Contreras, L. M., Solano, A., Cano, F. & Folgueira, J. (28 June-2 July 2021). *Efficiency gains due to network function sharing in CDN-as-a-Service slicing scenarios* [proceedings]. 2021 IEEE 7th International Conference on Network Softwarization (NetSoft), Tokyo, Japan.

DOI: 10.1109/NetSoft51509.2021.9492721

Efficiency Gains due to Network Function Sharing in CDN-as-a-Service Slicing Scenarios

Luis M. Contreras*, Alberto Solano*, Francisco Cano†, Jesus Folgueira*
*Transport and IP Networks, Telefónica I+D / CTIO, Madrid, Spain
†Video, Telefónica I+D / CCDO, Granada, Spain
{luismiguel.contrerasmurillo, alberto.solanorodriguez, franciscojose.canohila, jesus.folgueira}@telefonica.com

Abstract— **The consumption of video contents is currently dominating the traffic observed in ISP networks. The distribution of that content is usually performed leveraging on CDN caches storing and delivering multimedia. The advent of virtualization is bringing attention to the CDN as use case for virtualizing the cache function. In parallel, there is a trend on sharing network infrastructures as a way of reducing deployment costs by ISPs. Then, an interesting scenario emerges when considering the possibility of sharing virtualized cache functions among ISPs sharing a common physical infrastructure, mostly considering that usually those ISPs offer similar content catalogues to final end users. This paper investigates through simulations the potential efficiencies that can be achieved when sharing a virtual cache function if compared to the classical approach of independent virtual caches operated per ISP.**

Keywords—CDN, VNF sharing, CDNaaS, efficiency

I. INTRODUCTION

Video traffic is becoming nowadays the killer application for service providers' networks, and it will be probably the dominant component of the overall traffic in the future. The raise of multiple offers from multiple video platforms, either directly owned by Internet Service Providers (ISPs) or offered by Over-The-Top (OTT) content providers, such as Netflix, Amazon Prime, HBO, etc., is effectively changing the network demand landscape. Being this fact already true for streaming content, it will be even increased when considering in the near future other flavors of multimedia delivery, such as gaming or virtual reality.

According to analysis from the telco industry [1], video traffic in 2020 represents the 66% of all mobile data traffic, with the perspective of increasing up to 77% in 2026. The same occurs for fixed networks, where e.g. TVs generate the largest component of the traffic per device in Western Europe [2], with residential users moving from offline to online activity, in parallel with the improvement in video formats, especially the Ultra-High-Definition (UHD) or 4K. Some estimations consider that the number of installed flat-panel TV supporting UHD will raise from 33% in 2018 to 66% by 2023 [3], then favoring the demand of higher resolution on-line contents.

Such a huge amount of content is served leveraging on overlay Content Delivery Networks (CDNs), which allow servicing contents in a scalable manner. A number of distributed delivery points or caches store the content and deliver copies of it locally, alleviating the demand in terms of capacity needed in the transport network, since that content would be required to obtain remotely, otherwise. Those caches can be found at the border of the ISP networks or even internal to them. The latter is the current trend, in where multiple caches from different content providers (i.e., from the own ISP but also from third parties) are deployed internally to the network, delivering the content in proximity producing the reduction of bandwidth at higher layers in the network topology, but also the perceived latency, leading to which is named as sub-millisecond Internet [4].

The advent of network virtualization has brought the attention on the possibility of considering the CDNs and the caching of content as a relevant use case. For instance, it can permit the dynamic deployment and flexible instantiation of caches in the network on top of virtualized infrastructures. Thus, both Network Function Virtualization (NFV) and Multi-access Edge Computing (MEC) paradigms have look at the CDNs in their specifications [5][6]. Furthermore, traditional CDN providers have also moved into the virtualization arena by providing virtualized solutions of their CDN solutions like Akamai [7] or Amazon [8].

Thinking on the OTT content providers, the aforementioned trend of increase on the video consumption implies that subscribers from different ISPs in a given geographical area practically consume the same kind of content from a reduced number of content providers, if not actually the same, independently of the ISPs they are subscribed to.

On the other hand, it is becoming common the fact of sharing infrastructures among service providers [9][10][11]. This is due to the need of reducing investments for improving margins. Such a sharing implies to host services of competitor ISPs on top of a single and common infrastructure, and/or hosting directly Virtual Network Operators (VNOs) not having any infrastructure at all, or very limited.

The sharing scenario, as mentioned, can be implemented by one ISP sharing its infrastructure to others, or by neutral operators opening their infrastructures to third party ISPs. Such kind of operators are commonly known as Infrastructure Providers (InPs).

From an operational perspective, the effective way of implementing the sharing of infrastructures is expected to be through the adoption of network slicing [12][13], allowing the recreation of a virtually dedicated network for each of the ISPs or VNOs in the area, while using a common physical infrastructure (we will just refer from now on to ISPs in a general way for simplicity). Then, a network slice per ISP can contain all the services offered for their respective subscribers, including the necessary caches for providing video contents. This helps to reduce the traffic and associated costs in the transport network.

The network functions in the allocated slice, like the caches referred before, will be in the form of Virtual Network Functions (VNFs). In principle, those VNFs can be instantiated separately, dedicated per ISP. However, when looking at video cache function itself, thinking on the fact that most of the OTTs will be common to all the ISPs supported by a certain InP, and also considering that the content offering itself will be also the same, the necessity of having multiple instances of the same function instead of sharing a single one can be questioned. Reasons for this are the potential benefits and efficiencies that can be achieved in terms of consumed compute and storage resources, as well as others like energy efficiency.

Fortunately, NFV specifications consider the possibility of sharing VNFs among virtualized services [14], then open the door to optimize the deployment and usage of network functions, that in this particular case implies leveraging on the same virtualized cache instance for all the ISPs.

The motivation of this paper is to answer the question on what can be the achievable efficiency if following the approach of cache sharing. In order to answer that question we perform simulations by observing the number of contents consumed for a base of subscribers of distinct ISPs in different scenarios. This permits to understand on which conditions this can be desirable and derive some deployment guidelines.

The paper is structured as follows. Section II presents related work for each of the aspects this paper addresses, namely the integration of CDN and ISPs, the on-demand instantiation of CDN cache end points, and the sharing of VNFs in network slices. Section III describes the network scenario under evaluation, for the comparison of the shared vs non-shared virtual cache approach. The simulation framework used for analyzing the case of interest is introduced in Section IV. Next, Section V performs an analysis of the achieved efficiencies, discussing the obtained results. Section VI realizes assessment of the solution from an economic point of view. Finally, Section VII provides conclusions and proposes future lines of work.

II. Related work

This work is a confluence of several trends in the industry, namely the integration of CDNs on service providers' networks, the provision of virtualized CDN capabilities in a virtualized on-demand manner, and the sharing of virtual network functions on network slicing.

This section provides an overview of these areas of knowledge, providing state-of-the-art works for reference.

A. Integration of CDN on service providers' networks

CDNs have been for long a subject of research in the pursue of a tighter integration into ISP networks. This is mainly because CDNs permit the reduction of traffic volumes, especially for peering and transit layers. Different models of integration can be considered.

The most basic scenario is the one of interconnection between the ISP and the content provider through a peering or transit agreement [15]. With this approach the content enters the network in a higher layer in the topology. The main advantage of this scenario is basically shortening the transport paths from the origin of the content and stabilizing the delays on transmission [16]. Other alternative is the direct interconnection of caches, e.g. the ones from the OTT with the ones from an ISP, as proposed by initiatives like [17] or [18].

For caches being deployed in the internals of the ISP network, one model is the direct operation of the CDN by the ISP, who owns in consequence the CDN infrastructure together with the transport network, and then can provide harmonized control, network planning and operation of both. Analysis of the effectiveness of CDNs operated by ISPs can be found in [19], [20] and [21]. Reference [19] discusses the beneficial impacts on the reduction of traffic due to the deployment of caches and how characterizing the popularity of the content consumed in the network. In [20] the authors analyze the advantages of commonly planning the location of the CDN caches considering the topology of the network. Recent work in [21] proposes a further integration of the CDN and the ISP network by leveraging on BGP-LS [22] for advertising topological information and ALTO server [23] for automatically providing the CDN with up-to-date network and cost maps.

Finally, another model is given by the deployment on ISP networks of caches from OTTs (e.g., Netflix, Akamai, etc.) through bilateral agreements among the parties. As these services are quite popular, the advantage of doing so is the reduction of traffic internal to the network. Also, as described before, there is a notable reduction on the perceived delay [4]. For instance, reference [24] explores different collaboration schemas between ISPs and CDN providers in a virtualized environment. However, this other model of integration can become less effective since the logic for selecting the caches to deliver the content when a subscriber request is received is on the hands of the content provider, so that decision is usually not aware of the underlying network circumstances. This can provoke not optimal decision on some occasions impacting the perceived QoE [25].

B. CDN as a Service

The flexibility brought by NFV has generated the possibility of dynamically instantiating caches. This on-demand approach is generically termed in the literature as CDN as a Service (CDNaaS). Some previous works have concentrated in the possibility of virtualizing and dynamically instantiating CDN nodes in the cloud.

The authors in [26] propose an architecture for on-demand virtual CDN service instantiation that could be used by OTTs requesting resources dynamically on top of a telco cloud by means APIs. Similarly reference [27] describes a CDNaaS platform enabling the dynamic creation of slices for CDNs spanning multiple cloud domains and analyzing different strategies of placement for the virtualized cache nodes. This is extended to the concept of slicing in [28].

The work in [29] provides a mathematical framework to calculate the success rate of different tiers of virtualized CDNs taking into account the QoE of the subscribers and the resources used by the virtualized cache nodes. In [30] the authors propose a Big Data architecture fulfilling the ETSI NFV guidelines, allowing to control the virtualized components of a cloud-based CDN for minimizing the CDN costs while ensuring the highest quality on the service delivery.

C. VNF sharing

Network services are usually deployed as a concatenation or chain of network functions. When going in a virtualized manner, services are represented for instantiation as a graph of interacting and interconnected VNFs. Potentially, same kind of functions could be part of different network services. In that situations, common VNF instances could be potentially shared among services.

VNF sharing has been studied in the literature mainly from the perspective of the service deployment and placement, targeting the optimization of the network embedding problem,

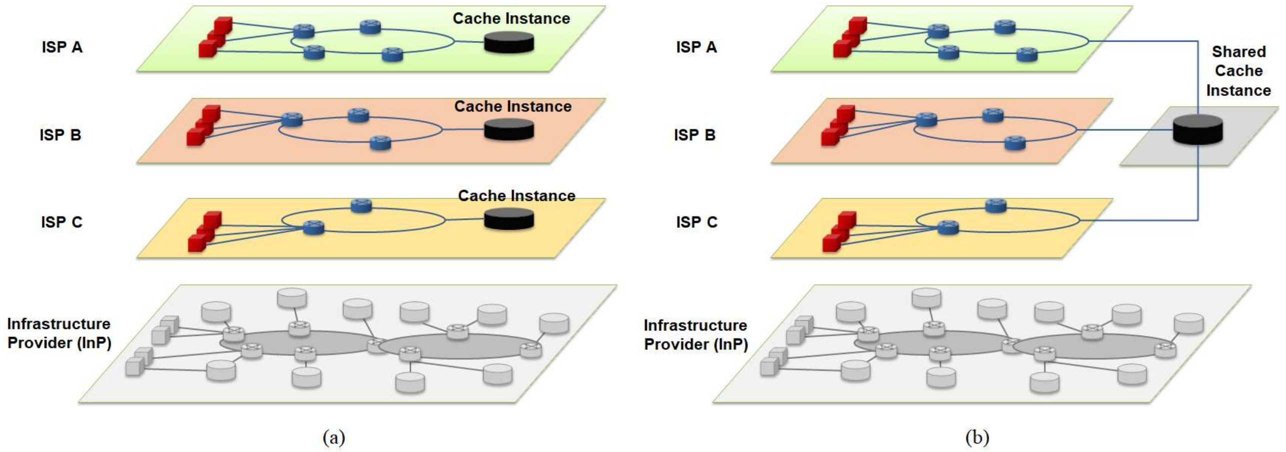

Fig. 1. Scenarios of comparison: (a) dedicated cache instances per ISP vs (b) shared cache instance

usually looking for the higher ratio of physical resource utilization and cost reduction, at the time of honoring more service requests, as studied in [31][32][33].

D. Summary

The related work overviewed here shows an important background in the exploration of mechanisms for improving the efficiency on the usage of resources in operational ISP networks. With the CDN case, the networking resources can be optimized through the reduction of the traffic delivered in the network. In addition to that, through the sharing of VNFs, compute resources can be optimized, as well.

The virtualization paradigm permits to deploy CDN caches in a virtualized fashion, as VNFs. Considering the fact that popular contents are concentrated on the hands of few content providers, ISPs deploying network slices on top of the facilities of the same InP could potentially share the VNF instances implementing such caches. Thus, it is possible to conjugate both the efficiency due to the deployment of CDNs and the efficiencies due to the sharing of caches.

In the following sections we quantify such efficiencies through the analysis of a variety of scenarios simulating diverse demands under different conditions.

III. Network scenario under evaluation

Figure 1 illustrates the scenario under evaluation. It is assumed that different ISPs make use of a common infrastructure provided by an InP. Note that one of those ISPs could play also the role of InP in a given geographical area.

It is also assumed that those ISPs in such area served a common set of contents, provided e.g. by an OTT that maintains separated commercial agreements with all those ISPs. Thus, the ISPs essentially have a similar content offer for all the base of end-users in the area, achieving service differentiation by other means (e.g., price competition, bundle of offers, etc.)

The OTT initially performs the content distribution by means of virtual caches (vCaches), in the form of VNFs, deployed on top of the InP infrastructure. Each vCache will store the contents demanded by the customer base of each of the ISPs in that particular area. When considering a separate instances of virtual cache per ISP, this means that each vCache will be dedicated for a given ISP.

However, since the catalogue of contents offered by the OTT are the same for all the set of ISPs, in principle it is possible to deliver all the demanded contents from a common virtual cache instance just managing the distinct subscriptions differentiating the delivery of the content for each of the ISPs, e.g. by using different Ethernet vlans in the connection to each network slice per ISP. That is, a certain content “Content-A”, instead of being replicated through different virtual cache instances (one per ISP) is contained in a single virtual cache instance but being delivered to the corresponding ISPs demanding such content.

This is relevant since the efficiency of caching increases with the number of end users being served, since there will be higher coincidence in the kind of content demanded by the users. This is evident for the more popular contents, but also happens for the less popular ones. Thus, sharing the vCache has the effect of concentrating the end-user demands, which should reduce the overall number of contents to be cached in an area.

IV. Simulation framework

In order to understand to what extent we can achieve efficiency when following the approach of cache sharing, we perform some simulations based on the number of contents consumed for a base of subscribers of distinct ISPs in different scenarios. This permits to understand on which conditions this can be desirable. A number of metrics are considered for the comparison, based on the different usage of resources in the two scenarios.

We will consider a network divided in areas as the ones represented in Fig. 1. The objective of the simulation is to understand what the amount of contents demanded in the network is, and how that number impacts in shared vs non-shared vCache scenario.

A. Content popularity model

The preference of content visualization by end users in IPTV, Video-on-Demand (VoD) and cache systems in general is commonly modeled by a power-law distribution known as Zipf function [34][35][36]. The Zipf function states that the occurrence of a certain event (here, the tuning of a multicast channel for IPTV or the selection of a certain unicast content in a VoD system) is determined by:

$$k\frac{1}{x^{\alpha}} \tag{1}$$

where k is a constant value, x the rank or popularity of the event in the distribution, and α the factor which characterizes the skewness of the distribution. Then, the frequency or probability that predicts the eligibility of an event is provided by:

$$\frac{\frac{1}{x^{\alpha}}}{\sum_{n=1}^{N}\left(\frac{1}{n^{\alpha}}\right)} \tag{2}$$

Where N is the total number of ranked elements. As α increases, the popularity of the first ranked events increases, while the distribution tail concentrates less occurrences. Here we will consider 0,6 and 0,9 as reference values for α in line with the observations in [36].

B. Number and size of contents

The amount of available on-demand contents to be consumed either live or in an on-demand fashion has continuously increased along the time. Even if such increase applies to both types of content, the order of scale differs. In the case of live content, usual values nowadays could stay around few hundreds of contents. For the VoD case, the quantity considered can be even higher than ten thousand.

Several factors have contributed to this. On one hand, the proliferation of OTT video platforms have augmented significantly the number of contents in their respective catalogues, generating a very broad multimedia offer. Secondly, it usually occurs that the same multimedia content is coded differently (e.g., Smooth Streaming, DASH, HLS, etc.) adapting it to multiple receiver platforms and players, then creating differentiated copies of the same content which are consumed also differently depending on the acceptance of a given player.

There are also differences between live and VoD contents regarding their lifetime. Live content usually is stored in the cache during few hours, as much, since further than that time the content can be considered no longer live. This time in the cache allows users accessing late to the content, but yet interested in a recent event, to be served. Once that time is exceeded, the profile in the consumption of that content can be considered as passing to the category of on-demand.

The on-demand content can usually stay in the caches for longer, typically up to the time that the capacity of the cache is exhausted, then requiring to make storage space available for newer content being demanded.

For the analysis in this paper we will generically consider on-demand content as the subject of interest for the end-users. In addition to that, a common and unique coding of the content will be assumed for simplification. Here, a typical content will be considered to be coded for an average bit rate of 5 Mbps and an average duration of 80 minutes, requiring then ~ 2,8 GB of storage in the vCache (as observed in an actual commercial ISP network).

C. ISPs and end-users

In open, competitive markets, the base of end-users will be divided among different competing ISPs addressing such market. The distribution is usually unbalanced, with some ISPs capturing higher share of users than others. Reasons for differentiation can be multiple, such as overall service offering pricing, etc.

For the analysis we will consider the presence of four ISPs with different market shares, as follows:

- Service Provider A: 40%
- Service Provider B: 30%
- Service Provider C: 20%
- Service Provider D: 10%

This differentiation represents a market where a dominant ISP is taken the majority of the share, with two major competitor ISPs as followers plus one challenger entering the market. The specific market share can differ in real scenarios but the assumption here permits to compare at different granular levels between the ISPs as a function of their relative share in a market. As reference, during Q1 2020 the share in Spain of the four main operators in the country was 38,3%, 25,2%, 20,6% and 10,3% for fixed broadband, and 29,6%, 24,4%, 22,4% and 13,9% for mobile access, respectively.

D. Number of vCaches in the network

The number of locations where to deploy vCaches is another axis of dimensioning. In the context of this analysis, it can be understood as the number of Points of Presence (PoPs), Central Offices (COs) or edge nodes (thinking on a more distributed deployment with high capillarity) where a vCache could be instantiated. It is assumed that on each of those locations there is sufficient compute infrastructure to host the vCache in terms of processing capacity, storage, etc.

The number of locations depends on the size of the country to be served, the distribution of the population (i.e., density) and the availability of physical infrastructures (sites, transmission, etc). The selected locations do not necessary need to be at the same hierarchical level in a layered network topology, that is, some of them could be considered at the edge of the network while others being more centralized. The criteria followed in this analysis will consider simply the total number of users being served by each vCache, where such total number will be divided among the ISPs as described before.

Here we will explore the behaviour of distributing the users in a number of locations with vCache ranging from 100 to 1000. The former can represent the number of PoPs concentrating main distribution areas (at region/province level) in a mid-size country, while the latter can represent the number of central offices in such a country. This is compatible, for instance, with the number of aggregation (the former) and pre-aggregation (the latter) sites considered in other reference networks like in [37]. The results however can be extrapolated to any other number of vCaches present in the network.

E. Modelling the assignment of end-users to desired content, ISPs and vCaches

For each of the simulated preference of end users, the simulation firstly identifies the content desired by the user, then associates that end user to one of the ISPs, and finally assigns him/her to one of the vCaches. With that, once the modelling is finished, it is possible to quantify how many different contents are stored per vCache. That quantification essentially considers that any solicited content is stored in the vCache with the expectation of being served from the cache for a second or higher request. In this way, we are modelling

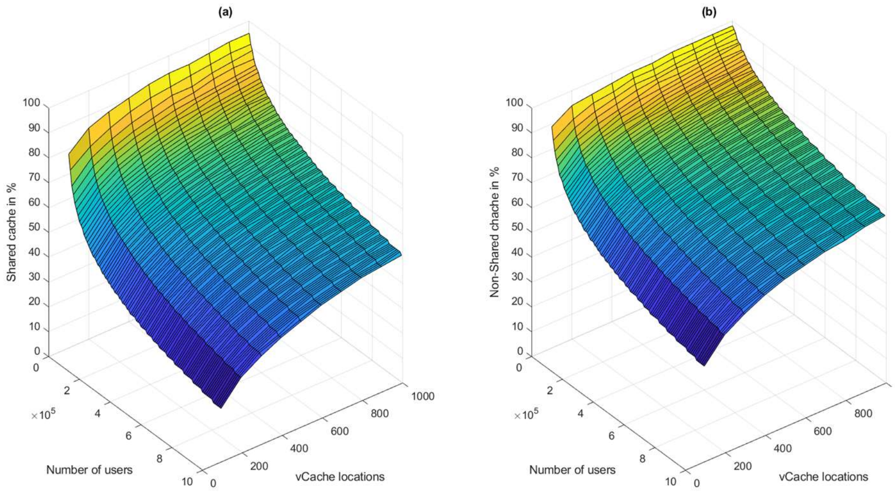

Fig. 2. Percentage of the total number of cached contents in shared (a) vs non shared (b) approaches with respect to the number of end users (10000 to 1000000), for α = 0,9 and 3000 selectable contents as a function of the number of vCache locations (100 to 1000)

what would result from a peak demand in the aforementioned conditions. This procedure generates two views: *(i)* the view of the solicited unique contents that would be generated in case each of the ISPs maintain separated vCaches, and *(ii)* the view of the solicited unique contents if the ISPs share the vCache. From the comparison of both views, conclusions about the efficiency of the shared cache approach can be obtained.

For the selection of the content, a content is identified according to the Zipf distribution as defined in sub-section IV.A. Then, the end user is associated to an ISP according to the market share percentages specified in IV.C. Finally, that user is assigned to a vCache node. The simulation, as first approximation to the problem, will assume a uniform distribution of end users among the vCaches. This implies that, roughly, the same number of end users is considered per vCache.

F. *Description and parametrization of the simulation*

In order to build up the simulations we have used the multi-paradigm numerical computing development environment and proprietary programming language MATLAB, running on a server counting on two 2.20GHz vCPUs, 16GB RAM and 200GB storage capacity.

Table I summarizes the parameters of the simulation. Each simulated scenario run on average for 2 hours and a half, generating 5MB of data. Twenty scenarios were run, with minor deviations among runs, confirming the validity of the obtained results. The results here presented are based on the average values of one of that runs.

Caching contents produces a clearer advantage in the reduction of traffic in the networks, from the peering and transit points up to the locations of the vCaches. However, such efficiency comes at the cost of deploying different levels of computing infrastructure in the network. In order to measure that trade-off, we define the following ratio R as

$$R = \frac{\#\, stored\ objects}{\#\, end\ users\ requesting\ contents} \qquad (3)$$

The lower R, less storage is needed for serving the same amount of end users at a given instant. Fig. 2 shows the results of the simulation, showing the percentage of the total number of cached contents with respect to the number of end users. In general terms, intuitively, the better R is achieved when the larger the number of end users and the lower the number of locations. Essentially, concentrating the contents in few locations reduce the need of having multiple replicas of the same content, mainly the popular ones. The counterpart of this approach is that having fewer locations for vCaches imposes certain degree of centralization, then implying more usage of networking resources for distributing the traffic from the vCache locations towards the end users.

TABLE I. PARAMETERS OF THE SIMULATION

Parameter	Values	Description
Skew factor (α)	0,6 and 0,9	Power-law factor of the Zipf distribution of content selection
Number of contents	[500, 5000]	Number of selectable contents in the scenario under evaluation
Number of end users	[10000, 1000000]	Population of users simultaneously demanding content
Number of ISPs	4	The ISPs have the following market share over the base of users, respectively: 40%, 30%, 20% and 10%
Number of locations	[100, 1000]	Locations where vCaches are deployed

It can be also observed that following the shared approach clearly improves the ratio of stored content in the network, that is, for the same number of locations and end users requesting contents, less storage is needed. This is also intuitive in the sense that concentrating the demands from the different ISPs, the number of replicas for the less popular content becomes also reduced.

The trade-off between networking and compute/storage resources for the overall network design with the shared and non-shared vCache approaches is left for further study. We now focus on understanding the particular efficiencies achieved when sharing vCache VNFs among all the ISPs versus maintaining separated VNFs per ISP in the network.

V. EFFICIENCY ANALYSIS OF VCACHE VNF SHARING

The following sub-sections analyze the impact of the distinct factors in the simulation from the perspective of

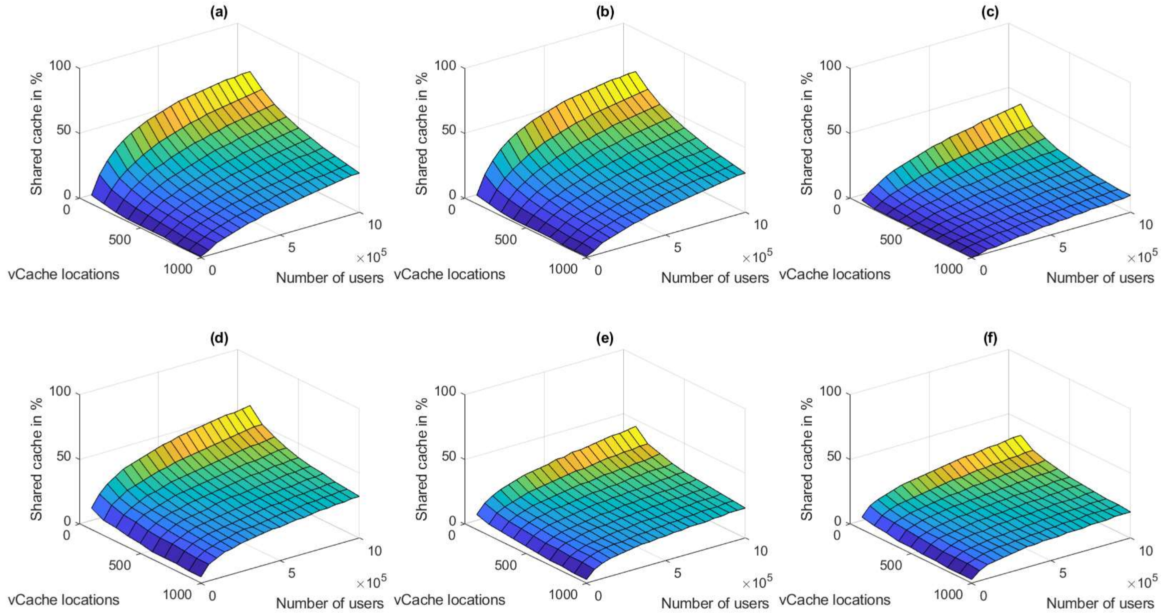

Fig. 3. Efficiency E (in percentage) in terms of the average contents cached per location considering 100 to 1000 vCaches, and 10000 to 1000000 users uniformily distributed among the vCaches. Graphs show the results when α = 0,6 for 1000 (a), 3000 (b) and 5000 (c) contents, and similarly when α = 0,9 for 1000 (d), 3000 (e) and 5000 (f) contents

shared VNFs performing the caching of contents. In this respect, it can be considered the achieved efficiency E as the ratio between the average stored contents per vCache in the both the shared vs non-shared network scenarios. Fig. 3 provides an overview of E when varying the number of contents offered, the user preferences on that contents, the number of locations where the vCaches are instantiated, and the number of end users simultaneously requesting contents.

A. *Impact of the content offer*

In general terms, as the content offer increases, i.e. as more contents are available to the end users, the efficiency gain of the shared cache approach diminishes. This is due to the fact that the more contents are selectable, the higher the dispersion of the chosen contents is. Thus, the coincidence of election of contents among ISPs is also reduced, which implies that in shared vCache more individual contents need to be stored.

B. *Impact of user preferences*

The end user preferences can be more or less disperse. The higher the dispersion, the larger the number of individual content selected.

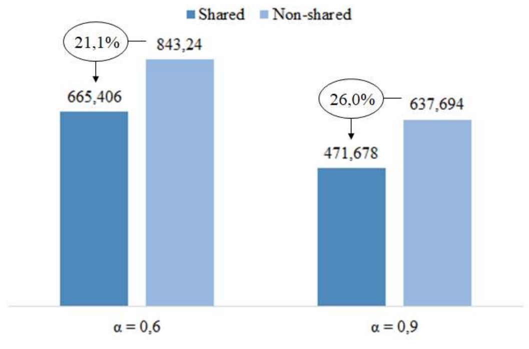

Fig. 4. Comparison of average number of contents in the shared and non-shared vCache scenarios for 500000 users uniformally distributed in 500 locations with 2000 selectable contents

In Fig. 3 the effect of the dispersion can be observed by comparing the charts with different skew factor. When the selection of content is more concentrated (i.e., higher value of α, or less variance of the distribution), the efficiency becomes higher for the shared vCache case.

As an example, Fig. 4 represents the efficiency achieved by the shared vCache for 500000 users distributed across 500 locations when 2000 contents are available with both α = 0,6 and α = 0.9. The absolute difference in terms of average cached contents are 177,8 and 166 respectively. Despite the number of contents decreases in absolute value, the relative efficiency increases with greater values of α. That is, the shared approach is better when the preference in the selection of the content is less dispersed.

C. *Impact of the number of locations*

As long as the number of vCache locations increases, the efficiency of the shared approach decreases. Two effects can be considered here. On the one hand, distributing end users among more vCaches implies the replication of the more popular contents in all the caches. Furthermore, the gain on the less popular contents obtained when concentrating them in less locations is diluted, also provoking the need of storing more objects across the network for those less demanded contents. Fig. 5 shows graphically that trend.

As observed, when the preferences of the users are more dispersed (α = 0,6) the gain is severely reduced especially for high number of locations. While the percentage of efficiency is similar for low number of locations (e.g., 32,9% when α = 0,9 and 32,6% when α = 0,6 for 100 vCaches), it becomes lessened for high number of them (e.g., 17,6% when α = 0,9 and 9,0% when α = 0,6 for 1000 nodes).

D. *Impact of the number of end users*

As the number of users increases, the number of requested contents also increases. Due to the different popularity of the contents, the number of contents will not grow at the same pace, since certain contents will be already cached. Fig. 6

shows this fact by illustrating the evolution in the number of cached contents for both shared and non-shared situations.

Interestingly, as the number of end users grows, the efficiency of the shared approach also increase. Clearly, the more popular contents in the shared approach will be cached only once instead of four times (one per each of the ISPs), clearly contributing to the reduction of cached contents. However, as the number of end users increases, also less popular contents will be subject of coincidence in the end users requests, thus contributing to the overall efficiency of the shared scheme.

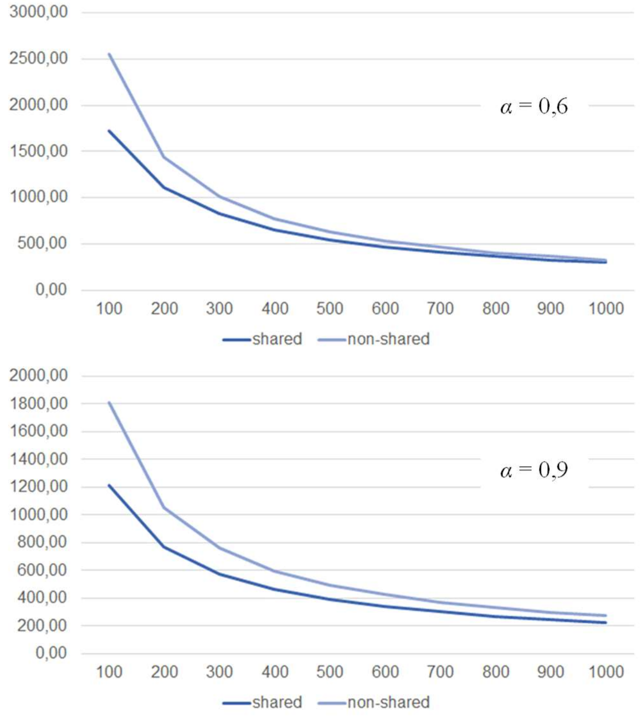

Fig. 5. Variation on the number of total cached contents when increasing of vCache locations for 3000 selectable contents for 350000 simultaneous end users

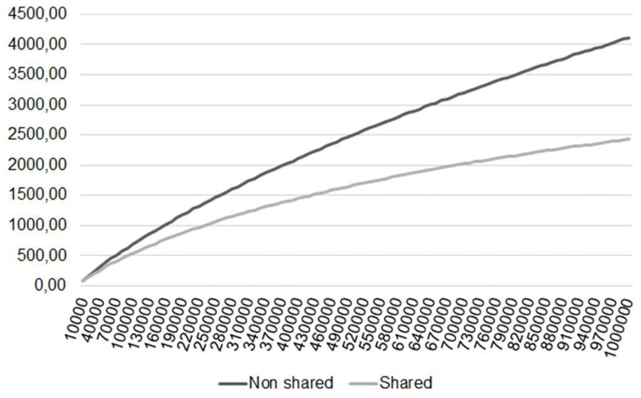

Fig. 6. Number of total cached contents as the number of end users grows from 10000 to 1000000 for shared and non shared approaches for $\alpha = 0.9$, 4000 selectable contents and 100 locations.

VI. Economic assessment

In order to perform an assessment of the economics of the shared approach we will consider the following cost function per individual vCache location.

$$Cost_{vCache} = \gamma + \rho \times \varepsilon \tag{4}$$

where Υ represents the costs associated to the instantiation of the vCache in terms of processing and volatile memory, ρ indicates the number of stored contents in the cache, and ε is the storage cost per content. For simplicity, all the vCaches are assumed to require similar CPU and RAM capacity, as well as all the contents are considered to have the same storage needs. For comparison, the total cost should consider the contribution to the cost of all the vCaches instantiated for satisfying the end users demand.

For the non-shared approach, the total cost will be the sum of the individual costs per ISP, i.e. four different VNFs each of them dimensioned to the specific needs of a particular ISP. In the case of the shared vCache, the cost will be the one of the shared VNF aggregating all the contents. How the split of costs is performed for each of the ISP in the latter case is out of scope of the paper, but it can be assumed that such split could be based on the actual number of content requests by the end users of each ISP, so proportional to the market share.

Fig. 7 presents one sample case of the average values of content stored per ISP, as well as the resulting stored contents when sharing the vCache.

To calculate the cost we apply a sensitivity analysis based on the relation among the unitary cost of the CPU and RAM per vCache, Υ, and the unitary cost of storage per content, ε. We then consider three scenarios

- Scenario 1: $\varepsilon = 0.01\ \Upsilon$
- Scenario 2: $\varepsilon = 0.1\ \Upsilon$
- Scenario 3: $\varepsilon = 0.5\ \Upsilon$

With that in mind, and assigning a unitary cost of 1 Cost Units [CU] to Υ, Table II presents a sample economic assessment considering 500000 end users accessing over 3000 selectable contents, and for 100, 500 and 1000 vCache locations. The assessment is calculated for the two skew factors in the simulation, 0.6 and 0.9 respectively.

The costs for the non-shared case is calculated as the sum of the individual costs per ISP, while in the case of the shared vCache, the cost is calculated considering a single vCache per node for all the ISPs. In order to estimate the costs per ISP the total cost of the shared case is divided proportionally to the assumed market share per ISP.

The table presents the overall [CUs] per solution as well as the percentage of savings for the shared vs non-shared cases. The calculation takes the absolute values for all the vCaches locations.

As can be observed from the analysis, there are important savings when the shared approach is followed. The savings are higher for lower values of ε (i.e. storage costs) mainly due to the fact that, when sharing, the overhead processing costs of the vCache are reduced from 4 times (one per ISP) just to 1 (the shared vCache). The processing costs impact more severely to the ISPs with lower market share, since that processing costs can be considered as fixed independently of the number of contents to be delivered.

Obviously, for a given number of locations, the total costs increase as ε grows, since the contribution of storage to the total cost increases as well. The costs also increase as the number of locations increases, essentially because both the

TABLE II. ECONOMIC ASSESSMENT (500000 END USERS, 3000 AVAILABLE CONTENTS, 100/500/1000 NODES)

		100			500			1000		
		0,01 x Υ	0,1 x Υ	0,5 x Υ	0,01 x Υ	0,1 x Υ	0,5 x Υ	0,01 x Υ	0,1 x Υ	0,5 x Υ
α = 0,6	***Non-Shared***	**3752,1**	**33921,2**	**168006,0**	**6388,6**	**45886,1**	**221430,5**	**8634,8**	**50347,6**	**235738,0**
	ISP-1	*1315,8*	*12258,0*	*60890,0*	*2195,6*	*17456,3*	*85281,5*	*2814,3*	*19143,4*	*91717,0*
	ISP-2	*1095,9*	*10059,3*	*49896,5*	*1812,4*	*13623,7*	*66118,5*	*2388,2*	*14881,6*	*70408,0*
	ISP-3	*830,7*	*7406,8*	*36634,0*	*1406,2*	*9561,8*	*45809,0*	*1946,7*	*10466,7*	*48333,5*
	ISP-4	*509,7*	*4197,1*	*20585,5*	*974,4*	*5244,3*	*24221,5*	*1485,6*	*5855,9*	*25279,5*
	Shared	**2161,5**	**20714,6**	**103173,0**	**4152,7**	**37027,4**	**183137,0**	**5118,2**	**42182,1**	**206910,5**
	ISP-1 (40%)	*864,6*	*8285,8*	*41269,2*	*1661,1*	*14811,0*	*73254,8*	*2047,3*	*16872,8*	*82764,2*
	ISP-2 (30%)	*648,4*	*6214,4*	*30951,9*	*1245,8*	*11108,2*	*54941,1*	*1535,5*	*12654,6*	*62073,2*
	ISP-3 (20%)	*432,3*	*4142,9*	*20634,6*	*830,5*	*7405,5*	*36627,4*	*1023,6*	*8436,4*	*41382,1*
	ISP-4 (10%)	*216,1*	*2071,5*	*10317,3*	*415,3*	*3702,7*	*18313,7*	*511,8*	*4218,2*	*20691,1*
	Sh* vs *N-Sh	**42,4%**	**38,9%**	**38,6%**	**35,0%**	**19,3%**	**17,3%**	**40,7%**	**16,2%**	**12,2%**
	ISP-1	*34,3%*	*32,4%*	*32,2%*	*24,3%*	*15,2%*	*14,1%*	*27,3%*	*11,9%*	*9,8%*
	ISP-2	*40,8%*	*38,2%*	*38,0%*	*31,3%*	*18,5%*	*16,9%*	*35,7%*	*15,0%*	*11,8%*
	ISP-3	*48,0%*	*44,1%*	*43,7%*	*40,9%*	*22,6%*	*20,0%*	*47,4%*	*19,4%*	*14,4%*
	ISP-4	*57,6%*	*50,6%*	*49,9%*	*57,4%*	*29,4%*	*24,4%*	*65,5%*	*28,0%*	*18,2%*
α = 0,9	***Non-Shared***	**2751,3**	**23913,4**	**117967,0**	**5354,0**	**35539,9**	**169699,5**	**7731,0**	**41309,5**	**190547,5**
	ISP-1	*945,2*	*8551,5*	*42357,5*	*1757,0*	*13069,8*	*63349,0*	*2416,3*	*15163,2*	*71816,0*
	ISP-2	*793,4*	*7033,5*	*34767,5*	*1496,5*	*10464,8*	*50324,0*	*2108,1*	*12081,0*	*56405,0*
	ISP-3	*612,4*	*5223,7*	*25718,5*	*1208,9*	*7588,8*	*35944,0*	*1783,7*	*8837,1*	*40185,5*
	ISP-4	*400,5*	*3104,7*	*15123,5*	*891,7*	*4416,5*	*20082,5*	*1422,8*	*5228,2*	*22141,0*
	Shared	**1593,1**	**15030,9**	**74754,5**	**3076,0**	**26259,6**	**129298,0**	**4011,7**	**31116,7**	**151583,5**
	ISP-1 (40%)	*637,2*	*6012,4*	*29901,8*	*1230,4*	*10503,8*	*51719,2*	*1604,7*	*12446,7*	*60633,4*
	ISP-2 (30%)	*477,9*	*4509,3*	*22426,4*	*922,8*	*7877,9*	*38789,4*	*1203,5*	*9335,0*	*45475,1*
	ISP-3 (20%)	*318,6*	*3006,2*	*14950,9*	*615,2*	*5251,9*	*25859,6*	*802,3*	*6223,3*	*30316,7*
	ISP-4 (10%)	*159,3*	*1503,1*	*7475,5*	*307,6*	*2626,0*	*12929,8*	*401,2*	*3111,7*	*15158,4*
	Sh* vs *N-Sh	**42,1%**	**37,1%**	**36,6%**	**42,5%**	**26,1%**	**23,8%**	**48,1%**	**24,7%**	**20,4%**
	ISP-1	*32,6%*	*29,7%*	*29,4%*	*30,0%*	*19,6%*	*18,4%*	*33,6%*	*17,9%*	*15,6%*
	ISP-2	*39,8%*	*35,9%*	*35,5%*	*38,3%*	*24,7%*	*22,9%*	*42,9%*	*22,7%*	*19,4%*
	ISP-3	*48,0%*	*42,5%*	*41,9%*	*49,1%*	*30,8%*	*28,1%*	*55,0%*	*29,6%*	*24,6%*
	ISP-4	*60,2%*	*51,6%*	*50,6%*	*65,5%*	*40,5%*	*35,6%*	*71,8%*	*40,5%*	*31,5%*

processing costs of the vCache (one per location) and the number of contents stored in the network increase, as well.

The impact of the skew factor does not vary too much the relative savings in percentage, with better behavior as α increases, but not significant. In absolute terms, the higher the concentration of user preferences (i.e., higher α), the lower the overall cost.

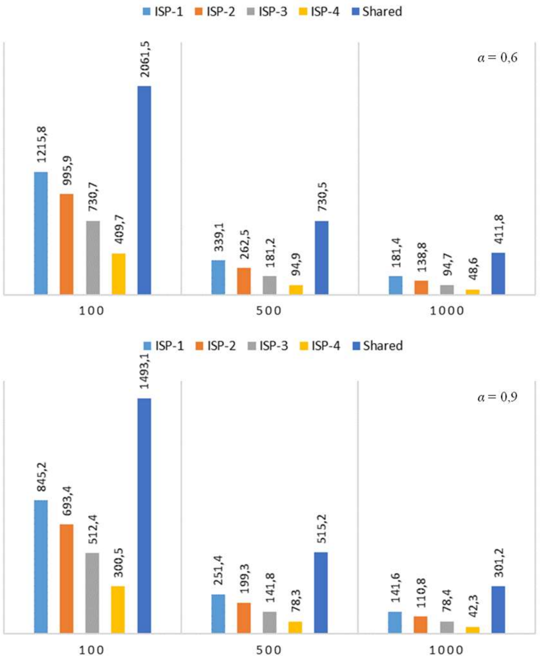

Fig. 7. Total average cached contents per individual ISP and when sharing the vCache for 100, 500 and 1000 locations, considering 500000 end users and 3000 available contents.

When looking at the impact on the ISPs, the most benefitted ISPs of following the shared approach are the ones with lower market share, showing very important savings. As the weight of the number of contents in the final cost increases, the savings are reduced. This is because the fixed costs of the vCache processing are diluted when summed up with the storage costs. This is more obvious on the ISPs with higher market share, since they contribute more to the total number of contents requested in the network.

VII. CONCLUSIONS AND FURTHER WORK

Video content is, and will be for long, the major component of traffic in existing ISP networks. Most of that content is provided nowadays by means of CDNs operated by a limited number of OTTs, then resulting in common content catalogues consumed by end users of different ISPs. The introduction of virtualization mechanisms are incentivizing, on one hand, propositions for virtualizing caches to deliver the streaming content, and in the other hand, facilitating the sharing of network infrastructures. Here we have explored the intersection of both comparing by simulation a network scenario where multiple ISPs shared virtualized caches instead of maintaining separate ones, on a common infrastructure.

The result of the analysis presents clear advantages when following the vCache shared scenario, in terms of consumed resources for the different situations considered. The sharing schema outperforms the non-shared one, especially relevant for those ISPs that could have less market share. The analysis considered on-demand content however the results can be extrapolated to whatever other scenario. However in some other cases, such as e.g. live content, the relevance of the storage component is not as high as in the on-demand case.

As further work, different lines of complementary research emerge. One of them is to analyze the impact of fixing the size of the cache in such a way that a limited number of contents can be stored. This will provoke that some of the contents should be obtained from remote locations, then impacting on the transport network costs. With that variable, the analysis can be focused on the convenient size of the share and non-shared vCaches for trading off the overall network costs. Another one, is to add the temporal dimension to the analysis, that is, the variation along the time on the use behavior. This can provide a view on the convenience of dynamically instantiate new caches in the network taking profit of the flexibility provided by virtualization. Finally, apart from the network costs, another interesting analysis can come from an analysis centered on the energy efficiency achieved with shared schemas.

Acknowledgment

This work has been partly funded by the project 5GROWTH (Grant Agreement no. 856709).

References

[1] Ericsson white paper, "Ericsson Mobility Report", November 2020. [Online]: https://www.ericsson.com/4adc87/assets/local/mobility-report/documents/2020/november-2020-ericsson-mobility-report.pdf

[2] R. Wood, J. Konieczny, "Fixed network data traffic: worldwide trends and forecasts 2019–2025", Analysis Mason, February 2020.

[3] Cisco white paper, "Cisco Annual Internet Report (2018–2023)", 2020. [Online]: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.pdf

[4] M. Trevisan, D. Giordano, I. Drago, M.M. Munafò, M. Mellia, "Five Years at the Edge: Watching Internet From the ISP Network", *IEEE/ACM Transactions on Networking*, Vol. 28, No. 2, pp. 561-574, April 2020.

[5] ETSI GS NFV 001, "Network Function Virtualizarion (NFV); Use cases", V1.1.1, October 2013.

[6] ETSI GS MEC 002, "Mobile Edge COmpueting (MEC); Technical Requirements", V1.1.1, March 2016.

[7] Akamai white paper, "The Case for a Virtualized CDN (vCDN) for Delivering Operator OTT Video", 2017. [Online]: https://www.akamai.com/us/en/multimedia/documents/white-paper/the-case-for-a-virtualized-cdn-vcdn-for-delivering-operator-ott-video.pdf

[8] Amazon white paper, "Secure Content Delivery with Amazon CloudFront", November 2016. [Online]: https://d1.awsstatic.com/whitepapers/Security/Secure_content_delivery_with_CloudFront_whitepaper.pdf

[9] A. Khan, W. Kellerer, K. Kozu, M. Yabusaki, "Network Sharing in the Next Mobile Network: TCO Reduction, Management Flexibility, and Operational Independence", *IEEE Communications Magazine*, Vol. 49, pp. 134–142, 2011.

[10] D.-E. Meddour, T. Rasheed, Y. Gourhant, "On the role of infrastructure sharing for mobile network operators in emerging markets", *Computer Networks*, Vol. 55, Issue 7, pp. 1576-1591, May 2011.

[11] I. Szczesniak, P. Cholda, A. R. Pach and B. Wozna-Szczesniak, "Interoperator fixed-mobile network sharing", *International Conference on Optical Network Design and Modeling (ONDM)*, Pisa, 2015.

[12] I. Afolabi, T. Taleb, K. Samdanis, H. Flinck, "Network Slicing and Softwarization: A Survey on Principles, Enabling Technologies, and Solutions", *IEEE Comms. Surveys & Tutorials*, Vol. 20, No. 3, pp. 2429-2453, Third Quarter, 2018.

[13] A..A. Barakabitze, A. Ahmad, R. Mijumbi, A. Hines, "5G network slicing using SDN and NFV: A survey of taxonomy, architectures and future challenges", *Computer Networks*, Vol. 167, February 2020.

[14] ETSI GS NFV-INF 001, "Network Function Virtualizarion (NFV); Infrastructure Overview", V1.1.1, January 2015.

[15] T. Böttger, F. Cuadrado, G. Tyson, I. Castro, S. Uhlig, "Open Connect Everywhere: A Glimpse at the Internet Ecosystem through the Lens of the Netflix CDN", *ACM SIGCOMM Computer Communication Review*, Vol. 48, Issue 1, January 2018.

[16] A.-J. Su, D.R. Choffnes, A. Kuzmanovic, F.E. Bustamante, "Drafting behind Akamai", *IEEE/ACM Transactions on Networking*, Vol. 17, pp. 1752–1765, 2009.

[17] L. Peterson, B. Davie, R. van Brandenburg, "Framework for Content Distribution Network Interconnection (CDNI)", RFC 7336, August 2014.

[18] Streaming Video Alliance, "Optimizing Video Delivery With The Open Caching Network", September 2018. [Online]: https://streamingvideoalliance.docsend.com/view/vfijwff

[19] G. Hasslinger, F. Hartleb, "Content delivery and caching from a network provider's perspective", *Computer Networks*, vol. 55, pp. 3991-4006, 2011.

[20] N. Kamiyama, T. Mori, R. Kawahara, H. Hasegawa, "Optimally Designing ISP-Operated CDN", *IEICE Transactions on Communications*, Vol. E96–B, No.3, pp. 790-801, March 2013.

[21] J. Zhang, L.M. Contreras, K. Gao, F. Cano, P. Diez Cano, A. Escribano, Y. R. Yang, "Sextant: Enabling Network-Aware Applications in Carrier Networks", accepted in *IFIP/IEEE International Symposium on Integrated Network Management*, 2021.

[22] H. Gredler, J. Medved, S. Previdi, A. Farrel, and S. Ray, "North-bound distribution of link-state and traffic engineering (TE) information using BGP", RFC 7752, March 2016.

[23] S. Kiesel, W. Roome, R. Woundy, S. Previdi, S. Shalunov, R. Alimi, R. Penno, Y. R. Yang, "Application-layer traffic optimization (ALTO) protocol", RFC 7285, September 2014.

[24] N. Herbaut, D. Negru, Y. Chen, P. A. Frangoudis and A. Ksentini, "Content Delivery Networks as a Virtual Network Function: A Win-Win ISP-CDN Collaboration," *IEEE Global Communications Conference (GLOBECOM)*, Washington, DC, 2016, pp. 1-6,

[25] P. Casas, A. D'Alconzo, P. Fiadino, A. Bär, A. Finamore, T. Zseby, "When YouTube Does not Work—Analysis of QoE-Relevant Degradation in Google CDN Traffic", *IEEE Transactions on Network and Service Management*, Vol. 11, No. 4, pp. 441-457, December 2014.

[26] P.A. Frangoudis, L. Yala, A. Ksentini, T. Taleb, "An architecture for on-demand service deployment over a telco CDN", *IEEE International Conference on Communications (ICC)*, pp. 1-6, Kuala Lumpur, 2016.

[27] I. Benkacem, T. Taleb, M. Bagaa, H. Flinck, "Optimal VNFs Placement in CDN Slicing Over Multi-Cloud Environment", *IEEE Journal on Selected Areas in Communications*, Vol. 36, No. 3, pp. 616-627, March 2018

[28] S. Retal, M. Bagaa, T. Taleb, H. Flinck, "Content delivery network slicing: QoE and cost awareness", *IEEE International Conference on Communications (ICC)*, 2017.

[29] U. Bulkan, T. Dagiuklas, M. Iqbal, "On the modelling of CDNaaS deployment", *Multimedia Tools and Applications*, Vol 78, pp. 6805-6825, 2019.

[30] M. Ruiz, M. Germán, L. M. Contreras, L. Velasco, "Big Data-backed Video Distribution in the Telecom Cloud", *Computer Communications*, Vol. 84, pp. 1-11, 2016

[31] C. Mei, J. Liu, J. Li, L. Zhang, M. Shao, "5G network slices embedding with sharable virtual network functions", *Journal of Communications and Networks*, vol. 22, no. 5, pp. 415-427, October 2020.

[32] F. Malandrino, C. F. Chiasserini, G. Einziger, G. Scalosub, "Reducing Service Deployment Cost Through VNF Sharing", *IEEE/ACM Transactions on Networking*, vol. 27, no. 6, pp. 2363-2376, December 2019.

[33] A. Mohamad, H. S. Hassanein, "On Demonstrating the Gain of SFC Placement with VNF Sharing at the Edge," *IEEE Global Communications Conference (GLOBECOM)*, Waikoloa, HI, USA, pp. 1-6, 2019.

[34] Z. Avramova, D. De Vleeschauwer, S. Wittevrongel, H. Bruneel, "Dimensioning Multicast-Enabled Networks for IP-Transported TV Channels", Proc. of the *International Teletraffic Congress (ITC20)*, pp. 6-17, June 2007.

[35] D. T. van Veen, M. K. Weldon, C. C. Bahr, E. E. Harstead, "An analysis of the technical and economic essentials for providing video over fiber-to-the-premises networks", *Bell Labs Technical Journal*, Vol. 10, No. 1, pp. 181-200, 2005.

[36] B. Berg, et al., "The CacheLib Caching Engine: Design and Experiences at Scale", *USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2020.

[37] B. Naudts, M. Kind, F.-J. Westphal, S. Verbrugge, D. Colle, M. Pickavet, "Techno-economic analysis of software defined networking as architecture for the virtualization of a mobile network", *European Workshop on Software Defined Networking*, 2012.