Shared Cache as a Service in Future Converged Fixed and Mobile Network

Zhe Li, Yaning Liu,
Jean-Charles Point
JCP-Connect, Rennes, France
{firstname.lastname}@jcp-connect.com
pointjc@jcp-connect.com

Selami Ciftci, Onur Eker
Argela, Ankara, Turkey
{Firstname.Lastname}@argela.com.tr

Marco Savi, Massimo Tornatore,
Giacomo Verticale
Politecnico di Milano, Italy
{firstname.lastname}@polimi.it

***Abstract*—Thanks to the widespread deployment of LTE and WiFi networks, connectivity at anytime, anywhere on any device is no longer a dream. However, we can hardly say that today the QoS/QoE provided by the current network is satisfactory. This is mainly caused by the separated management of the different infrastructures, which leads to sub-optimal usage of network resources. This work presents a Shared Caching System (SCS) in the context of Fixed Mobile Converged (FMC) networks that allows network-infrastructure operators to offer Caching as a Service (CaaS) to OTT service/content providers or virtual-network operators. We highlight the benefits of deploying such a holistically-managed in-network caching over different type of network infrastructures.**

I. INTRODUCTION

Internet traffic keeps growing at pace as a consequence of the steadily increasing number of users and the adoption of new bandwidth-intensive services (such as video services) by these users. According to recent studies [1], global IP traffic has increased more than fivefold in the past 5 years, and will increase threefold over the next 5 years. Moreover, due to the increasing popularity of smart phones and emerging mobile applications, mobile Internet is dramatically expanding. The Cisco Visual Networking Index [2] shows that Internet traffic from wireless and mobile devices will exceed traffic from wired devices by 2016 and nearly half of Internet traffic will originate from non-PC devices by then. Global mobile traffic will increase nearly 11-fold in 2018.

To meet the growing bandwidth demands of mobile users, commercial LTE networks are being deployed to significantly increase the bandwidth and reduce the latency experienced in the mobile network. Though LTE improves the network performance between end users and mobile network, the bandwidth is still not enough to deal with the exploding traffic volume. Consequently, various technologies such as caching in the air [3] and traffic offloading [4] have emerged as potential solutions. Despite the substantial work in these aspects, the proposed solutions are still sub-optimal due to the separated management of different types of networks, and thus yields an inefficient resource utilization. For example, in the traffic offloading solution, a vertical handover has to be performed to switch the user equipment (UE) from one type of network to another. But a seamless handover is not always possible, especially from mobile to fixed network, as the two network architectures are typically separated at both structural and functional level. Therefore, a Fixed Mobile Converged network architecture will be essential for the future 5G network.

In this work, we take a step ahead and we assume the FMC network is already available. In this context, virtualization of converged content caches and network resources will enable new added-value services and pricing schemes for FMC network operators. Leveraging virtualization, the infrastructure and the content in the same network can be managed separately, as the network-management layer can be controlled by the FMC network operator while the content-management layer is controlled by either the network operator itself or the content provider. The Shared Caching System (SCS) allows network operators to provide novel content-based bundle services to content or service providers, namely to offer Caching as a Service (CaaS). On the other hand, leveraging the sophisticated management of caching, quality-based differential pricing is also seen as a promising solution for operators to increase their revenue.

II. SHARED CACHING ARCHITECTURE

Current fixed and mobile networks are separate. The fixed network is composed of three segments: the access network, the metropolitan/regional aggregation network, and the core network. The mobile network is composed of 2 segments, which are different from the fixed network segments: the RAN (composed itself of the mobile access network and the mobile aggregation network) and the mobile backbone network.

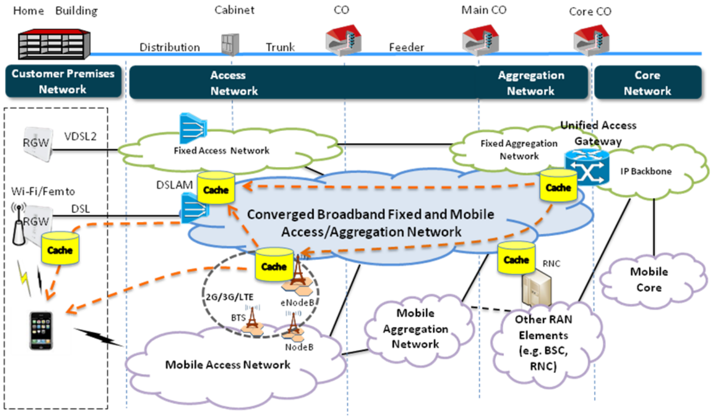

Fig. 1. FMC Framework

A vision of the unified access and aggregation network architecture enabling FMC is shown in Fig. 1. The key enabler of this FMC architecture is the Unified Access Gateway (UAG). It is a network element which logically supports service creation for fixed and mobile users. Central to this concept is the ability to support fronthaul architectures such as Cloud-RAN (Cloud-Radio Access Network). The Cloud-RAN model implies that the Baseband Unit (BBU) function of the base station is centralized in a server pool hosted at the UAG. As such, the UAG is the first location of confluence for fronthaul mobile services and additionally for the convergence of mobile and fixed line services. The UAG also supports other functions such as very tight coupling of LTE and WiFi networks [4] and unified authentication, authorization, accounting so as to guarantee the seamless switch between fixed and mobile networks. In other words, it ensures the convergence of fixed and mobile network.

The UAG plays the pivotal point in the FMC operator network and it will serve as the main entity providing interactions between FMC operator and other operators and/or service providers. Its control plane is composed of both network- and content-management layers. While the network-management layer deals with traffic-related functions as topology management and routing, the content management layer handles content-related functions such as cache and metadata management. In this respect, the UAG in a FMC operator network provides unified APIs/interfaces to mobile and fixed network operators as well as content providers.

The UAG acts as the universal coordinated convergence controller in FMC networks. In this respect, the UAG is the main interaction entity in the network between FMC network operator and other stakeholders. Therefore, the UAG is also the heart of the caching framework in FMC networks, as it does not only store the cached content, but it is also responsible for the delivery to end users of content cached at any FMC nodes. For example, in a vertical 4G-to-Wi-Fi handover scenario, the content cached at UAG, Wi-Fi access points and eNodeBs, can be managed/delivered via intelligent pre-fetching or any other caching techniques without the need for any content transmission in the core network. According to this example, the content-management layer takes decisions on where to cache the content using its cache and metadata management modules, while the network-management layer deals with transmission of the content from the caching location to eNodeB or Wi-Fi access point using its topology and routing management modules, transparently to the end users.

To realize content management, a cache controller is integrated in the UAG. The cache controller receives inputs from both content providers and network operators, and gives as outputs optimal caching and real-time pre-fetching plans. Directives will be sent to the distributed cache nodes to execute these plans. We leverage the solution proposed in [5] and extend it to the shared FMC infrastructure to build the cache controller. As it is shown in Fig. 2, we deploy shared cache nodes on various network equipments to form a hierarchical caching system. The cache controller keeps the mapping of content and its location (*i.e.,* in which cache node the content is stored). The OTT service/content provider sends its requirements to UAG or FMC network operator, and then these requirements, together with the current network conditions, are forwarded to the cache controller. Upon receiving the information, the cache controller updates the cache plan according to the information given by the FMC network operator. Finally, the caching plan is transformed into forwarding rules and caching (pre-fetching) directives and sent to UAG and cache nodes respectively to deploy the new cache setting. The UAG is responsible for setting the appropriate forwarding rules in the OpenFlow switches so that the user requests will be forwarded to the right cache nodes.

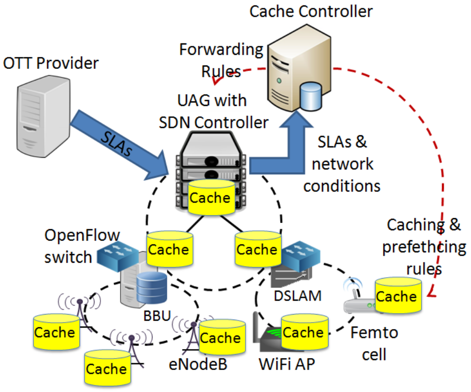

Fig. 2. Shared Caching Architecture

III. PERFORMANCE EVALUATION

Our objective here is to show how the FMC operator can leverage SCS to improve the QoS/QoE at user side via NS3 simulator. Therefore, we assume that the operator has a full control of the network, and it is able to obtain all the traces of user's motion information, *i.e.,* the operator knows when a user will change its WiFi interface to LTE interface (and vice versa), and which access node it will attach to. On the contrary, since current mobile and WiFi networks are managed separately, the mobile operator and the WiFi operator do not have the same knowledge as the FMC operator, so they cannot detect the interface switching action of a UE. On the other side, when the UE is covered by both the WiFi and LTE, it will select the WiFi network as preference. Moreover, seamless handover between different interfaces in the FMC network is assured, while in the current network a change of interface will always introduce a cut of session (*i.e.,* interruption of video playback).

The test scenario is related to the example described in section II. We used BonnMotion [6] to generate the random street map and the random walk model for 100 active mobile users. We assumed that the mobile users consume short movies with the sizes ranged from 5 MB (a 5 minutes, 200 kbps video) to 60 MB (a 20 minutes, 400 kbps video). Every video is divided into several clips of 5 seconds each. Overall, there are $10,000$ videos in the content pool whose popularity follows a Zipf distribution with $\alpha = 0.8$. When the UE moves, it will switch between LTE and WiFi network according to the coverage of WiFi APs. At the same time, the FMC network operator monitors the user behavior. Once it detects the possible switch from one eNB or AP to another, the SCS will pre-fetch the content before the switch happens so that when the user connects to the new access point, the content is already accessible in the local cache. We use the simple Least Recently Used (LRU) policy for caching replacement.

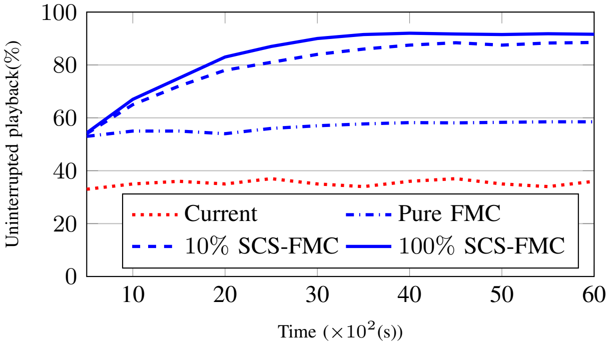

Fig. 3. Playback continuity

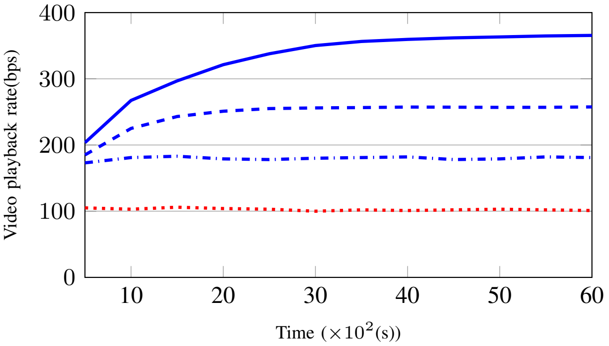

Fig. 4. Playback rate (legends are the same in Fig. 3)

We deployed 8 APs and 2 eNBs over a geographical space around $2km^2$. The maximum capacity of each AP is 72 Mbps, and the AP serves around 10 users simultaneously in average. Therefore the link capacity between APs and end users is fast enough so that, if a video is cached in the AP, the user can obtain it immediately. However, we assumed that the downlink that connects these APs to the rest of the Internet is the bottleneck. Its capacity is limited by 10 Mbps so that the downlink cannot afford the video traffic when the 100 users consume video content simultaneously via WiFi network. This may produce some interruptions of video playback even if a user is in the coverage of an AP.

Two eNBs are deployed in the network such that they can provide a full coverage of the territory. However, they will be connected by UE only when WiFi network is not available. The average mobile link capacity is set to 1 Mbps and the backhaul bandwidth is 10 Mbps. Therefore, the LTE network cannot afford 100 simultaneous connection either. Cache storage is deployed on eNBs and WiFi APs. The total size of the equipped cache varies from from 10% of the size of the total video data (around 80 GB) to the whole data set (around 800 GB). The duration of our simulation was one hour and a half, and there is no video stored in SCS at the beginning of the simulation.

Fig. 3 displays the video playback continuity in different systems. Please note that in the *current network* and *Pure FMC* setting, there is neither cache nor pre-fetching function available. As we can see from the figure, a pure FMC network can already improve the percentage of the video playback continuity up to 55%, which means 55% of the videos watched by the users in the network can be finished without any interruption. Conversely in the current network the percentage of continuity is only around 30%. In this pure FMC network setting, there is no intelligent traffic offloading between LTE and WiFi network. Therefore, if most of the users are connected to the WiFi network, there will be a congestion in the WiFi downlink which leads to playback interruption. Intelligent traffic offloading mechanisms can be deployed to further improve the performance of the FMC network.

The key benefit shown in Fig. 3 is the significant improvement of the video playback continuity gained by deploying SCS. Using only a SCS with small cache size (*i.e.,* 10% of the total content quantity), around 90% percent of the video playback session can be finished without any interruption. However, increasing the cache size of the SCS does not help much in terms of playback continuity.

The average video playback rate at user side is given in Fig. 4. While increasing the cache size of SCS does not impact much on the playback continuity, it does lead to a dramatic increase of video playback rate. The playback rate in 100% *SCS-FMC network* is tripled compared to the playback rate in the current network, which indicates significant improvement of QoE.

IV. CONCLUSION AND FUTURE WORK

It is commonly agreed that the fixed-mobile converged network will be the architecture of future 5G network. In-network caching will be an indispensable technology to deliver high service quality in the FMC network. In this work, we showed the great advantage of deploying such a Shared Caching System in terms of improving user QoS/QoE. The next step is to realize the real world implementation of the system and investigate the business models to delpoy SCS.

ACKNOWLEDGMENTS

The work leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement NO. 317762 "COMBO project".

REFERENCES

[1] Cisco, "Cisco visual networking index: Forecast and methodology, 2013-2018," June 2014, http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.html.

[2] ——, "Cisco visual networking index: Global mobile data traffic forecast update, 2013-2018," Feb 2014, http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html.

[3] 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*, vol. 52, 2014.

[4] X. Lagrange, "Very tight coupling between lte and wi-fi for advanced offloading procedures," in *Workshop on Interference and Design Issues for Future Heterogeneous Networks (IEEE WCNC)*, 2014.

[5] P. Georgopoulos, M. Broadbent, B. Plattner, and N. Race, "Cache as a service: Leveraging sdn to efficiently and transparently support video-on-demand on the last mile," in *IEEE International Conference on Computer Communications and Networks (ICCCN)*, 2014.

[6] BonnMotion, "A mobility scenario generation and analysis tool," http://sys.cs.uos.de/bonnmotion/.