoS requirements from multiple MUEs may also affect the scheduling at the eBBU pool. The H-CRAN with packet scheduling exposes better delivery performance than conventional heterogeneous networks with only HS scheduling.

The video can be also cached at the local eBBU pool, based on the technology of content awareness caching for 5G networks, thus reducing the traffic amount from original HS. More, both the video encoding and transmission can be adapted to the online NSI of multiple RATs. The eBBU pool can even work as a Service Provider (SP) with the units encoding the source video, controlling the frame rate, and managing the pre-caching content and buffering in MUEs. More accurate online NSI can determine the encoding redundancy and the size of pre-caching content could be minimized, thus saving the scarce spectrum resource. More accurate NSI at the eBBU pool may lead to decisions to reduce encoding redundancy and therefore increase the efficiency. Caching (replica servers) can be also placed in HPNs, eNodeBs, and even in RRHs or MUE if sufficient storage resources are available [5] [19].

A multiple-source adaptive streaming (MS-stream) solution is proposed in [9], targeting to enhance the consumer's perceived quality. Compared with traditional single-server approach this solution can to better exploit expanded bandwidth, link diversity, and reliability. It is codec agnostic, DASH compliant, and receiver-driven, thus being a pragmatic and evolving solution for QoE enhancement. The content is split into multiple independent sub-streams providing the opportunity to achieve easy-to-design bitrate adaptation and server-switching mechanisms. This approach can be used also in H-CRAN environment, and we considers such a solution, where several caching entities are distributed over the BBU pool or in the RANs (see Fig. 1), or even in MUEs.

Fig. 2 shows a high level view of the architecture proposed in this paper; it introduces MS-stream approach to 5G H-CRAN environment, while additionally taking benefit from MEC support to achieve global optimization of server-path resources. Different islands having heterogeneous RATs are connected in hub-style to the eBBU pool. At its turn the eBBU pool is connected to the mobile core network and through this to the general internet. Several caching nodes can be placed in different places following different policies

of the Service Provider and other criteria (popularity, time-life, cost, etc.). MEC servers are supposed to be installed close to each HPNs of the H-CRAN. The specific Control Plane of our system is composed mainly by the Service Provider (SP) entity placed at eBBU pool level and several functional blocks called RAN Monitors (RAN-Mon), which are installed as application instances over MEC servers. The SP gets the video content requests from the user terminals and optimizes server utilization. The RAN-Mon block role is similar as in [13], i.e., it interacts with MEC server in order to collect RAN statistics (NSI, i.e., cell load information, channel state information provided by channel state indicator, etc.).

Fig. 2 considers a variant where the MEC server is collocated with a Macro Base Station (MBS). DASH clients can run in mobile terminals. The SP communicates in the Control Plane with RAN-Mon and is aware of network resources status; such information is usually available at RAN level in 4G or 5G.

An user client content object request is addressed (similar to DISEDAN system discussed in Section II) to SP entity. Based on user request the SP performs a selection phase (based on multiple criteria algorithm) of a set of servers containing DASH descriptions (see [9] for details) of the required media object media. Then the user (after making the final filtering/selection and performing a local-information based selection) starts a set of parallel Data Plane dialogues with the caching servers selected. During the sessions, individual adjustments of the flow rates can be applied by using DASH algorithms and/or changing the current server set (server switching). Also in an action of selecting an updated set of servers, the multiple criteria optimization algorithm can be applied. The main difference from DISEDAN system and also from approach presented in [13] is the fact that not only individual, but overall optimization can be achieved while taking benefit from RAN information.

V. MULTI-SERVER SELECTION OPTIMIZATION FOR H-CRAN

This section is devoted to propose a solution for multiple server set selection to serve a given user request, coming form an mobile end user terminal, to the Service Provider. It will be supposed that SP has enough knowledge about caching servers placed in a given region, and also about distance and channel status between a given server and mobile terminal of the requesting user. This paper will not detail the signaling messages between the SP and different MEC servers placed in RAN.

Several multi-objective optimization algorithms can be considered. In this work, the optimization is based on a previously used procedure - Multi-Criteria Decision Algorithms (MCDA) - which has been proved powerful and efficient in [23][24]. Note the important fact that the method proposed has no limitation in number of parameters to be considered as input. The multi-criteria algorithm can use more or less parameters as they are available in the system.

The multi-objective optimization algorithm tries to find $min\ F(x) = [f_1(x), ..f_k(x)]$ where $x \in X^i$, the decision variables space, and $f_1(x), ..f_k(x)$, are a set of objectives, [23] [24].

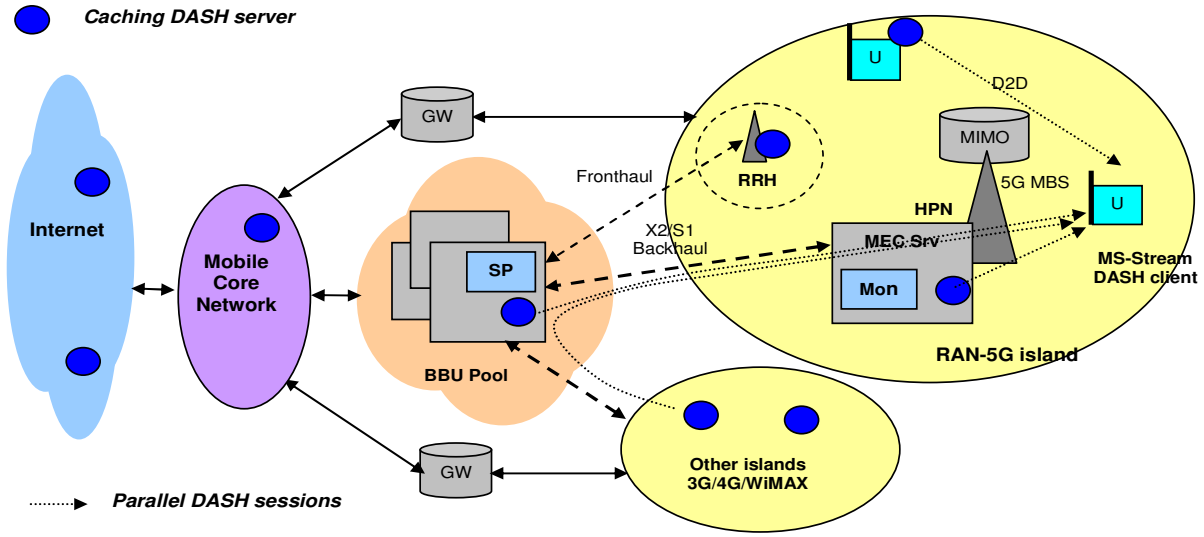

Figure 2. Architecture based on H-CRAN for multiple source streaming and MEC support (variant: MEC implemented at MBS); GW- gateway; RRH- Remote Radio Head; MBS- Macro Base Station; D2D- Device to Device; HPN- High Power Node; MS- multiserver; BBU – Baseband Unit; MEC Mobile Edge Computing; X2/S1 – Interfaces imported from 4G technology; U- generic notation for an user having a mobile terminal

One method to solve MCDA problem is offered by *reference level decision algorithm* [24], which considers a decision space R^m and the decision parameter/variables: v_i, i=1, ..m; $\forall i$, $v_i \geq 0$. A candidate solution is an element $S_s=(v_{s1}, v_{s2}, .., v_{sm}) \in R^m$. Let S be the number of candidates indexed by s = 1, 2, ..S. The value ranges of decision variables might be bounded by given constrains.

The algorithm searches a solution satisfying a given objective function, conforming a particular metric. Two reference parameters are defined: r_i =*reservation level*=the upper limit for a decision variable which should not be crossed by the selected solution; a_i=*aspiration level*=the lower bound for a decision variable, beyond which the solutions are seen as similar. For each decision variable v_i, r_i *and* a_i, will be computed among all solutions s = 1, 2, ..S: $r_i = \max [v_{is}]$, $a_i = \min [v_{is}]$, where s = 1, 2, ..S.

Two modifications of the decision variables are applied in [24]: a. *replacement of each variable* with *distance from its value to the reservation level*: $v_i \rightarrow r_i - v_i$; (higher v_i will decrease the distance); b. *normalization* is also introduced to get non-dimensional values, which can be numerically compared. For each variable v_{si}, a ratio is computed, for each solution *s,* and each variable i: $v_{si}' = (r_i - v_{si})/(r_i - a_i)$, where the factor $1/(r_i - a_i)$ - plays also the role of a weight. To support a variety of SP policies, a modified formula can be used, i.e.:

$$v_{si}' = w_i(r_i - v_{si})/(r_i - a_i) \quad (1)$$

where the factor $w_i \in (0,1]$ represents a weight (associated to a priority) that can be established from SP policy considerations. Such weights can significantly influence the final selection. The optimization algorithm presented below is derived from that applied in [7]:

1. Compute the matrix $M\{v_{si}'\}$, s=1…S, i=1…m
2. Compute for each candidate solution *s*, the minimum (worst case) among all its normalized variables v_{si}':

$$min_s = min\{v_{si}'\}; \; i=1...m \quad (2)$$

3. Make selection among solutions by computing:

$$v_{opt} = max\{min_s\}, \; s=1, ..S \quad (3)$$

This v_{opt} is the optimum solution, i.e it selects the best value among those produced by the Step 1.
4. Repeat the algorithm for the servers left, until a desired set of "best" servers is obtained, or the list is exhausted. (this step is necessary to determine the set of active servers for MS-stream).

The performance of such optimization algorithm has been already proven in [6][7]. In the context of H-CRAN its efficiency depends finally on the accuracy of the network parameters delivered by MEC server to SP.

A simplified example shows the optimization procedure. One supposes that decision variables are those defined in Table 1. The variable *v1* is estimated directly by the SP, by inspecting the servers. The other variables are provided by the MEC server to SP. Table 2 presents six candidates solutions (entries are native not-yet normalized values). Priority examples are introduced in Table 1, derived from SP policy. Here, the server load and numbers of RAN cells crossed are considered the most important.

In this example one can define: a1= 0, r1=100; a2=0, r2=10; a3=120, r3 = 20; a4=0, r4=100; a5=0, r5=30.

TABLE I. DECISION VARIABLES EXAMPLE

Decision variables	**Semantics**	**Units**	**Priority**
v1	Load of the caching server	(%)	1- max
v2	Number of RAN cells or sub-networks to be crossed	Integer	2
v3	Average capacity available on the channel server- client	Mbps	2
v4	Load of the cell of the server	(%)	3
v5	Estimated server-client delay	ms	4- min

TABLE II. CANDIDATE SOLUTIONS EXAMPLE

	s_1	s_2	s_3	s_4	s_5	s_6
v_{s1}	0	20	40	70	80	50
v_{s2}	2	3	1	3	4	5
v_{s3}	60	30	50	80	50	60
v_{s4}	30	10	20	60	20	30
v_{s5}	15	20	10	10	20	5

Applying the basic algorithm (i.e., with no priorities) simple computation will show that formula (4) is *max{0.5, 0.3, 0.5, 0.3, 0.2, 0.5},* showing that solutions s1, s3, s6 are equivalent. Suppose we want *n* servers for MS-stream delivery. Then the step 4 of the algorithm simply means to select the first *n* servers of the list, considering the order given by the step 3 of the algorithm; if n=3, they are { s1, s3, s6}.

If some decision variables are considered more important in the selection process, then introduce policies, can be defined. An example of priorities assigned is given in the last column of the Table 1. To these priorities the SP can associate weights (acting as compression factors) defined, e.g., w_1= *0.5,* w_2= *0.7,* w_3= *0.7,* w_4= *0.8,* w_5= *1.0.* Then the step 3 of the algorithm will produce the *{0.5, 0.3, 0.3, 0.15, 0.1, 0.25}.* It is seen that s1 solution is the best, followed by s2 and s3.

VI. CONCLUSIONS, EXTENSIONS AND FUTURE WORK

This paper proposed an architectural solution for optimizing video content delivery in 5G Heterogeneous Cloud RAN environment. A previously developed multi-server video streaming system, based on DASH adaptation subsystem has been taken and combined here with Mobile Edge Computing (MEC) capabilities, in order to optimize the resource usage in RAN and enhance the quality of experience (QoE) seen by the end users.

Specific work developed here is on the initial best path-server selection, producing a subset of servers (which will serve the DASH sessions of the users). While the efficiency of Multi-criteria decision algorithms has been already proven in such types of problems, the contribution here is the

extension of such an approach to MS-stream + MEC cooperation in H-CRAN environment. Due to network related information, both QoE increase and global optimization of RAN resource usage is expected.

Future work will be done to evaluate the system performance in a large network environment, and extension of algorithm applicability during the DASH sessions, when problems appear to switch the set of caching servers. More in depth study should be also done to embed the RAN Monitoring subsystem in mobile edge computing environment.

REFERENCES

[1] J.G. Andrews, .et al., "What Will 5G Be?", IEEE Journal on Selected Areas in Communications, Vol. 32, No.6, pp. 1065-82, June 2014.

[2] M. Peng, Y. Li, Z. Zhao, and C. Wang, "System Architecture and Key Technologies for 5G Heterogeneous Cloud Radio Access Networks", *IEEE Network Magazine*, vol. 29, no. 2, pp. 6–14, Mar. 2015.

[3] "Cisco Visual Networking Index: Forecast and Methodology 2013–2018", White Paper, June 2014.

[4] P. A. Khan and B. Rajkumar, "A Taxonomy and Survey of Content Delivery Networks", Department of Computer Science and Software Engineering, University of Melbourne. Australia : s.n., 2008, www.cloudbus.org/reports/CDN-Taxonomy.pdf, [retrieved: Dec., 2015].

[5] X.Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. M.Leung, "Cache in the Air: Exploiting Content Caching and Delivery Techniques for 5G Systems", IEEE Communications Magazine, pp.131-139, February 2014.

[6] http://wp2.tele.pw.edu.pl/disedan/ [retrieved: May, 2016]

[7] E. Borcoci, M. Vochin, M. Constantinescu, J. M. Batalla, and D. Negru, "On Server and Path Selection Algorithms and Policies in a light Content-Aware Networking Architecture", ICSNC 2014, http://www.iaria.org/conferences2014/ICSNC14.html [retrieved: July, 2016].

[8] A. Bęben, J. Mongay Batalla, P. Wiśniewski, and P. Krawiec (WUT), "ABMA+ : lightweight and efficient algorithm for HTTP adaptive streaming", ACM Multimedia Systems (MMSys), Klagenfurt (Austria), May 2016, [doi: http://dx.doi.org/10.1145/2910017.2910596]

[9] J. Bruneau-Queyreix, D. Négru, J. M. Batalla, and E. Borcoci, "Multiple Description-DASH: Pragmatic video streaming maximizing End-Users' Quality of Experience" IEEE International Conference on Communications, 23-27 May 2016, Kuala Lumpur, Malaysia, http://icc2016.ieee-icc.org/content/symposia, [retrieved: July, 2016].

[10] I. Sodagar, "The MPEG-DASH Standard for Multimedia Streaming Over the Internet," MultiMedia, IEEE, vol. 18, no. 4, pp. 62 – 67, 2011,.

[11] ISO/IEC 23009-1, "Information technology -- Dynamic adaptive streaming over HTTP (DASH) -- Part 1: Media presentation description and segment formats," ISO/IEC, Geneve, second edition, 2014.

[12] J. Choi, J. Han, E. Cho, T. Kwon, and Y. Choi, "A Survey on Content-Oriented Networking for Efficient Content Delivery", IEEE Communications Magazine, pp. 121-127, March 2011.

[13] J. O. Fajardo, I. Taboada, and F. Liberal, "Improving Content Delivery Efficiency Through Multi-Layer Mobile Edge Adaptation", IEEE Network Magazine, pp.40-46, November/December 2015.

[14] P. K. Agyapong, M. Iwamura, D. Staehle, W. Kiess and A. Benjebbour, „Design Considerations for a 5G Network Architecture" IEEE Communications Magazine, pp. 65-75, November 2014.

[15] M. Peng, Y. Sun, X. Li, Z. Mao, and C. Wang, "Recent Advances in Cloud Radio Access Networks:System Architectures, Key Techniques, and Open Issues" IEEE Communications Surveys and Tutorials, pp. 1 - 27, 2016, http://arxiv.org/abs/1604.00607, [retrieved: July, 2016].

[16] N. Panwar, S. Sharma, and A. K. Singh, "A Survey on 5G: The Next Generation of Mobile Communication" , accepted in Elsevier Physical Communication, pp.64-84, 4 Nov 2015, http://arxiv.org/pdf/1511.01643v1.pdf, [retrieved: July, 2016].

[17] A. Checko et al., "Cloud RAN for Mobile Networks—A Technology Overview", IEEE Communications Surveys & Tutorials, Vol. 17, No. 1, pp. 405-426, First Quarter 2015.

[18] China Mobile Research Institute, "C-RAN White Paper: The Road Towards Green RAN", June 2014, http://labs.chinamobile.com/cran/wp-content/uploads/2014/06/20140613-C-RAN-WP-3.0.pdf, [retrieved: July, 2016].

[19] T.X. Tran, A. Hajisami, and D.Pompili, "Cooperative Hierarchical Caching in 5G Cloud Radio Access Networks (C-RANs)", https://arxiv.org/pdf/1602.02178, [retrieved: October, 2015].

[20] M. Patel *et al.*, "Mobile-Edge Computing Introductory Technical White, Paper," 2014, https://portal.etsi.org/portals/0/tbpages/mec/docs/mobile-edge_computing_-introductory_technical_white_paper_v1%2018-09-14.pdf, [retrieved: October, 2015].

[21] H. S. Matharu, "Cloud RAN and Mobile Edge Computing, a dichotomy in the making", http://www.microwave-eetimes.com/content/cloud-ran-and-mobile-edge-computing-dichotomy-making, 2016, [retrieved: July, 2016].

[22] M. Sheng, W. Han, C. Huang, and S. Cui, 'Video Delivery in Heterogenous Crans: Architectures and Strategies', IEEE Wireless Communications, pp.14-21, June 2015.

[23] J. Figueira, S. Greco, and M. Ehrgott, "Multiple Criteria Decision Analysis: State of the Art Surveys", Kluwer Academic Publishers, 2005.

[24] A. P. Wierzbicki, "The use of reference objectives in multiobjective optimization". Lecture Notes in Economics and Mathematical Systems, vol. 177., Springer-Verlag, pp. 468–486.