WHY AND HOW EDGE CLOUD COMPUTING CAN ADDRESS PERFORMANCE AND ECONOMIC SUSTAINABILITY ISSUES FOR TELCO DOMESTIC NETWORKS

Gianfranco Ciccarella[1], Romeo Giuliano[2], Franco Mazzenga[3], Francesco Vatalaro[3], Alessandro Vizzarri[3]
[1]Telecommunications Consultant, L'Aquila (Italy) [2]Guglielmo Marconi University, Department of Engineering Science, Rome (Italy), [3]University of Rome Tor Vergata, Department of Enterprise Engineering "Mario Lucertini," Rome (Italy)

NOTE: Corresponding author: Alessandro Vizzarri, alessandro.vizzarri@uniroma2.it

Abstract *– This paper analyzes some of the telecommunication companies' (telcos) domestic networks' main objectives and issues related to the application services Quality of Experience (QoE), i.e., "technical QoE," and the economic sustainability of the Very High Capacity (VHC) networks. Telco domestic networks and Over The Top (OTT) networks are the Internet segments connecting the end-user equipment to the servers/clouds that provide application services. With the advent of VHC networks, telcos afford difficulties in managing effectively and efficiently the application services and their business sustainability. This paper affords some issues related to traditional telcos' domestic network architectures and the approach to managing the application services' performance improvement. Some telcos started changing their architectures, although a massive transformation had not yet started in the industry. Traditionally, telcos focus on network services, i.e., the transport of IP packets. Performance improvement is obtained by QoS-based traffic management techniques, such as bandwidth reservation and packet prioritization. To manage application performance improvement, the use of layer 4 techniques and Edge Cloud Computing (ECC) is effective, as demonstrated by the multiyear experience of OTTs which already use these technologies. In telco domestic networks, layer 2 tunnels increase the complexity of deploying ECC. However, some vendors provide non-standard solutions to make the IP layer 3 user plane visible and to deploy ECC. ECC is the key factor for the transformation. It is a mini/micro-data center that distributes some applications and content closer to the end users. The distribution can provide a paradigm shift in a telco's business. In the paper, we first highlight the need for regulators to appreciate the need to encourage this industry transformation fully. Then, we present solutions for the telco industry based on network architecture transformation and ECC at the telco's edge to improve technical QoE, reduce cost and possibly generate new revenue streams. We also provide analysis results on network cost saving with ECC.*

Keywords – 5G, application services KPIs, edge cloud computing, network cost, network KPIs, new telco business models, uRLLC

1. INTRODUCTION

The primary Internet ecosystem's objective is to provide application services to the end users with the Quality of Experience (QoE) required by the different services (e.g., e-commerce, video streaming, gaming, 360-degrees augmented/virtual reality). A rather broad definition of QoE, provided by Recommendation ITU-T P.10/G.100 (2017) Amd. 1 (06/2019) [1], consists of «*the degree of delight or annoyance of the user of an application or service.* » This definition encompasses both system-level (i.e., related to overall customer experience) QoE and the so-called "technical QoE," which defines the quantitative requirements and measurements for the application services performance. Technical QoE is one of the most significant factors in assessing customer experience. The technical QoE needed for the different services depends on the end-to-end connection between the end-user devices and the servers/clouds that offer the services. The increase of technical QoE depends on the source location and network's extent that content and applications must cross to reach the end user.

This paper considers only the primary application services technical QoE Key Performance Indicators (KPIs) as defined below (Section 2), i.e., those strictly related to the end-to-end quality of the service transport [2]. Other system-level KPIs, such as those related to service authorization and accounting, availability of the connection from the end-user device to the server/cloud, security/privacy, etc., are out of the scope of this paper, and we will not consider them.

Quality of Service (QoS), which differs from technical QoE, measures service parameters by averaging some main technical network KPIs. Traditionally, a telco is focused on QoS and sometimes tries predicting QoE based on a construct of QoS-QoE models. However, a telco's predicted QoE turns out to be a loose function of the set of network performance (averaged) KPIs. Then QoS techniques have limited or no effect on individual customer service experience.

Telco domestic networks and Over The Top (OTT) networks are the Internet segments connecting the end-user equipment to servers/clouds that provide the application services. OTTs can manage effectively and efficiently the application services technical QoE, and have economic sustainability. On the contrary, telcos have had severe difficulties with application service quality and economic sustainability since the advent of ultra-broadband services, especially video services [2-4].

Ever-growing ultra-broadband traffic demand has increased access network capacity. Regulators firmly push telcos to deploy expensive Very High Capacity (VHC) networks in their access networks. However, they are mainly focused on the bit rate increase in the communication channel and do not address the application services' performance end-to-end. This lack of consideration may hinder the application services' performance and the economic sustainability of the telco business. The main objectives for VHC networks should be:

a) to improve the end-user technical QoE, i.e., to provide the KPIs required by the ultra-broadband services [5-7];
b) to define solutions that satisfy the economic sustainability objectives for telcos' fixed and mobile networks sector. Economic sustainability is related to telco networks' cost savings and revenues increase.

The telco sector's economic sustainability is a crucial issue. Credit Suisse analyzes the European Telecoms industry's long-standing challenges and gives the following view: «*we find the main European telcos generate returns ~250bp below the cost of capital with annual value destruction ~5% of market cap. Major telcos would need to increase FCF by ~75% (€13bn) to reach a level where the sector was no longer destroying economic value.*» [3]

In this paper, we present solutions based on network architecture transformation and Edge Cloud Computing (ECC) at the edge of telecommunication networks, i.e., close to end users, that can improve the management of application services. ECC, employed for many years by OTTs in their networks, improves application performance through layer 4 techniques that manage application transport.

ECC is a mini/micro-data center that distributes content and application services close to end users. Distribution enables improved application service performance and network cost savings. Improved performance of application services is provided by applications and content partially managed by the ECC, which perform better than applications and content fully managed by the centralized server/ cloud. Performance improvement is achieved with network KPIs from the end-user equipment to the ECC, which are better than end-to-end network KPIs (from the end-user equipment to the server/cloud). In this paper we also show that network cost savings can be achieved by reducing traffic peaks in the network segment from the ECC to the interconnection with OTT networks.

Telcos focus on network services, i.e., the transport of IP packets (layers 1, 2, and 3 of the IP protocol stack). The traditional telco architecture utilizes layer 2 tunnels, from the end-user device to the telco core network, and layer 3 for the interconnection to other networks and autonomous systems.

One main reason for the telco sector affording difficulties in their business is related to the application services' quality and the sector's economic sustainability. Telcos attempt to improve service quality and performance through QoS-based traffic management techniques, such as bandwidth reservation and packet prioritization at layers 2 and 3. QoS techniques are much less effective than layer 4 ECC techniques in providing the technical quality needed by the applications and do not ensure network cost savings. Moreover, using layer 2 tunnels from/to the end-user device to/from the telco core network does not give visibility of the end-user layer 3 needed to deploy ECC platforms. Then, to have layer 3 visibility, some core network functions should be distributed. This requirement increases the complexity of the network and the costs but, as this paper shows, the deployment of ECC platforms allows us to achieve network cost savings and paves the way for new business opportunities between telcos and OTTs (i.e., double sided platform business).

In many cases, telco networks limit the technical QoE for applications and content. For example, video streaming provided by OTTs (e.g., Netflix, Amazon Prime, and DAZN) to the telcos has different resolutions from lower than standard definition to higher than full high definition and up to 4K ultra-high definition. However, only a limited number of telcos' end users today experience full HD quality.

This paper's primary goals are:

- To analyze problems the telco industry faces that limit applications' technical QoE and hinder economic sustainability.
- To present possible solutions based on telco network architecture transformation and solutions to speed up the ECC deployment in access networks.
- To evidence the main advantages for telcos concerning the deployment of ECC platforms in their networks, i.e., network cost savings and double-sided platform business opportunities.

The paper outline is as follows. Section 2 presents the application services, network services, and the relevant KPIs; we also discuss some main challenges for regulators and domestic telcos. Then, Section 3 provides the main concepts of ECC, discusses application performance improvement and cost reduction, and provides a cost analysis. Section 4 considers the main challenges for the telco domestic networks related to the ECC deployment and highlights some possible solutions. Finally, in Section 5, we draw our main conclusions.

2. APPLICATION SERVICES, NETWORK SERVICES, AND RELEVANT KPIS

Services provided on the Internet are often called application services or applications. The reason is that applications are managed end-to-end by software executed in the server/cloud and the end-user devices. Hence, application KPIs are related to the end-to-end connection from the end-user device to the server/cloud to provide a service. The main technical KPIs for the applications, including content delivery, are first defined and then some comments/examples are given to show that these KPIs can be used to give the application services target requirements. Applications' KPIs are latency measured through round-trip time, *RTT*, throughput, *TH*, download time, *DT*, and video delay, *VD*. Latency (in the order of ms) is the time a packet goes from source to destination and vice versa. The *TH* (order of *Mbps* or ten *Mbps*) is the 'speed' (a.k.a., bandwidth) measured at the application layer and may differ from the available bit rate. The *DT* (seconds) is a performance parameter related to the response time to requests from an end user, for example, the response time of a web server. Finally, the *VD* (seconds) is the time between the instant when capturing a frame and when displaying that frame at the end-user device. *VD* is an important KPI for livestreaming services.

Typical examples of KPIs for some application services are as follows:

- For 4K Video on Demand (VoD) streaming, the main KPI is the throughput. The *TH* value defines the video quality. Typically, *TH* for 4K VoD is about 15 Mbps, for full high definition it is 7 Mbps and for high definition it is 5 Mbps. The *DT* is also a KPI for VoD and livestreaming services; however, it can be easily managed.
- For 4K livestreaming, the main KPIs are throughput and video delay. *TH* is about 20 Mbps, and *VD* is about 30 s. Livestreaming *TH* is higher than 4K VoD to reduce *VD* (the full VoD video compression requires higher computing time and increases *VD*, which should be much lower than 30 s for live events).
- For autonomous driving, the main KPIs are RTT (value is about 1 ms) and throughput (value is about 10 Mbps).
- For 360 degrees augmented reality, the main KPIs are *RTT* (value is about 1 ms) and throughput (value is 150 Mbps-1 Gbps).

When Internet services only were narrowband or broadband, latency was generally not critical. Its centrality has begun to be recognized for ultra-broadband services to be carried out with VHC networks. Any real-time or near real-time communication on the Internet is inherently two-way, which means that one elementary communication (outbound packet) ends when a feedback loop is closed (return packet). High latency in bidirectional communications can be a problem in itself. This fact has been well-known since geostationary satellite communications introduced latencies of several hundred *ms*, which are annoying for human communication and echo cancellation. The limit for real-time perception and reaction for future haptic communications (e.g., telesurgery, remote-controlled robots, autonomous driving) dramatically lowers to 1 *ms* or less; therefore, it is much more critical [8]. This essential criticality starts to be recognized for future mission-critical networks, especially those for ultra-Reliable

Low Latency Communication (uRLLC): «*Early efforts that are just starting for the development of 6G are also emphasizing this aspect (...) While URLLC support has also been emphasized for 5G, that support focuses on the latency incurred at the edge and between end device and network (antenna and front-/mid-/backhaul), not between end-to-end communication peers and across the network core, which remains an open problem.*» [9] See the appendix for additional considerations on how ECC can alleviate and possibly solve the end-to-end latency problem for mission-critical uRLLCs.

The effect of latency is twofold in networks, as it can also affect the throughput of applications. If the available bit rate is not the bottleneck for *TH*, the limit is given by congestion control algorithms that manage the transport of the applications between the end-user's device and the server/cloud. The *TH* depends on network KPIs utilized by the congestion control algorithm, such as *RTT* and Packet-Loss, *PL* [10-12] or Bottleneck Bandwidth and *RTT* for congestion control BBR (Bottleneck Bandwidth and Round-trip propagation time) [13], or other KPIs for congestion control based on machine-learning techniques [14]. The network end-to-end KPIs' deterioration, such as higher latency and packet loss, reduces the *TH* and worsens all the other applications KPIs.

The primary end-to-end network KPIs are available bit rate, packet-loss, and latency (i.e., *RTT*). *PL* is the fraction of IP packets not received within a defined time interval due to congestion of network routers or bit transmission errors. Latency is both a major network and application KPI.

The bit rate, *BR*, is different from the *TH*, which is the speed of the application. The available BR (order of *Mbps* or ten *Mbps*) is the transmission channel speed for the application in question, available between the end user's device and the server/cloud. Many IP flows share the network link capacity. Hence, the *BR* available to an application depends on the number of active applications (i.e., the number of IP flows) and the applications' KPI requirements. The available *BR* gives the upper limit to the *TH* for any application.

Network and application KPIs are Random Variables (RVs) or RV functions. Hence, mathematical models must consider distribution functions (or moments) and density measured at peak times, which are difficult to obtain. Also, the available bit rate is a function of RV. This article only considers architectural issues and does not present mathematical performance [15]. For more details, see [6, 7].

Layer 4 of the IP protocol stack (end-to-end application transport) mainly manages application KPIs that depend on network KPIs, i.e., the end-to-end transport of IP packets between the server/cloud and the end-user device. Layers 1 (physical) to 3 (network) of the IP protocol stacks handle the transport of IP packets.

The technical QoE required for the different services depends on the application KPIs of the telco and OTT networks. In addition, technical QoE also depends on the performance and processing power of the clouds and end-user devices. Telco networks connect the end user's device to OTT networks or other networks, such as Content Delivery Networks (CDNs) and clouds. OTT clouds provide end users with most of the application services.

2.1 Regulators and KPIs

Regulators are not yet correctly addressing application performance KPIs in their institutional activities related to VHC telco networks. The reference to application and network KPIs is essential to putting into perspective regulatory policies that seriously affect the cash flow of telcos and the sustainability of the industry's businesses.

Changes to the regulatory definitions for VHC networks are needed. The European Commission has promoted the concept of telco VHC networks with the European Electronic Communications Code (EECC) to accelerate the European Gigabit Society (EGS) target for the year 2025 [16]. The concept revolves around the development of optical fiber or equivalent access networks and considers latency by itself; see recital 13 of the EECC: «*The requirements concerning the capabilities of electronic communications networks are constantly increasing. While in the past the focus was mainly on growing bandwidth available overall and to each individual user, other parameters such as latency, availability, and reliability are becoming increasingly important. The current response to that demand is to bring optical fiber closer and closer to the user, and future 'very high capacity networks' require performance parameters that are equivalent to those that a network based on optical fiber elements at least up to the distribution point at the serving location can deliver.* (...)»

Some of the main high-level comments on the European regulatory approach are as follows. Very few regulatory documents provide clear definitions for network and application KPIs, in particular:

- The difference between bit rate and throughput.
- The minimum bit rate available per active user during the peak hour.
- The dependence of throughput on network KPIs (such as available bit rate, latency, and packet loss utilized by congestion control algorithms.
- The need to ensure end-to-end QoE from the server/cloud to the end user's device and vice versa.

For example, recital 13 of the EECC refers to the total physical transmission of the available *BR* and the available *BR* per user ("*growing bandwidth available overall and to each individual user*") and "*to bring optical fiber closer and closer to users*." However, despite substantial investment in underground infrastructure, this alone cannot provide a marked improvement in application performance. The main requirement for a VHC network in ECC recital 13 and related documents is the implementation of optical fiber in the access network without any references to end-to-end application KPIs. More stringent requirements related to available *BR* and other network KPIs are indicated for mixed copper-fiber access networks (see Berec, [17]). Furthermore, for any communication media, including fiber, Berec makes reference to a "*multi-dwelling building*" point, not the home, and "*high typically achievable data rates*." In no case is the dependence of *TH* on latency and other network KPIs recognized.

While for old narrow/broadband systems, this dependence has no practical effect in ultra-broadband systems on the Internet, in modern VHC fixed and mobile telco networks, such as 5G-based systems for present eMBB (enhanced Mobile Broadband) services, and, even more so, future advanced ones for uRLLC services, the situation tends to reverse. Depending on the mechanism adopted for congestion control and as a function of the network KPIs in a hypothetical bandwidth-latency plane for requirements (such as the one represented in [18]), the services fall into two regions: band or "*bit rate-limited*" and latency or "*RTT-limited*" [19]. This consideration has profound implications for the success of regulatory policies and, ultimately, for customer satisfaction and industry's sustainability.

Said differently, interventions on the BR only and other network KPIs cannot substantially improve customer satisfaction. At the same time, "*the sector [goes on] destroying economic value*" [3] if the regulations do not consider the end-to-end technical QoE of the applications at all.

The requirements for the access networks refer to the available BR. They should be provided end-to-end for all network and application service KPIs, including domestic telco and OTT networks:

- The endpoints for domestic telco networks are the end-user device and the interconnection with OTT networks.
- The endpoints for the OTT networks are the interconnections with the telco networks and the server/cloud.

For any Internet ecosystem stakeholder, the main goal of the VHC network should be improving the technical QoE of applications, as this is one of the most crucial drivers for GDP growth.

Therefore, regulators should provide indications, or even incentives, to accelerate the transformation of the telco network architecture based on the deployment of layer 4 platforms within the access networks close to the end users. Layer 4 platforms improve all the application KPIs, reduce the network Total Cost of Ownership (TCO) in many cases (for conditions to obtain savings see Section 3 below) and enable new business models and revenues. Unfortunately, application service KPIs and layer 4 platforms remain largely beneath the regulators' radar today.

2.2 Telcos and KPIs

To improve the application's performance, a telco typically works on the transport of IP packets which, for the traditional architecture and from the end-user device to the telco core network, is a layer 2 tunnel. The transport of IP packets is improved by QoS-based traffic management techniques, such as bandwidth reservation and packet prioritization at layers 2 and 3 of the IP protocol stack. While these actions can help restore network faults and congested network paths, they can provide a limited application performance improvement because applications' KPIs are much more effectively managed at layer 4, as we'll discuss in the following.

Through bandwidth reservation, QoS can help manage or eliminate bottlenecks related to available bit rate. However, this can lead to regulatory problems related to net neutrality. These problems can arise if, in the daily management of connectivity, the QoS parameters used for the telco network and application services differ from those used for OTTs and other Content Provider (CP) application services. According to the principle of fairness, steady-state congestion control algorithms provide the same *BR* to all IP flows with the same end-to-end network KPIs. However, for an Internet network that handles many IP flows, the steady-state is never fully achieved due to the start and end of IP flows.

Though technical QoE mainly focuses on the application services and QoS focuses on the network services, they are related. Traditionally, a telco tries predicting QoE based on a construct of QoS-QoE models. However, telco's predicted QoE turns out to be a loose function of the set of network (averaged) performance KPIs. Telcos traditionally manage the IP packet transport and do not manage the transport of the applications (i.e., layer 4 for any service and end user). Consequently, a telco manages averaged network KPIs and generally cannot capture the quality perceived by its end users one by one.

Other main differences between the approaches to improve the performance of applications based on the technical KPIs of QoE at layer 4 and the techniques of QoS at layers 2 and 3 are as follows. First, QoS can provide some IP streams with a higher available bit rate (via bandwidth reservation) and lower RTT (via higher priority) to reduce queuing delays. However, this result penalizes IP flows without reserved bandwidth and at high priority. Furthermore, the improvement of the application KPIs concerns a limited number of IP flows and is very 'low', while the performance degradation of the penalized IP flows is much higher.

Second, QoE techniques are mainly based on the partial distribution of applications and content close to end users in the ECC mini/micro-data center. The distribution provides application KPI improvement as a function of:

- The fraction, *HR*, of the applications and content being managed in the mini/micro-data center. This fraction is called Hit Ratio ($0 \le HR \le 1$)

- The ratio $q = RTT_{\text{Tot}}/RTT_{\text{ECC-to-End-user}}$, where RTT_{Tot} is the *RTT* from the server to the end-user device and $RTT_{\text{ECC-to-End-user}}$ is the *RTT* from the ECC to the end-user device. This ratio is always > 1 and usually has values that provide a much more extensive application KPIs improvement than the improvement obtained through QoS techniques.

Third, QoE techniques do not penalize other applications and provide savings on network costs. Deployment reduces the peak *TH* from the mini/micro-data center to the OTT network interconnect, and the network cost depends on the peak *TH*.

3. EDGE CLOUD COMPUTING

The crucial technology for the needed paradigm shift for the telcos' industry is Edge Cloud Computing (ECC). ECC is an evolution of cloud computing. It brings hosting of some applications and content from OTTs' data centers (a.k.a. centralized data centers) down to mini/micro-data centers at the network "edge," closer to end users. OTTs have hyperscale clouds/data centers very far from the telco networks and different levels of medium-small clouds/data centers. Some medium-small clouds/data centers are not far from the telco core network. The ECC connection to OTTs' data centers (i.e., centralized clouds) is needed as the applications and content are dynamically distributed in the ECC platform according to the end-user requests and usually are not fully managed in the ECC.

The applications and content distribution improves all the application KPIs quoted above, as the network KPIs utilized by the congestion control algorithms are improved. We have already provided some comments related to *TH* in the previous section, and similar comments apply to all other application KPIs [2].

3.1 ECC principles

ECC provides applications' performance improvement by distributing applications and content in the ECC server close to the end-user device. Different criteria manage the distribution. We will not analyze them in this paper. In synthesis, the most important are: the need to improve the application KPIs, the cost versus performance trade-off, the number of end users that watch content on demand and live video streaming, and the processing power available in the ECC. The latter can be increased by dedicated processors, such as, e.g., deep learning processors, and field programmable gate arrays based on accelerators.

We can also achieve application performance improvement by terminating (i.e., 'closing') the congestion control algorithms, usually managed by layer 4 and in some cases by layer 5 protocols, at intermediate points between the end-user device and the server/cloud. This solution does not use physical distribution at a server close to the end-user device. It adopts protocol accelerators that close the congestion control algorithms at intermediate points [20]. In a telco network, they are in the path from the access POP to the interconnection with the OTTs networks. Shorter network segments for the congestion control algorithm improve the end-to-end applications' KPIs which are limited by the most extended segment. Protocol accelerators can provide applications' technical QoS improvement without using a micro/mini-data center and, for the telco networks, without the need to distribute some of the core functions to have layer 3 visibility. However, they increase the network TCO as the peak TH from the end user to the server/cloud grows.

The distribution of content is based on transparent caching both for video on demand and live video. Zipf's law provides the theoretical foundations for the advantages of using transparent caching. The relative frequency with which end users request content follows a Zipf-like distribution, where the relative probability of a request for the *n*-th most popular content is inversely proportional to n^{α} with α taking on some value less than unity, typically ranging from 0.64 to 0.83 [21]. Therefore, it is enough to store a limited amount of content in the nearby transparent cache to achieve high values for the probability that a cache delivers content (i.e., the hit-ratio, HR), avoiding a high number of requests and content to traverse the network. For $\alpha = 0.8$, only 10% of content stored provides $HR \approx 50\%$, which means cutting in half the heavy video traffic.

Application distribution and transparent caching effectively reduce the IP downstream data traffic in the network between the server/cloud and the ECC. Then, the network upgrades needed to manage the growth of the IP traffic volume have a much lower cost. Consequently, in many cases, the ECC network architecture TCO is lower than the legacy network TCO [6, 7]. The following section provides a high-level analysis defining the saving conditions.

The ECC also plays an essential role in transforming telco's business models. Thanks to this, telco networks can become versatile service enabling platforms for the telecommunication industry and OTTs. ECC supports this transformation, as it opens the telco networks edge for applications and content, including those from third parties [22]. For the telcos, this business transformation enables new revenues, based on application performance improvement, from both the telco end users and the OTTs (two-sided platform business model).

The telco industry is entirely in line with the importance of ECC to improve technical QoE, obtain network cost saving under conditions that are usually met, and enable incremental revenues. ETSI presented Multi-access Edge Computing (MEC), a standardized ECC architecture. In a white paper, ETSI introduces MEC in the following words: «*Multi-access Edge Computing is regarded as a key technology to bring application-oriented capabilities into the heart of a carrier's network, in order to explore a wide range of new use cases, especially those with low latency requirements. When it comes to deploying MEC, there are many potential scenarios where MEC can fit in, and – as the name clearly spells out – these are not limited to 4G or 5G at all! As a universal access technology, MEC offers itself to any application that has locality requirements like a shopping mall or a sports arena, or wherever low latency is required.*» [23] The network KPI considered is the latency, as obtaining technical QoE improvement by latency reduction is more manageable than working on other network KPIs such as packet loss. Relying on 'low latency' is more technically sound than dwelling on QoS techniques.

OTT networks "edge" is the interconnection point between autonomous systems. However, in this paper, application service performance is addressed. Then the focus is on the interconnection between OTTs and telco networks that can be at the telco network border or inside the telco's core network.

The "edge" for the telco networks is usually at the fixed network access POPs, the 4G and 5G mobile Radio Access Network (RAN) sites, or the end-user sites [4]. The mobile network edge is usually the base station aggregation point for the centralized RAN architecture. For the distributed RAN architectures (cloud, virtual and open RAN), the edge can be in the centralized unit or, to meet the KPIs of some very demanding new services, in the distributed unit. Other definitions of the telco networks edge are also given [24].

In domestic telcos' networks, the above definition is sometimes called the telco's "far edge" (far refers to the network core, also called IP Edge), and the "near edge" is a network POP between the access and the core network, such as a metro/regional POP [25].

Deploying ECC platforms in more centralized locations reduces the technical QoE improvement due to worsening network KPIs from the end-user equipment to the ECC. However, as the ECC unitary cost depends on the managed peak *TH*, it can provide network cost savings due to a higher peak *TH* in the far edge. We cannot obtain this result by ECC platforms installed in access POPs with a very low peak *TH*. However, the peak *TH* in the far edge can save network costs in most cases. This result is because of the peak *TH* growth in the high-speed telco networks, the growing traffic aggregation due to the reduction of the fixed networks access POPs number, the cloud RAN architecture, and the increase of the mobile networks' base stations.

ECC has been widely used since about 2005 by OTTs (such as Amazon Web Services, Google, Microsoft, and Netflix) and content delivery providers (such as Akamai, Limelight, and CD Network) to improve service KPIs and reduce network costs. The ECC platforms were one of the main drivers that enabled the growth of Internet services [4]. They allowed the provisioning of OTT services all over the world to end users that are thousands of *km* far from the centralized clouds (medium, big, and hyperscale data centers).

3.2 Application performance improvement and cost reduction in telco networks

ECC can effectively address application performance and cost issues of the fixed and mobile high-speed (VHC) telco networks. Moreover, ECC is acknowledged as one of the key pillars for meeting the demanding KPIs needed for new and future services and distributing telco networks' AI-based operations and management. For this reason, 5G has been designed to provide enablers to ease ECC deployment. In 5G, we can use the User Plane Function (UPF) and the flexibility in locating UPF to distribute applications and content to the telco edge. According to [5]: «*The control of that data plane, i.e., the traffic rules configuration, now follows the NEF-PCF-SMF route* (Network Exposure Function-Policy Control Function-Session Management Function, *ed.*). *Consequently, in some specific deployments, the local UPF may even be part of the MEC implementation*.»

The throughput-over-available bit rate ratio, i.e., the bandwidth utilization ratio $r = TH/BR$, is significantly reduced in telco's high-speed networks having much larger bit rate values than narrow/wideband networks. A low value for r harms telco's economic sustainability as it increases TCO and does not enable network monetization, mainly based on application performance improvement. When throughput service requirements become more severe, as it is with time when moving towards new and future services, retaining the legacy network architecture while increasing the available *BR* in the access network causes the telco to invest in infrastructures and access technologies only. Consequently, the TCO increases, while the customer satisfaction (i.e., the technical QoE) tends to be very poor.

Data related to the bandwidth utilization ratio can be obtained by Ookla and Netflix that provide respectively average and median monthly available bit rate measurements and averages over the last six months for on-demand (VoD) streaming throughput. For Italy, Germany, and France, Table 1 shows May 2022 Ookla average and median bit rate measurements [26] and Netflix *TH* average measurements from December 2021 to May 2022 [27].

All bandwidth utilization ratios are very low (the highest is 8.8% in Italy, which has the lowest median mobile available *BR*). Then, it is evident that the applications use a very 'small' percentage of the communications channel available *BR*.

Note that utilization ratios and performance for any application service, including services where both uplink and downlink have an impact on performance (such as file transfer, video conference, remote teaching, and 360 degrees augmented reality) depend on applications and network KPIs, on 'how' the applications are managed (at layers 4 and 5 of the IP-IETF protocol stack) and on the end-user device (operating system and processing power). Then applications' performance and utilization ratios (downlink and uplink) can differ. In a specific network and geographical area, the differences are given by the platform used to manage the application (such as the adaptive streaming protocol) and by the end-user device. Zoom throughput requirements are group video calling *TH* (up/down) = 600 Kbps/1 Mbps; HD video *TH* (up/down) = 3,8 Mbps/3 Mbps or 2,6 Mbps/ 1,8 Mbps; gallery receiving view *TH* from 2 Mbps to 4 Mbps; screen sharing 50/150 Kbps. Note that Table 1 gives data for the VoD streaming with the throughput as the main KPI and no live constraints on video delay. The absence of these constraints allows the use of 'big' video delays (i.e., video buffers, up to more than 1 minute). At the same time, throughput is higher (better video quality) than for livestreaming services.

Some years ago, the telco industry started working on cloud-based architectures, Network Function Virtualization (NFV), Software Defined Network (SDN), and MEC. The architecture evolution towards ECC celebrates today's first successes and is becoming progressively more important. However, a strong legacy in telco architectures and protocols is delaying this transformation so that the «*cloud-native shift will require a comprehensive overhaul of network architecture and is unlikely to happen fast.*» [28].

Table 1 – Throughput over bit rate measurements from Netflix and Ookla data in Italy, Germany and France

Ookla average/median download measurements			
Italy			
	TH [Mbps] Netflix	Bitrate [Mbps] Ookla	*TH*/Bitrate [%]
Average mobile download	3.4	65.2	5.2
Average fixed download		127.5	2.7
Median mobile download		38.5	8.8
Median fixed download		53.9	6.3
Germany			
	TH [Mbps] Netflix	Bitrate [Mbps] Ookla	*TH*/Bitrate [%]
Average mobile download	3.4	96.4	3.5
Average fixed download		148.0	2.3
Median mobile download		53.7	6.3
Median fixed download		74.2	4.6
France			
	TH [Mbps] Netflix	Bitrate [Mbps] Ookla	*TH*/Bitrate [%]
Average mobile download	3.2	113.7	2.8
Average fixed download		247.5	1.3
Median mobile download		60.2	5.3
Median fixed download		113.0	2.8

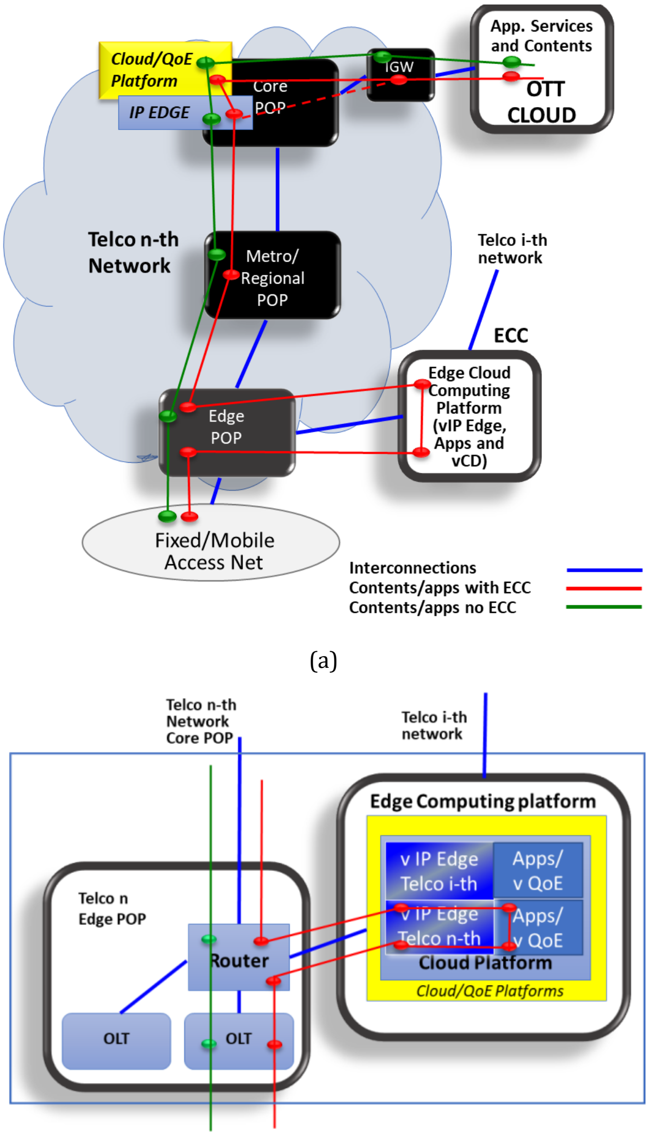
In short, the VHC telco network issues depend on the telco network architecture, which does not manage applications' technical QoE. This condition may limit network performance, not allow network cost reduction, and not enable VHC monetization based on a two-sided platform business model. Therefore, the traditional telco architecture requires significant transformations, such as control and user plane separation, layer 3 (IP) transport all over the network, and ECC platforms deployment. Especially the latter improves application service performance and can, in many cases, reduce the network TCO. ECC platforms in the telco networks can provide the technical QoE needed by the different applications and are easily scalable, i.e., they can improve the application performance if needed.

The ECC high-level architecture (Fig. 1) presents the interconnection of the ECC to the telco edge POP and the traffic flows for applications and content not managed by the ECC and the main ECC components. Two or more telcos manage the shared ECC.

The main components of the ECC, and MEC, are:

- The micro-mini-cloud (or data center), i.e., HW (server and memory) and SW to create and manage virtual machines (or containers) to execute the virtual IP edge SW, applications, software platforms that provide services, and content delivery platforms (such as virtual CD platforms, transparent cache, accelerators, and front-end optimization platforms).
- The virtual IP edge distributing by virtualization of some of the core functions to obtain the visibility of layer 3 of the IP protocol stack (the IP edge is the core of the telco network: 'IP' indicates the layer 3 visibility and 'Edge' is related to the interconnection with other networks).
- Applications and software platforms that offer ultra-broadband services and virtual CD platforms.

The figure places the ECC platform outside the telco networks. However, the platform can be colocated at the edge POP. This architecture effectively shares the platform built by a neutral host among telcos that deploy their ECC platforms by Infrastructure-as-a-Service (IaaS) and Platforms-as-a-Service (PaaS). The platform manages downstream and upstream traffic flows as follows. The router at the edge POP makes traffic classification and steering and sends/receives traffic to/from the end users,



(a)

(b)

Fig. 1 – (a) Interconnections and traffic flows among telco networks and Edge Cloud Computing (ECC) platform; (b) Telco access POP and ECC platform architecture main components.

other OTT and telco networks, and the ECC platform (red line). Other traffic (green line) does not reach the ECC. The IP edge first manages the traffic sent to the ECC to have layer 3 user plane visibility. By traffic classification, we can deliver it to applications, software platforms, or CD platforms that, after processing, give the traffic back to the IP edge and then to the router in the telco POP.

The main drivers for ECC deployment are related to the applications and content distribution that improves downstream applications performance (technical QoE) and network TCO savings.

Downstream applications performance improvement enables VHC network monetization, new business models, and new revenues. ECC always provides an application performance improvement. In many cases, we can obtain network TCO savings for RAN and the segment from the ECC site (access network) to the interconnections with OTTs. We can obtain RAN saving by increasing the bandwidth utilization ratio, *r*. In 5G, larger values for r provide more intensive utilization of the radio-link bit rate and reduce the number of RAN sectors needed to manage the total TH. The saving for the segment from the ECC site to the interconnections with other networks is obtained by application and content distribution, which reduces peak TH.

3.3 Application performance improvement and cost reduction in telco networks

For a specific network with its topology, technologies, and network segments unitary costs we define the cost saving, S%, as:

$$S\% = 100\,\frac{NC(noECC; TH = THq) - [NC(with\,ECC; THq) + ECC\;cost\;(THq)]}{NC(noECC; TH = THq)} \tag{1}$$

where:

- *NC*(no ECC; *TH*=*TH*$_q$) is the cost of the traditional IP network (i.e., centralized IP edge and no ECC platforms). TH is the peak network throughput without ECC, and THq is the peak network throughput with ECC. *TH*$_q$ > *TH* as the traffic managed by the ECC is closer to the end-user equipment. The cost of the traditional network is evaluated for *TH* = *TH*$_q$, to make a fair cost comparison because the network cost depends on total peak throughput. Note that total peak throughput grows according to volumes (*Pbyte/month*) and the ratio total peak throughput over total average throughput.
- *NC*(with ECC; *TH*$_q$) is the ECC network architecture cost (i.e., the cost of the network with the ECC platform). *NC*(with ECC; *TH*$_q$) is a function of the speed up *SU* = *TH*$_q$/*TH* and of the network components' costs and is lower than the costs of the network segments without ECC.
- *ECC cost* (*TH*$_q$) is the cost of the ECC platform.

Eq. (1) is general and provides conditions for network cost savings for any application service and scenario, both fixed and mobile, including vehicular technology and autonomous driving.

We obtain a saving if the following condition is satisfied:

$$\textit{ECC cost}(\mathrm{THq}) < \textit{NC}(\text{no ECC; TH=THq}) - \textit{NC}(\text{with ECC; TH=THq}) \tag{2}$$

i.e., the ECC platform cost is lower than the network costs saving obtained by the ECC platform. This condition is usually satisfied, by considering the present average network and ECC costs if *TH* > 6-8 *Gbps* (note that ECC costs have a higher reduction versus network costs).

To give examples and a range of the applications' performance improvement and the network cost saving that can be obtained by ECC platforms, some results of simulation case studies are presented. More details are reported in [6].

The primary objective in the design of an end-to-end VHC network architecture is to achieve high application services' performance, while minimizing the TCO across the entire infrastructure. To achieve this goal, the design approach, based on ECC and network modeling, asks for a solution of the ECC location problem. The analysis is focused on how different options for ECC deployment affect cost savings and throughput improvement.

A fixed ultra-broadband VHC network with 5 core nodes, 25 regional nodes and 250 access nodes is considered. The simulations analyze the results obtained by ECC platforms deployed in the access nodes that manage video content delivery by transparent caches. The network without ECC was modeled by network KPIs (to obtain the network *TH*) and by network segments costs (to obtain the network cost). The network with ECC was modeled by considering the speed-up, i.e., *SU* = (*TH* with ECC)/(*TH* without ECC) > 1, to obtain TH_q , the throughput for the network with ECC. The transparent cache hit ratio, *HR*, is 50%. Finally, it is also assumed that the ECC platform can manage up to 10 *Gbps* (i.e., when the total throughput in an access node is >10 *Gbps*, the ECC cost is evaluated for a capacity of 20 *Gbps*).

In the following figures, we present the network speed-up and savings as a function of the (increasing) number of access nodes with ECC. The ECC platforms are in the access nodes starting from the node providing the highest saving.

In Fig. 2 we report the speed-up of each single access node (*SU*(*i*), dashed lines) and the network *SU* (*NSU*(*i*), solid lines), i.e. the *SU* related to the total network throughput and obtained by increasing the number of access nodes where the ECC platform is deployed.

The access nodes in the figures are ordered according to the savings, i.e. the first node has the highest saving and the last has the lowest. When ECC platforms are in all the access nodes the NSU has the highest value. Two scenarios with different *RTTs* between the end users and the access nodes and between the access nodes and the Big Internet are considered, the packet loss for the two scenarios are the same (i.e., no *PL* improvement is obtained by ECC). Scenario A refers to a lower network *SU* = 1.75 and scenario B refers to a higher network *SU* = 3. In the simulations we assume that the network bit rate is not the bottleneck and then the increase of the number of access nodes equipped with the ECC improves network *SU* and the network application throughput.

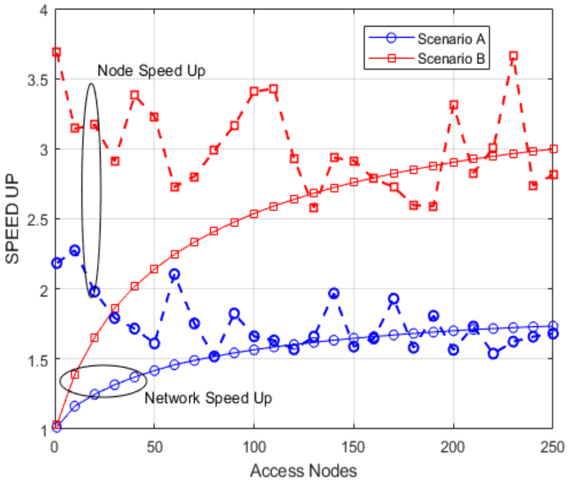


Fig. 2 – Network speed-up vs access nodes with ECC for scenario A (*NetSU*=1.75) and scenario B (*NetSU*=3)

In figures 3 and 4 the network saving as a function of the nodes with ECC [NS%(*i*)] and the network saving coefficients (*Cx*%, *Cx* < 100% indicates cost saving) are reported for total network throughput without ECC (TH_{tot}).

The difference between the two simulations is only the network speed-up. *Cx* is the value obtained when the ECC is in all the access nodes. TH_{tot} has three values 0.2, 0.5, 5 *Tbps*, is exponentially distributed in the access nodes, and the average *TH* value for the access nodes are respectively 0.8, 2, 20 *Gbps*.

The main results for network savings related to the average total network *TH* and the network speed-up are the following. By increasing TH_{tot}, a higher number of network nodes can provide savings. This is due to unitary ECC cost that is a decreasing function of the total *TH* managed by the access node. By increasing *SU*, the network saving is higher.

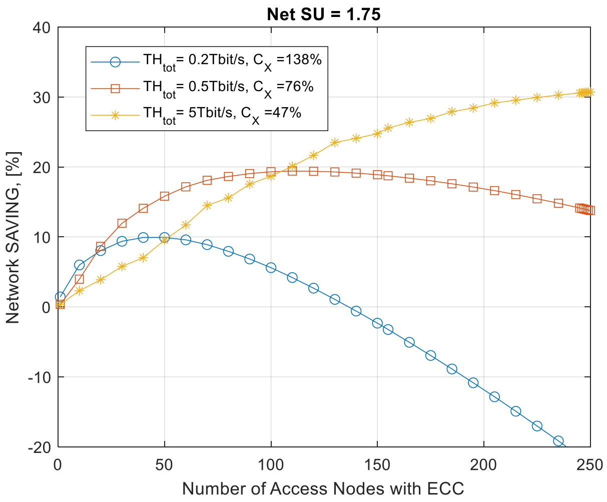


Fig. 3 – Network savings vs access nodes with ECC, for different values of total *TH* = 0.2, 0.5 and 5 *Tbps*

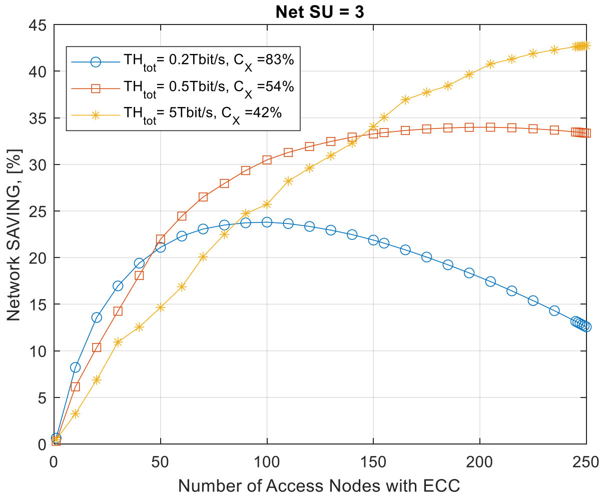


Fig. 4 – Network savings vs access nodes with ECC, for different amounts of total network *TH* = 0.2, 0.5, and 5 *Tbps*.

The simulation results show that the introduction of ECC platforms in the network nodes is always able to provide performance improvement and that cost savings can be obtained if unitary ECC cost is lower than unitary network cost between the ECC site and the Big Internet. The unitary network costs for the ECC and the network depend on the costs of the ECC HW and SW components and the network elements, the nodes speed-up, the total throughput managed by the nodes ECC platforms and the hit ratio, that in general is the fraction (or percentage) of the applications and content downstream traffic that is managed in the ECC micro/mini-data center (the model based on *HR* is general and is not limited to content delivery by transparent caches).

Cost saving in the range 25-40 % can be obtained also having "high" network saving coefficients (see figures 3 and 4) and, in general, unitary ECC cost is lower (in many cases much lower) than unitary network cost. Moreover, ECC cost can be significantly reduced by ECC platform sharing. The network cost savings obtained by ECC are presented in [6] for a fixed network, in [29] for the 5G network deployment in the UK, and [30] for CAPEX/OPEX active and passive network component sharing.

The ECC platform sharing is in line with one of the main trends for telco networks. They are deploying both the traditional passive infrastructure sharing (such as ducts, dark fiber, colocation) and the active network components sharing (e-NodeB, active backhaul components). The sharing can be done among telcos, or a Neutral Host (NH) can offer it. In both cases, the main ECC sharing driver is platform cost saving. ECC platform sharing is an incremental saving added to the network cost saving, S%, obtained by the ECC architecture. ECC platform sharing provides saving as it reduces the Operation and Maintenance (O&M) cost, which also depends on the number of ECC platforms, and improves the ECC platform resource utilization. It provides lower unitary costs due to a higher peak *TH* managed by the ECC. Peak *TH* increase derives from the aggregation of more than one telco traffic.

The NH services that a telco buys are IaaS and PaaS. By sharing the NH mini/micro-cloud, IaaS provides the dedicated cloud infrastructure the telco uses to build its ECC, using telco's SW platforms (IP edge, content delivery, IoT, analytics, machine learning), applications and content. The NH manages the cloud infrastructure, and the telco manages all the SW platforms, applications and content. PaaS provides the dedicated cloud infrastructure and some, or up to all the SW platforms the telco uses to build its ECC. The telco fully controls SW platforms, applications, and content.

The ECC and any active platform-sharing allow the telcos to deploy active network components using the "buy" approach based on IaaS and PaaS. The cloud business widely uses the "buy" approach. In the traditional telco "make" approach, the telco buys, deploys, and manages the HW and SW network platforms (such as the fixed and mobile access and the core platforms). The "make" approach, compared to the "buy" approach, requires specific know-how and competencies. It has higher costs for operations and management activities (that are not shared), can cause economic sustainability problems due to a non-optimal platform utilization, and usually has a higher time-to-market. The "make" approach is the preferred solution in some geographic areas where "make" has a lower cost than "buy," or if it provides competitive advantages and then higher market share/revenues.

In the following section, we analyze the main issues and the speed-up of the ECC deployment in both fixed and mobile telco networks and propose some solutions.

4. ISSUES AND SOLUTIONS FOR THE ECC DEPLOYMENT IN TELCO DOMESTIC NETWORKS

The ECC is a mini/micro-data center with HW and SW components, including the operating system and the infrastructure virtualization. It is connected to centralized data centers, stores content, and executes applications dynamically loaded from the centralized data centers. It works mainly at layer 4 (transport of the applications of the IP protocol stack) and then needs layer 3 (network layer) visibility.

In OTT networks, the network layer is always visible because it is used to interconnect different networks (autonomous systems) and components of one network. Moreover, the OTTs manage applications and content delivery end-to-end, from a single end user to the server/cloud. The functions 'equivalent' to the telco core network are more straightforward and fully distributed. Layer 3 visibility allows the deployment of ECC platforms in any network POPs because the interworking with centralized functionalities is very limited, and then it is not an issue.

On the contrary, in traditional telco networks, layer 3 visibility requires the distribution of some core functions, as the transport of IP packets is made by a layer 2 tunnel from the end-user device to the core network and vice versa.

The main issues in distributing telco core functions in the ECC platform depend on the existing telecommunications standards that define the interfaces to interconnect the core to the fixed and mobile access networks. Today standard interfaces to interconnect the different core entities do not exist or are not widely used. For 4G networks, the entities are Home Subscriber Server (HSS), Serving Gateway (S-GW), Packet Gateway (P-GW), and Mobility Management Entity (MME). However, some vendors already provide non-standard solutions to open layer 2 tunnels, and make IP packets visible (e.g., [31]).

Some of the EPC and the Broadband Remote Access Server (BRAS) functions must be distributed for mobile and fixed networks. The ECC and the telco core network have a strong interaction, difficult to manage due to the existing standards that do not provide the visibility of the interfaces between the core entities (or components) needed to distribute core functions. Moreover, the distributions of core functions give challenges and limitations related to session and mobility management, lawful interception, security, charging, and identifying specific end users at the ECC platform [23].

To give evidence, we present an example. For the 4G and 5G NSA networks, layer 3 visibility requires distribution at the edge site of all, or some, of the EPC entities [23]. The main solutions for distributing the EPC entities are:

- Distribution of all the EPC entities (or components): Home Subscriber Server (HSS), Serving Gateway (S-GW), Packet Gateway (P-GW), and Mobility Management (MME). In this case, the EPC is connected to the RAN, ECC, and OTT/telco networks by standard interfaces (S1-U and S1-MME from RAN to S-GW and MME, respectively; SGi from P-GW to ECC and OTT/telco networks). This solution requires fewer changes to the operator's network. However, each ECC site has all the EPC components (including HSS being distributed), and this gives a higher cost compared to solutions that distribute only some of the EPC functions.
- Distribution of the S-GW and P-GW entities: The control plane functions (MME and HSS) are located at the telco's core site. This solution requires an internal EPC interface (S11, not visible outside the centralized EPC) to connect the distributed S-GW to the centralized MME. Moreover, the S5/S8 interface is used at the ECC to connect S-GW to P-GW.
- Distribution of the S-GW with local breakout: This solution requires internal EPC interfaces (S11 and S5/S8, not visible outside the centralized EPC) to connect S-GW to the centralized MME and P-GW.

5G was designed to provide enablers to ease ECC deployment. However, this solution can be effectively used when the 5G Stand Alone (that manages fixed and mobile networks' core entities) will be massively deployed. This deployment will take a long time, and short-term solutions are needed to enable the effective ECC deployment at a telco network's edge.

The distribution of some core functions in the ECC mini/micro data center at the telco network edge provides the user plane layer 3 visibility. It then allows the application services performance (technical QoE) improvement by distributing applications and content close to the end user. The performance improvement enables incremental revenues from telcos' end users and the OTTs (two-sided platform business model).

The consensus of the telecommunications industry on ECC is now consolidated. However, telco operators have different visions essentially related to "how," "when," and "where" to implement the ECC. "How" has important impact on "when" and "where." Moreover, it can allow the rapid development of the ECC without increasing costs and, in many cases, with a significant saving for the TCO.

For the first set of fixed and mobile access sites, the ECC deployment in VHC networks should be completed in a few months to reduce or eliminate the criticalities related to the technical QoE application services and improve economic sustainability. ECC platforms are now available and can be deployed by changing the traditional approach of telcos and telecommunication vendors to reduce complexity, costs, and implementation times.

Cultural change and innovative approaches are required to address the critical issues related to the ECC architecture deployment. Telcos need know-how (on the cloud and all the ECC components) to manage the ECC deployment in some target geographical areas without significant changes to the traditional network architecture.

We can achieve this by separating ECC platforms from network services and by a fully transparent approach to distributing the core functions. "Transparent" means that the core functions distributed at the ECC are interconnected to the centralized core functions without changing the central core interfaces. This separation allows the adoption of different vendors for the distributed and centralized core functions and avoids vendor lock-in (open-architecture solution). The main requirement to build open architectures is the definition and the use of standard interfaces among the different entities (or components) of the access network (such as radio unit, distributed unit, and centralized unit) and the core network (such as in 4G HSS, S-GW, P-GW, and MME).

Regarding this framework to speed up the ECC deployment, the main approaches are the following ones:

- The core functions are distributed by the centralized core vendor, which can manage the non-standard interfaces between the core entities. This not open solution creates vendor lock-in and can increase the ECC cost.
- The distribution of the core functions is made by an 'Edge-core' platform provided by a vendor different from the core network vendor.

The edge-core platform is an edge platform installed in the edge POPs. It executes core functions that provide user plane layer 3 visibility and traffic classification and steering (e.g., see [32]).

The edge-core platform in the ECC is connected to a support platform, installed in the centralized core, needed to manage the interconnection between the edge-core functions and the centralized core and preserves centralized core functions, including lawful interception, charging, and policy control. This solution is open. However, it needs a support platform in each core site due to the lack of standard interfaces among core entities. The interface between the support platform and the centralized core is the standard interface between the access and the core. The centralized core then sees the edge-core as access POP. The functions of the support platform are not installed in the edge-core as the number of centralized core sites is much lower than the number of edge-core sites. Then the support platform is used to improve the edge-core performance and to reduce its cost (because it reduces the processing power and the complexity of the edge-core).

Telcos also need to identify where the ECC platform will be distributed, taking into account the performance improvement and the network cost savings that, in many cases, can be obtained (as shown in the previous section). Under typical conditions, an ECC platform allows for savings if it manages sufficiently high peak throughput (more than 6-10 *Gbps*). In addition, the saving increases by distributing the ECC closer to the end users. If the minimum throughput condition is not met, some access POPs can share the same ECC platform. The cost reduction is obtained without significant performance reductions if the maximum distance among access POPs is up to 20-30 km. It is not difficult to get throughput values in an access POP allowing for savings, considering the number of active end users during the peak hour and the volume of video traffic. Cisco VNI 2019 data [15] indicates that, globally, IP video traffic, which in 2017 was equal to 75% of all IP traffic, will be 82% in 2022 (a fourfold increase, with CAGR = 29%). The content delivery platforms, which in 2017 managed 56% of total Internet traffic, will manage 72% of total traffic in 2022. CD platforms will be closer to end users, and ECC will play an increasingly important role in improving the quality of applications.

In addition to solving the ECC location problem, the telcos should manage the ECC platform for the fixed/mobile network, completely separate from the legacy network. The ECC architecture, as highlighted in ETSI documents [5] and GSMA [4], favors the platform-sharing between multiple operators and the creation of the ECC by new players, such as tower companies, wholesale companies, and cloud service providers.

As in the hyperscale data center market, the network and platforms, including the ECC (as shown in Fig. 1), will be and, in some cases, are already being developed by third parties. Telcos will therefore be able to create their ECC network and offer ECC services also using third-party platforms by IaaS and PaaS ("buy" deployment model), according to the following guidelines: (*i*) defining sharing agreements with other operators, (*ii*) using third-party platforms, (such as tower, wholesale, and cloud companies, which provide IaaS and PaaS, based on a micro-data center, virtual IP edge, and content platforms' delivery distributed close to end users), (*iii*) defining agreements with vendors or cloud companies that build and manage ECC platforms.

The strategy of quickly introducing the ECC into the mobile and fixed VHC networks and the transformation project of the telco network and application services should be based on these innovative approaches.

In their January 2019 report [24], GSMA Intelligence analyzes the role of telcos in the edge cloud ecosystem: «*The edge cloud ecosystem is broad; it is unclear whether operators will play a dominant role. Operators have been at the center of discussions around the edge opportunity because of their connectivity networks, including network assets at the edge. The broader opportunity, however, relies on a myriad of players - some of which bring assets operators do not possess (platforms) and some of which will compete head-to-head in the network.*»

5. CONCLUSIONS

This paper analyzed the opportunities offered by ECC to deliver application services with the required performance while ensuring the economic sustainability of telco domestic networks enabling new business models and incremental revenues.

Specifically, for VHC networks, the article presented:

1. The need to manage application services.
2. The rationale behind improving technical QoE with ECC architecture.
3. Some comments on how to implement ECC to initiate the transformation of network architectures.

In addition, we provided some preliminary comments on new business models, enabled by network architecture transformation and ECC implementation and capable of generating incremental revenues from end users of domestic telco and OTT networks (two-sided platform business models).

Today, the telecom sector is still too focused on network services. It assumes bit rate as the primary KPI without considering that, especially for VHC networks, the leading indicators are all application KPIs. Instead of bit rate, end-to-end throughput should be considered, which is often much lower than the available bit rate. As a confirmation, the paper reports updated throughput measures in some European countries.

To this day, regulators for VHC networks focus on the bit rate. A high bit rate in fixed and mobile domestic network access is undoubtedly essential. Still, all application service KPIs must be managed and ensured end-to-end from the end-user device to the server/cloud, and vice versa, to improve the quality of application services. Implementing the ECC architecture can guarantee application KPI values required by new and future applications. A new cultural approach to quality of experience and network architecture design by telcos and traditional vendors is also required.

This paper presented some data on network cost savings. The savings can improve the economic viability of telecommunications: a conservative assessment yields savings between 20% and 40% from sharing passive and active network elements and savings between 20% and 50% from the ECC architecture. We can achieve above high total savings with ECC platforms, considering both the savings due to the ECC architecture and those achieved by sharing the ECC platform.

Domestic telco networks should include ECC platforms for both fixed and mobile networks. Finally, solutions based on ECC platform sharing and deployment by third-party platforms, including neutral hosts, using an IaaS and PaaS approach should also be used.

REFERENCES

[1] Recommendation ITU-T P.10/G.100 (2017) Amd. 1 (06/2019), "Vocabulary for performance, quality of service and quality of experience,".

[2] F. Vatalaro and G. Ciccarella, "A Network Paradigm for Very High Capacity Mobile and Fixed Telecommunications Ecosystem Sustainable Evolution," IEEE Access, pp. 135075-135090, 2020.

[3] Credit Suisse, "European Telecoms. Getting out of the rut II," 2022.

[4] GSMA, "The Internet Value Chain 2022," 2022.

[5] ETSI White Paper No. 28, "MEC in 5G networks," 2018.

[6] G. Ciccarella, R. Giuliano, F. Mazzenga, F. Vatalaro, and A. Vizzarri, "Edge Cloud Computing in Telecommunications: Case Studies on Performance Improvement and TCO Saving," 2019 Fourth International Conference on Fog and Mobile Edge Computing (FMEC), 2019, pp. 113-120.

[7] G. Ciccarella, D. Roffinella, M. Vari, and F. Vatalaro, "Performance improvement and network TCO reduction by optimal deployment of caching," 2014 Euro Med Telco Conference (EMTC), 2014, pp. 1-6.

[8] G. P. Fettweis, "The Tactile Internet – Applications and challenges," IEEE Vehicular Technology Magazine, 2014, pp. 64-70.

[9] A. Clemm, M. F. Zhani, and R. Boutaba, "Network Management 2030: Operations and Control of Network 2030 Services," J. Netw. Syst. Manage., 2020, pp. 721–750.

[10] M. Mathis, J. Semke, J. Mahdavi, and T. Ott, "The macroscopic behavior of the TCP congestion avoidance algorithm," ACM SIGCOMM Computer Communication Review, 1997, pp. 67-82.

[11] S. Basso, M. Meo, A. Servetti, and J. De Martin, "Estimating packet loss rate in the access through application-level measurements," Proceed. 2012 ACM SIGCOMM workshop on Measurements up the stack, 2012.

[12] S. Ha, I. Rhee, and L. Xu, "CUBIC: a new TCP-friendly high-speed TCP variant," ACM SIGOPS operating systems review, 2008, pp. 64-74.

[13] N. Cardwell, Y. Cheng, C. S. Gunn, S. H. Yeganeh, and V. Jacobson, "BBR: congestion-based congestion control," Communications of the ACM, 2017, pp. 58-66.

[14] T. Zhang, and M. Shiwen, "Machine learning for end-to-end congestion control," IEEE Communications Magazine, 2020, pp. 52-57.

[15] D. Bertsekas, R. Gallager, *Data Networks*, 2nd Ed., Prentice-Hall Int.l, 1992.

[16] Directive (EU) 2018/1972 of the European Parliament and of the Council of 11 December 2018 establishing the European Electronic Communications Code.

[17] BEREC Guidelines on Very High Capacity Networks, BoR (20) 165, 1 October 2020.

[18] Commission Staff working document accompanying the document Communication From the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, "Connectivity for a Competitive Digital Single Market – Towards a European Gigabit Society," {COM(2016) 587 final}, Brussels, 14.9.2016.

[19] G. Santella and F. Vatalaro, "An approach to define Very High Capacity Networks with improved quality at an affordable cost," 2020 Global Information Infrastructure and Networking Symposium (GIIS), 2020, pp. 1-6.

[20] S. Ladiwala, R. Ramaswamy, and T. Wolf, "Transparent TCP acceleration," Computer Communications, 2009, pp. 691-702.

[21] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web Caching and Zipf-like Distributions: Evidence and Implications," INFOCOM'99 – Proceedings of Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies, 1999, pp. 126-134.

[22] G. Ciccarella, F. Vatalaro, and A. Vizzarri, "Content Delivery on IP Network: Service Providers and TV Broadcasters Business Repositioning," 3rd Int.l Conf. on Recent Advances in Signal Processing, Telecommunic. & Computing (SigTelCom), 2019, pp. 149-154.

[23] ETSI White Paper No. 24, "MEC Deployments in 4G and Evolution Towards 5G," 2018.

[24] GSMA Intelligence, "Distributed Edge Cloud: definitions, dynamics, and drivers," 2019.

[25] GSMA Whitepaper, "Telco Edge Cloud Value & Achievements," 2022.

[26] https://www.speedtest.net/

[27] https://ispspeedindex.netflix.net

[28] M. Lore, "NFV Struggles With Its Six-Year Itch," 2018, url: https://www.lightreading.com/nfv/nfv-strategies/nfv-struggles-with-its-six-year-itch/a/d-id/746707

[29] HSBC Global Research Equities Europe, "5G: What's the use...?" March, 2018.

[30] K. Samdanis et al., "From Network Sharing to Multi-tenancy: The 5G Network Slice Broker," NEC Europe Ltd, url: https://arxiv.org/pdf/1605.01201.pdf

[31] A. Bhalgat, "Accelerating Intelligent Video Analytics Using Ultra-Efficient 5G Core with Mavenir and NVIDIA Edge AI", Oct 05, 2020, url: https://developer.nvidia.com/blog/accelerating-iva-using-ultra-efficient-5g-core-with-mavenir-and-nvidia-edge-ai/

[32] Wind River white paper, "Enabling MEC as a New Telco Business Opportunity," 2017, url: https://events.windriver.com/wrcd01/wrcm/2017/08/Wind-River-WP-Enabling-MEC-as-a-New-Telco-Business-Opportunity.pdf

APPENDIX

Effectiveness of ECC for mission-critical long-distance communications

This appendix analyzes the conditions and solutions to achieve end-to-end latency (*RTT*) lower than 1 *ms* in long-distance communications. We only analyze the problem under very simplified conditions. However, most assumptions can be removed or restated, particularly the deterministic assumption.

Our problem statement is as follows. We assume:

- Two sites, A and B, and a bidirectional link between them; $L/2$ [*km*] is the distance.
- A fiber optical link between A and B (no switch/router, initially).
- Only one active application exists, so only one bidirectional flow between A and B.
- The application has congestion control, and the available bit rate does not limit throughput, *TH*.

We only account for steady-state analysis. Our target is a latency not greater than that allowing us to deliver tactile Internet services, or similar, between A and B (any site can be the source or the destination), i.e., *RTT* is not greater than $RTT_{\text{target}} = 1\ ms$.

Under classical (i.e., non-quantum) physical assumptions, the transmission latency, RTT_{tx}, depends on the speed of light therefore:

$$RTT_{\text{tx}}\ [ms] = L \cdot D_{\text{tx}}/1000 \qquad (3)$$

We account for an A-to-B, or vice versa, latency of the optical transmission of $D_{\text{tx}} = 6.6\ \mu s/km$. Considering only the optical transmission latency, the maximum distance between A and B to have transmission latency $RTT_{\text{tx}} = 1\ ms$ is 150 *km*.

We also account for the latency present in nodes A and B due to processing. This component is added to the latency due to the optical transmission. For simplicity, we lump up the total latency of nodes A and B, assuming it is constant and equal for flows from A to B and from B to A.

The total latency is $RTT_{\text{tot}} = RTT_{\text{tx}} + RTT_{\text{node}} \leq RTT_{\text{target}}$, and thus, the constraint on RTT_{node} is:

$$RTT_{\text{node}} \leq RTT_{\text{target}} - \mathrm{L} \cdot D_{\text{tx}}/1000 > 0 \qquad (4)$$

Therefore, $L_{\text{target-tot}}$, the target distance between A and B to have a latency RTT_{target}, must be:

$$L_{\text{target-tot}} \leq (RTT_{\text{target}} - RTT_{\text{node}}) \cdot 1000/D_{\text{tx}} \qquad (5)$$

If $RTT_{\text{node}} \leq 1\ ms$, $D_{\text{tx}} = 6.6\ \mu s/km$ and $RTT_{\text{node}} = 0.85\ ms$ (i.e., 85% of RTT_{target}), we get $L_{\text{target-tot}} \leq 22.7\ km$ and the distance between A and B can be no greater than about 10 *km*.

We now analyze the case $L_{\text{target-tot}} > L$ to see if and how we can meet the constraint on $L_{\text{target-tot}}$. We rule out bringing sites A and B closer so that *L* is less than, or equal to, $L_{\text{target-tot}}$. In many cases, this is impossible because of physical problems (e.g., the link between New York and Rome or the link via a geostationary satellite). Still, it would be very expensive even if it were possible.

One solution is based on the following consideration. It is not necessary to move all the applications to A or B. If we want to ensure RTT_{target} = 1 *ms* from A to B and vice versa, we need to distribute the applications that must have KPIs with $RTT_{\text{tot}} \leq 1$ *ms* (all KPIs, not just the latency) closer to the destination node (A or B).

In other words, as mentioned in the paper, it is sufficient to distribute closer to sites A and B only the application services (and content), which require KPIs that we cannot meet because of the distance between sites A and B and the latency at nodes A and B (RTT_{node}).

In the present deterministic case, the mathematical models are simple. However, the concepts are common to cases described with increasingly accurate and realistic models and confirmed in network experiences. Regarding the considered case, therefore, we distribute only a portion of the application (e.g., class 5 autonomous driving) or the streaming service (e.g., 8K) to meet the KPIs of the application service requiring the target latency (RTT_{target}) at sites A or B.

Distributing applications near the target site reduces latency and improves all application service KPIs. We must now evaluate the latency reduction obtained by the distribution and, in general, the improvement obtained for all the other application KPIs (*TH*, *DT*, and *VT*) that depend on low latency in a more or less significant way.

The distribution is relative to a portion of the single applications (and content) that require the target latency. However, even if the deployment were complete, the connection to the centralized cloud is always necessary.

We denote by the hit-ratio *HR* the fraction of the application distributed near A and B (by assumption we have only one data stream). For *n* flows, the approximate *HR* value is obtained with the average made for the different application types of the various *HR*s of the individual applications, weighted with the peak *TH* values of the individual applications. A more precise value is obtained by considering different values for *HR* and *TH*. Therefore, a more precise evaluation requires working on the different service families and is more complex.

We call RTT_{eccA} the latency between the site where we install the ECC micro/mini data center and the destination site A (ECC A). We only consider one site since the same evaluations apply to the other site. RTT_{eccA} is never equal to 0 since the deployment requires processing activities for traffic classification, steering, and running part of the ECC applications.

We now introduce the total latency RTT_{q} between the ECC A (close to site A) and site A obtained by distributing the *HR* fraction of the application in ECC A.

To meet the latency requirement, we must have $RTT_{\text{q}} \leq RTT_{\text{target}}$.

RTT_{q} includes the optical transmission and the processing latencies (processing latency is related only to ECC A, as processing latency in node A is considered in RTT_{node}). We have $RTT_{\text{q}} = RTT_{\text{tot}}/SU$, where *SU*, the network's Speed Up [6], is a function of the congestion control algorithm. The following is straightforwardly obtained under the condition provided in [10]:

$$RTT_{\text{q}} = RTT_{\text{tot}}/[HR\cdot(RTT_{\text{tot}} / RTT_{\text{eccA}} -1) +1] \quad (6)$$

We obtain different relationships for *SU* with other congestion control mechanisms for TCP (e.g., cubic and BBR) and others for UDP, such as QUIC.

The limit values for RTT_{q}, obtained by the Mathis *et al.* model [10], are with:

- $HR = 0$ (no application distribution), which gives the maximum value $RTT_{\text{q}} = RTT_{\text{tot}}$;
- $HR = 1$ (limiting case in which the application is fully distributed and thus runs only on ECC and not on the centralized cloud), which gives the minimum value $RTT_{\text{q}} = RTT_{\text{eccA}}$.

Finally, the following Figure A-1 shows RTT_{q} as a function of the hit ratio, *HR*, and the round-trip time with ECC A.

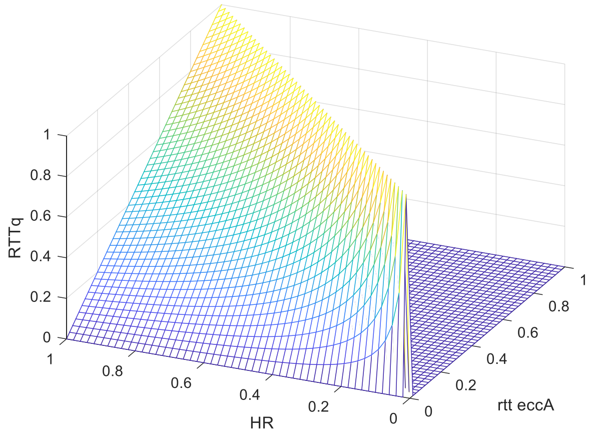


Fig. A-1 – Values of $RTT_{\text{q}} \leq 1$ ms as a function of *HR* and RTT_{eccA}.

AUTHORS

Gianfranco Ciccarella received a university degree (cum laude) in electrical engineering. He has more than 35 years of experience in telecommunications, working in the Telecom Italia Group Companies with growing technical and managerial responsibilities. Since January 2016, he has started a consultancy activity on TLC strategy, IP services and networks. In the Telecom Italia Group, he has held positions and responsibilities, including Corporate CTO for Telecom Argentina and TIM Brazil Companies, vice president for Next Generation Access Networks (NGANs) and Partnerships at the Strategy Department of Telecom Italia, and Telecom Italia Sparkle CTO and CIO with responsibilities to drive the design, deployment, and operations of the Telecom Italia International Wholesale Services and Network. He was the director of the Post Graduate Training and Technical Department, "Scuola Superiore Guglielmo Reiss Romoli," L'Aquila, Italy, did research and teaching at the EE Department, University of L'Aquila, Italy (where he became an associate professor), and was an adjunct associate professor at New York Polytechnic University, NY, USA. He was a member of the board of directors, the CEO, and the chairman for a number of companies within the Telecom Italia Group and a member of the Akamai Advisory Board. He has authored several articles and was invited many times to deliver speeches to international conferences and round tables.

Romeo Giuliano received a Ph.D. in telecommunications and micro-electronic engineering in 2004, from the University of Rome Tor Vergata, Rome, Italy. He is currently an associate professor at the University of Guglielmo Marconi (USGM, www.unimarconi.it). He has been involved in several European research projects within FP5, FP6, FP7, H2020, on topics such as Ultra-Wideband (UWB) applications, 4G wireless technologies for aeronautical applications, development of waveforms for air vehicles and remote control of trains. His main research topics are: 5G, B5G and 6G wireless systems, the fixed access unbundling and fiber optic network (e.g. VDSL, G.fast), IoT systems and C-V2X, indoor and outdoor positioning systems and machine learning on communication systems.

Other research interests: UWB technology, Unmanned Aerial Vehicles (UAVs/drones), remote train control. He is author of about 140 papers in international journals and conferences.

Francesco Vatalaro is a full professor in telecommunications at the University of Rome Tor Vergata, Rome, Italy, and holds more than 40 years of experience in the information and communication technology sector as an academic, consultant, and in the industry. He graduated with full marks in electrical engineering from Bologna University in 1977. After an industrial experience, in 1987, he joined the university. As founder in 2001 of the RadioLabs Research Center, Rome, Italy, he was President of the Center till 2008. He was a member of the Scientific Committee of Thales Alenia Space (France-Italy), President of the "Next Generation Network Committee" for the Italian authority for communications, President of the Italian section of the IEEE (Institute of Electrical and Electronics Engineers), and a member of the Strategic Committee of the IEEE Communications Society, Piscataway, New York. He is a Senior Life Member of the IEEE, and he is author or co-author of more than 200 scientific papers and several international patents.

Franco Mazzenga received a PhD degree in telecommunications in 1997. Since 2006 he has been an associate professor of telecommunications in the Department of Enterprise Engineering at the same university. Since 2001 he has been CTO of RadioLabs (www.radiolabs.it) and from 2012 a member of the board of directors. He is the author of 150 scientific papers in international journals and conferences. His research interests are in 5G/6G wireless and wired communication technologies. He has been involved in several European projects.

Alessandro Vizzarri. Alessandro Vizzarri received a degree in electronic engineering from the University of Bologna, a PhD in telecommunication engineering and microelectronics from the University of Rome Tor Vergata and post-lauream diploma from the Superior Institute of Communication (Rome). He is an adjunct professor of "Radio Spectrum Management" at the University of Rome Tor Vergata. He worked as a consultant engineer and manager for vendor, manufacturing and consultancy companies operating in the ICT market, in particular wireless networks (UMTS/HSPA, LTE, WiFi), broadcast networks (DVB-S/T/H), digital transmissions, multimedia systems (IMS), and project management. He has been a manager of several international projects related to ICT technologies. His research interests include wireless and broadcast networks, radio network planning, QoS and QoE, interoperability, regulation and standardization.