A Green Video Control Plane with Fixed-Mobile Convergence and Cloud-RAN

Ramon Aparicio-Pardo and Lucile Sassatelli
Université Côte d'Azur, CNRS, I3S - Sophia Antipolis, France
Emails: {raparicio,sassatelli}@i3s.unice.fr

***Abstract*—Video traffic is a key-challenge for fixed and mobile operators facing variable and massive load and variety of Over-The-Top (OTT) videos. Energy consumption is also a heavy Opex component, where the Fixed-Mobile Convergence is a promising solution, built on on economical optic fibers and LTE base-band operations consolidation. By combining Future Internet Architectures (FIA) principles such as ubiquitous caching, SDN and NFV, and FMC, we propose a complete, fully dynamic setup which optimizes both for power consumption and Quality of Experience (QoE), by choosing proper infrastructure (turning on a minimum number of computing and networking equipments) and operational (routing and caching) configurations. Our control plane named ViRCA is scalable thanks to data analytics techniques, and fully reactive to the dynamics of load and catalog both in time and space. Numerical assessments in realistic settings show power gains of up to 30% while the scores on different QoE metrics are maximized. Enabling elastic co-location of caches and radio base-band operations turns out to be crucial for both power and QoE objectives.**

I. Introduction

ISPs, and Mobile Network Operators (MNO) specifically, are facing services' variety and traffic increases. Telcos infrastructure needs to simultaneously support these services, ensure their required QoE, and possibly monetize them. Video traffic is in particular a key-challenge for telcos, due to the share of video streaming in the Internet traffic expected to reach 82% of all IP traffic by 2020 [1].

For MNOs, the network segments vulnerable to congestion and hampering QoE are the backhaul and Radio Access Network (RAN). This is specifically due to IP tunneling used to make the backhaul transparent (3GPP standard). To meet the QoE requirements, a first strategy is to skip the congested areas by employing in-network caching (e.g., iQstream startup), where content is stored at PDN-GWs or eNodeBs in order not to break the mobility management imposing tunnelling between those end-points. Proxy solutions are also used by telcos, in partcular to resize web content for mobile device to save bandwidth. The Evolved Packet Core (EPC) infrastructure shall be overhauled to meet the challenges brought by 5G services. Optical infrastructure and fine-grained (flow-based and location-based) monitoring to feed real-time decisions are key aspects[1]. To this aim, Big data is seen as a top strategic investment by a number of telcos[2]. Key-enablers to this much needed automated orchestration are SDN and NFV which, by decoupling control plane from data plane and software from hardware, respectively, allow to leverage the flexibility and scalability of cloud resources to provide a fine-grained and responsive control of the flows, while scaling up or down the computational resources (e.g., the Elemental^{TM} company provides software-defined video deployment to IPTV providers). To solve the problems (interference due to spectrum limitations and base stations' density, mobility, etc.) at the wireless last-hop, software-defined centralized control is also planned to enable 5G, and is generally referred to as Cloud-RAN, where both controllers and radio elements are hosted in the cloud [2]. The concept of Fog computing brings together these principles to add "a hierarchy of elements between the cloud and endpoint devices [...] to meet these challenges in a high performance, open and interoperable way"[3].

Energy consumption on another hand is a heavy Opex component. One promising solution for energy-efficient aggregation/access is the Fixed-Mobile Convergence (FMC) principle [3] (see Fig. 1: Left). The idea is to manage jointly the heterogeneous access technologies (e.g., FTTH/B, WiFi, 4G) to consolidate within the cloud/fog fixed and mobile optical head-ends as well as most of Base Station (BS) (base-band) processing. This is known as Base Band Unit (BBU) hosting, to mutualize usage of the physical resources (optical/electronic networking equipment and cooling). However, doing so in turn entails costs in bandwidth and opto-electronic-opto conversions (to compute the base-band digital signals up in the network way before the BS), thereby requiring fine control to truly yield energy savings without loss of performance.

We design a control plane addressing OTT video distribution for ISPs/MNOs facing variable and high video loads. By combining FIA principles such as ubiquitous caching, SDN and NFV, and FMC, we propose a complete, fully dynamic setup which optimizes both for power consumption and QoE, by choosing proper infrastructure (turning on a minimum number of computing and networking equipments) and operational (routing and caching) configurations. Our contributions are:

- Based on a realistic power model of micro-Data Center (DC) and networking equipments, high-level video transcoding and low-level base-band LTE operations, we first model the multi-objective (QoE and power) optimization problem which accounts for reactive caching. Indeed, considering recent

This work was partly funded by the French Government through the Investments for the Future Program reference ANR-11-LABX-0031-01.

[1] e.g., https://www.ovum.com/need-real-time-decisions-telcos/

[2] McKinsey: https://tinyurl.com/hq8xcr9

[3] https://www.openfogconsortium.org/

findings advocating for server-controlled video rate through continuous-rate encoding (as opposed to DASH representations selection), we address the case of massive and volatile OTT content for which ISPs do not plan pre-fetching as for Subscription-based VoD (SVoD), and rely on Fog computing.
• A dynamic orchestration with infrastructure-level and operation-level re-optimizations is devised from a primal decomposition to track load variations in time, space and content features. The scalability for massive video data is addressed with clustering techniques which prove efficient in simulations.
• Extended numerical simulations in realistic settings show power gains of up to 30% while the scores on different QoE metrics are maximized thanks to elastic consolidation of caches/transcoders and radio base-band operations. A comparison with ICN is also drawn.

After the related works are presented in Sec. II, Sec. III and IV detail the node and power consumption models. Our control plane is detailed in Sec. V and numerical results are analyzed in Sec. VI before the conclusion.

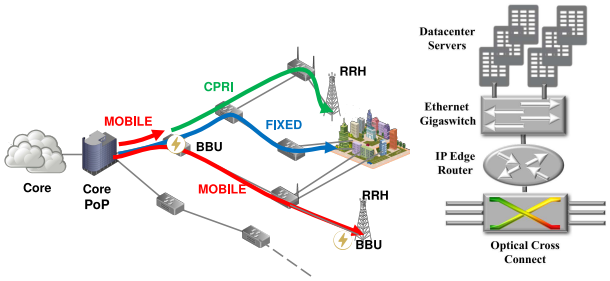

Fig. 1: Left: the FMC concept. Right: A micro-DC node.

II. Related works

We organize this section into three paragraphs, each covering one aspect in our work: joint routing and caching management, congestion minimization and energy-efficient caching.

To minimize transmission costs, [4] expresses the caching and routing subproblems separately, and derives routines for each. So-called "time-slot routing" is used to avoid greedily routing each incoming request by delaying it a bit. In [5], Ruiz et al. design a management system for a telco-CDN to serve a certain set of channels/contents. They express separately different subproblems aimed at reconfiguring the virtual resources to minimize the link costs, HTTP server and packager resources, establishing connections and re-allocating users. While we also explicitly optimize for video quality with a more refined QoE model, we consider easier-to-implement reactive caching for non-SVoD content and proper primal decomposition from the multi-objective problem to come up with subproblems without heuristics, as well as bundle-of-request routing (see Sec. V). In [5], video load prediction is used (which our plane does not encompass for simplicity, but can easily benefit from).

In [6], the ability to cache and transcode is considered at the eNodeB for MPEG DASH discrete representations. Only targeting the sum of the videos' qualities, the authors determine heuristically the caching and rate decisions, and the scheduling both between the core and the eNodeB, and the eNodeB and user via the wireless LTE link. In [7] (close to [8]), the problem of which video representation to cache is thoroughly investigated when MPEG DASH is used. The authors show the effectiveness of caching the one highest representation per video, so that popular videos are served with higher quality from the cache. The work however considers neither the last-hop bandwidth limitation where lower rates videos are needed nor the incurred cost of transcoding. By considering the iProxy solution [9] (each proxy is a server/cache/transcoder generating continuous video rates), our framework is relieved from the burden of DASH representation choices. In [10], the concept of Information-Centric Networks (ICN) is employed to show that, in the backhaul of Orange France, HTTP traffic can be reduced by 60% or more by adding only a few hundreds of GBs of storage overall. However Multi-Path (MP) transfer for video streaming has been shown to be tricky [11], [12], the more so with multi-source ([13] resorts to H.264/SVC and not AVC for that reason).

Energy-efficient caching (or CDNs) has also been investigated (not for video specifcally), in [14] analytically, and in [15] who consider that the static power component can be taken off by turning off links and network cards, which we also consider among other levers. In [16], an online cache-cooperation mechanism is designed so that the nodes make their caching decisions based on their local estimate of the global energy benefit. In [17], we sketched the idea of leveraging FMC with video distribution. We however did not consider caching, transcoding, decomposition and reactiveness to handle real requests and reconfigure. Finally the potential of 5G Cloud-RAN architectures co-locating video and BBU processing was suggested in [18]. From this idea we build a complete control plane enabling such potential.

III. Node and power consumption models

We assume future access/aggregation networks based on fog computing, therefore assuming each node is a micro-DC (represented in Fig. 1: Right) equipped with a few servers (with storage and CPUs), electronic and optical switching capabilities. The power consumption models of each of these elements is detailed in our recent survey [19]. In a nutshell, each is the sum of a static and dynamic component. For servers, the latter is dependent on the number of activated cores, counted in number of Virtual CPUs (vCPUs). For electronic switches and routers, it depends on traffic, and owing to the values in [19], we neglect it. The whole consumption of the optical equipment can be neglected as well. The number of vCPUs, entailing the number of servers switched on, and the number of activated switches and IP routers (where not only the optical cross-connect is used) therefore determine the total power consumption.

Each server may perform video transcoding and base-band tasks, where a task is a Virtual Machine (VM) requesting d vCPUs. A base-band VM hosted at a server is simply the virtualized BBU, or Virtual Digital Unit (VDU), of at least one BS (co-located at the BS in today's LTE RAN). The computational cost associated with each type of task is derived in GOPS and pass-marks, then translated into number of vCPUs in [19]. The LTE traffic is transported between the mobile end node and the possibly deported BBU by means of Common Public Radio Interface (CPRI). That implies to transform the radio band signals from analog to digital, entailing CPRI bit rate C^{CPRI} higher than C^{LTE}.

IV. Video processing and QoE model

We characterize the users' QoE as a function depending on the coding parameters (rate and resolution) and on the video content, by means of the Video Quality Metric (VQM) [20], shown to correlate with human visual perception . We consider four resolutions: $360p$, $720p$, $1080p$, and $2160p$ ($4K$), and three content types with increasing complexity (toon, movie, and sport). From the QoE functions (VQM vs. encoding rate), we extract a linear approximation to later obtain a linear problem (Sec. V-A). The reader is referred to [19] for details.

While MPEG-DASH aims at adapting the served video quality to the available network resources by providing versions encoded at certain different rates, the storing overhead and difficulty of representation choices with a discrete set has led to a sequence of works since 2013 showing that continuous instead of discrete bit rate adaptation enables higher QoE at the client, specifically in mobile networks [9], [21], [22]. In our work, we leverage these findings and assume the client adaptation and cache policy management of the iProxy solution [9] (also detailed in Sec. V-B). We enforce that, upon handling a request, the iProxy instance fetches (if needed) the highest-bitrate video in resolution $1080p$ (resp. $4K$) if the request is for $320p$, $720p$ or $1080p$ (resp. $4K$). DCT-based representation is stored [9], from which transcoding to any lower resolution and bitrate can be made.

V. Design of the ViRCA control plane

The key components of the Virtualized Infrastructure, Routing and Cache Assignment (ViRCA) control plane are presented. First ViRCA is formalized as Mixed Integer Linear Programming (MILP). We next show how the video catalog only serves as a formalism and is not a scale limitation thanks to data analytics. The dynamic orchestration is then designed based on a primal decomposition. Finally we detail how the optimization outcome is used to handle requests and how monitoring is performed. The terms iProxy, server, transcoder and cache are used interchangeably thereon.

A. Optimization formulation of ViRCA

For an access/aggregation network using the micro-DC node architecture in Fig. 1 and for a given video catalog, we search for the allocation of the VMs performing radio and video processing, the content cache placement (proxy instance selection with reactive caching) and the video stream routing maximizing users' QoE and minimizing consumption, jointly.

Let $G(\mathcal{N},\mathcal{L})$ be the graph of a virtual topology of optical paths (lightpaths). The capacity of virtual link $l \in \mathcal{L}$ is the number c_l of lightpaths in the bundle. We consider a set $\mathcal{E}^{\mathcal{F}} \in \mathcal{N}$ of optical head-ends connecting FTTH/B subscribers, and a set $\mathcal{E}^{\mathcal{M}} \in \mathcal{N}$ of cellular BS, and $\mathcal{E} = \mathcal{E}^{\mathcal{F}} \cup \mathcal{E}^{\mathcal{M}}$. The (regional) Point of Presence (PoP) is the highest hierarchy level. All the nodes are assumed to be composed as in Section III. We consider the CPRI for transporting these base-band signals from the BBU hotels (VDU) to the base stations. The constraints on CPRI routing are the same as in [23]. As mobile processing in the cloud requires strict delay limits, any lightpath from node $i \in \mathcal{N}$ to $e \in \mathcal{E}$ cannot exceed a certain reach. The possible set of lightpaths is $\mathcal{L}^{CPRI} \in \mathcal{L}$. The set of IP paths is referred to as $\mathcal{P}$. The set of paths: joining a node pair $(i,j) \in \mathcal{N} \times \mathcal{N}$ is denoted as $\mathcal{P}_{ij}$, traversing a virtual link $l \in \mathcal{L}$ as $\mathcal{P}_l$, traversing a node $i \in \mathcal{N}$ as $\mathcal{P}_i^{node}$, coming into a node $i \in \mathcal{N}$ as $\mathcal{P}_i^{in}$, and going out of a node $i \in \mathcal{N}$ as $\mathcal{P}_i^{out}$. The catalog of contents is $\mathcal{M}$. The demand is the average number v_{ems} of parallel requests for content $m \in \mathcal{M}$ issued from $e \in \mathcal{E}$, at resolution $s \in \mathcal{S} = (360p, 720p, 1080p, 4K)$. Instead of having a request for a commodity, a commodity $\{ems\}$ is hence a bundle of those. This coarser granularity allows having MP routing at the level of bundles, if not at the level of request (each using a single path, see Sec. V-D).

As shown in [19], we consider the transcoding power only depends on the output resolution and video type, whereby the definition of d^{ms} in Table I. A content requested after the last catalog update (i.e. not present in the current version) is denoted as $u \notin \mathcal{M}$. Its explicit consideration enables using optimally the resources not actually allocated for the foreseen catalog contents. Tables I and in II give all detail.

Trading between power savings and QoE improvements is a multi-objective optimization that we address by scalarization in Eq. (1a) with parameter γ. We could use a Nash Bargaining Solution (NBS) formulation to remove the parameter by minimizing the objectives' ratio, but keeping the problem linear helps convergence speed even in high-dimensional problems. The motivation of model in Eq. (1) is introduced in Sections III and IV, and detailed in [19].

$$\min_{\{x,y,f,vf,h,z,g,r,k,w,t\}} power - \gamma QoE \tag{1a}$$

$$power = P^{CPU} + P^{EGS} + P^{IPR} \tag{1b}$$

$$P^{CPU} = \sum_{i\in\mathcal{N}} [106.4k_i + 10.417(w_i + t_i)] \tag{1c}$$

$$P^{EGS} = \sum_{i\in\mathcal{N}} 2020 g_i, \quad P^{IPR} = \sum_{i\in\mathcal{N}} 4550 r_i \tag{1d}$$

$$QoE = \sum_{\substack{e\in\mathcal{E}\\ m\in\mathcal{M}\\ s\in\mathcal{S}}} \alpha_{ms} \sum_{i,j} x_{ij}^{ems} + \sum_{\substack{e\in\mathcal{E}\\ s\in\mathcal{S}}} \alpha_s \sum_{i,j} x_{ij}^{eus} \tag{1e}$$

1) QoE constraints: The linear approximation of the QoE does not provide intrinsic fairness among the bundles, as the concave function does. We therefore add the next constraints for bounding the QoE for each triplet $\{ems\}$:

$$\sum_j x_{ij}^{ems} \geq b_{min}^{ms} v_{ems} f_i^{ems}, \quad i \in \mathcal{N}, e \in \mathcal{E}, m \in \mathcal{M}, s \in \mathcal{S} \tag{2a}$$

Name	Description
$\lambda_{im} \in \mathbb{R}^+$	Poisson arrivals' intensity of request for content $m \in \mathcal{M}$ at node $i \in \mathcal{N}$
$v_{ems} \in \mathbb{R}^+$	Average number of parallel video requests for content $m \in \mathcal{M}$ at resolution $s \in \mathcal{S}$ from end point $e \in \mathcal{E}$
$wd_{ems} \in \mathbb{R}^+$	Average watching duration of content $m \in \mathcal{M}$ at resolution $s \in \mathcal{S}$ from end point $e \in \mathcal{E}$
$o_i^{ems} \in \mathbb{R}^+$	Traffic estimate needed for reactive caching at node $i \in \mathcal{N}$ for video requests of content $m \in \mathcal{M}$ at resolution $s \in \mathcal{S}$ from end point $e \in \mathcal{E}$
$b_i \in \mathbb{R}^+$	Minimum bandwidth to have between PoP and node $i \in \mathcal{N}$ to ensure connectivity
$b_{min}^{ms}(b_{max}^{ms}) \in \mathbb{R}^+$	Encoding bit rates corresponding to the minimum (resp. maximum) quality for content $m \in \mathcal{M}$ at resolution $s \in \mathcal{S}$ (in $Mbps$)
$\alpha_{ms}(\alpha_s) \in \mathbb{R}^+$	Parameter for QoE linear approx. at resolution $s \in \mathcal{S}$ (resp. for unknown content)
$C_e^{LTE} \in \mathbb{R}^+$	Overall capacity of LTE radio links at base station $e \in \mathcal{E}_m$ (in $Mbps$)
$C^{WDM} \in \mathbb{R}^+$	WDM channel capacity (in $Mbps$)
$C^{IPR} \in \mathbb{R}^+$	IP router switching capacity (in $Mbps$)
$C^{EGS} \in \mathbb{R}^+$	Ethernet gigaswitch switching capacity (in $Mbps$)
$c_l \in \mathbb{Z}^+$	Capacity of virtual link l (in number of lightpaths)
$s_m \in \mathbb{R}^+$	Size of content $m \in \mathcal{M}$ (iProxy representation size)
$S_i \in \mathbb{R}^+$	Total storage capacity at node $i \in \mathcal{N}$
$d_e \in \mathbb{R}^+$	Number of vCPUs (CPU fraction) required for the BBU tasks for BS $e \in \mathcal{E}_m$ (i.e. number of VDUs processing the traffic destined to node e)
$d^{ms} \in \mathbb{R}^+$	Number of vCPUs (CPU fraction) required to produce a representation of content $m \in \mathcal{M}$ at resolution $s \in \mathcal{S}$ from its stored version
$d^{us} \in \mathbb{R}^+$	Number of vCPUs (CPU fraction) required to produce a representation of content $u \notin \mathcal{M}$ at resolution $s \in \mathcal{S}$ from its stored version
$T \in \mathbb{Z}^+$	Number of cores (vCPUs) per physical server
$K \in \mathbb{Z}^+$	Number of physical servers per data center node

TABLE I: MILP notation. Input Parameters

Name	Description
$x_{ij}^{ems} \in \mathbb{R}^+$	Total traffic rate for request bundle $\{ems\}$ served from node $i \in \mathcal{N}$ to node $j \in \mathcal{N}$ (in $Mbps$)
$x_p^{ems} \in \mathbb{R}^+$	Traffic rate for request bundle $\{ems\}$ served on path $p \in \mathcal{P}$ (in $Mbps$)
$x_{ij}^{eus} \in \mathbb{R}^+$	Total traffic rate for request bundle $\{eus\}$ served from node $i \in \mathcal{N}$ to node $j \in \mathcal{N}$(inMbps)
$x_p^{eus} \in \mathbb{R}^+$	Traffic rate for request bundle $\{eus\}$ served on path $p \in \mathcal{P}$ (in $Mbps$)
$y_p \in \mathbb{R}^+$	Background traffic on path $p \in \mathcal{P}_{PoP}^{out}$ (in $Mbps$)
$f_i^{ems} \in [0,1]$	Fraction of requests v_{ems} served from node $i \in \mathcal{N}$
$vf_i^{eus} \in \mathbb{R}^+$	Number of requests for non-cataloged content served from node $i \in \mathcal{N}$
$h_{im} \in [0,1]$	Hit ratio of content $m \in \mathcal{M}$ at node $i \in \mathcal{N}$ (probability for i to store m)
$z_i^e \in \{0,1\}$	1, if node $i \in \mathcal{N}$ hosts BBU of base station $e \in \mathcal{E}$ 0, otherwise
$g_i \in \{0,1\}$	1, if node $i \in \mathcal{N}$ is switched on; 0, otherwise
$r_i \in \{0,1\}$	1. if IP router used at node $i \in \mathcal{N}$; 0, otherwise
$k_i \in \mathbb{Z}^+$	Number of active servers at node $i \in \mathcal{N}$
$w_i \in \mathbb{Z}^+$	Number of vCPUs at node $i \in \mathcal{N}$ performing the BBU processing tasks
$t_i \in \mathbb{Z}^+$:	Number of vCPUs at node $i \in \mathcal{N}$ performing the video transcoding (iProxy) tasks

TABLE II: MILP notation. Decision variables

$$\sum_j x_{ij}^{ems} \leq b_{max}^{ms} v_{ems} f_i^{ems}, \quad i \in \mathcal{N}, e \in \mathcal{E}, m \in \mathcal{M}, s \in \mathcal{S} \tag{2b}$$

$$\sum_j x_{ij}^{eus} \geq b_{min}^{s} v f_i^{eus}, \qquad i \in \mathcal{N}, e \in \mathcal{E}, s \in \mathcal{S} \tag{2c}$$

$$\sum_j x_{ij}^{eus} \leq b_{max}^{s} v f_i^{eus}, \qquad i \in \mathcal{N}, e \in \mathcal{E}, s \in \mathcal{S} \tag{2d}$$

2) Routing constraints:

$$\sum_{\substack{i \in \mathcal{N}, s \in \mathcal{S} \\ m \in \mathcal{M} \cup u \\ p \in \mathcal{P}_{ij}}} x_p^{ems} \leq C_e^{LTE} z_j^e, \qquad e \in \mathcal{E}^{\mathcal{M}}, j \in \mathcal{N} \tag{3a}$$

$$\sum_{p \in \mathcal{P}_{ij}} x_p^{ems} = x_{ij}^{ems}, \qquad \begin{matrix} e \in \mathcal{E}, m \in \mathcal{M} \cup u \\ s \in \mathcal{S}, (i,j) \in \mathcal{N} \times \mathcal{N} \end{matrix} \tag{3b}$$

$$x_{ij}^{ems} = 0, \qquad \begin{matrix} e \in \mathcal{E}^{\mathcal{F}}, m \in \mathcal{M} \cup u, s \in \mathcal{S} \\ i \in \mathcal{N}, j \in \mathcal{N} \setminus \{e\} \end{matrix} \tag{3c}$$

$$\sum_{p \in \mathcal{P}_{PoPi}} y_p \geq \sum_{\substack{m \in \mathcal{M} \\ e \in \mathcal{E} \\ s \in \mathcal{S}}} o_i^{ems} f_i^{ems} + b_i, \qquad i \in \mathcal{N} \tag{3d}$$

$$y_p = 0, \qquad j \in \mathcal{N}, p \notin \mathcal{P}_{PoPj}^{IP} \tag{3e}$$

$$\sum_{p \in \mathcal{P}_l} \Big(\sum_{\substack{e \in \mathcal{E} \\ m \in \mathcal{M} \cup u \\ s \in \mathcal{S}}} x_p^{ems} + y_p \Big) \leq C^{WDM} c_l \qquad l \in \mathcal{L}, \tag{3f}$$

$$\sum_{p \in \mathcal{P}_i^{node}} \Big(\sum_{\substack{e \in \mathcal{E} \\ m \in \mathcal{M} \cup u \\ s \in \mathcal{S}}} x_p^{ems} + y_p \Big) \leq C_i^{IPR} r_i \qquad i \in \mathcal{N}, \tag{3g}$$

$$\sum_{p \in \mathcal{P}_i^{in} \cup \mathcal{P}_i^{out}} \Big(\sum_{\substack{e \in \mathcal{E} \\ m \in \mathcal{M} \cup u \\ s \in \mathcal{S}}} x_p^{ems} + y_p \Big) \leq C_i^{EGS} g_i \qquad i \in \mathcal{N}, \tag{3h}$$

In (3d), b_i is set to a low value (e.g. 1Mbps) for fixed end nodes and BBU hotels, to ensure connectivity with the PoP. Contrary to previous works [5], [7], [24], [16], we are able to model the reactiveness of our caching policies (which relieves from pre-fetching to consider any OTT service) by estimating the required bandwidth w_{PoP-i} from PoP to cache i to serve cache misses. It was shown in [25] that FIFO and LRU caching policies can have their hit ratios modeled by those of TTL caches with proper timer value. Let us consider LRU to approach the considered iProxy policy. From Little's law we get (see notation in Table I):

$$\lambda_{im} = \sum_{es} \frac{v_{ems} f_{iems}}{wd_{ems}}$$

The hit ratio is given by [25]: $h_{im} = 1 - \exp(-\lambda_{im} T)$, with T such that $\sum_m (1 - \exp(-\lambda_{im} T))\, s_m = S_i$. We therefore get $w_{PoP-i} \geq \sum_m (1 - h_{im}) \lambda_{im} s_m$. Approximating $(1 - h_{im})$ with a constant κ (between 0.5 and 0.8 as in Orange traces [10, Sec. II]), we obtain $o_i^{ems} = \kappa \frac{v_{ems} s_m}{wd_{ems}}$.

3) VDU placement constraints:

$$z_i^e = 0, \qquad e \in \mathcal{E}^{\mathcal{F}}, i \in \mathcal{N} \tag{4a}$$

$$z_i^e = 0, \qquad e \in \mathcal{E}^{\mathcal{M}}, i \in \mathcal{E} \setminus \{e\} \tag{4b}$$

$$z_i^e = 0, \qquad e \in \mathcal{E}^{\mathcal{M}}, i \in \mathcal{N} \quad | \quad \mathcal{L}_{ie}^{CPRI} = \{\emptyset\} \tag{4c}$$

$$\sum_{i \in \mathcal{N}} z_i^e = 1, \qquad e \in \mathcal{E}^{\mathcal{M}} \tag{4d}$$

$$\sum_{e \in \mathcal{E}^{\mathcal{M}}} d_e z_i^e \leq w_i, \qquad i \in \mathcal{N} \tag{4e}$$

4) Transcoder and cache placement constraints:

$$\sum_{i\in\mathcal{N}} f_i^{ems} \leq 1, \qquad e\in\mathcal{E}, m\in\mathcal{M}, s\in\mathcal{S} \tag{5a}$$

$$h_{im} \geq f_i^{ems}, \qquad i\in\mathcal{N}, e\in\mathcal{E}, m\in\mathcal{M}, s\in\mathcal{S} \tag{5b}$$

$$\sum_{m\in\mathcal{M}} s_m h_{im} \leq S_i, \qquad i\in\mathcal{N} \tag{5c}$$

$$\sum_{\substack{e\in\mathcal{E}\\ s\in\mathcal{S}}} \left(\sum_{m\in\mathcal{M}} d_{ms} v_{ems} f_i^{ems} + d_{us} v f_i^{eus} \right) \leq t_i, \ \ i\in\mathcal{N} \tag{5d}$$

5) VM placement constraints:

$$w_i + t_i \leq T k_i, \qquad i\in\mathcal{N} \tag{6a}$$

$$k_i \leq K g_i, \qquad i\in\mathcal{N} \tag{6b}$$

B. Scaling up catalog with video analytics

Our system targets OTT distribution, where ISPs/MNOs are faced with variable and high video loads of large and variable catalogs, for which content pre-fetching is not planned. Other works employing MILP formulation for caching and routing assignments are plagued with the curse of dimensionality: a catalog of only 100 content is considered in [24], [5] considers a limited number of channels and heuristics, while [7] designs a heuristic reactive caching policy based on the insights from the low-dimensional MILP. We take a different approach. We posit that key features impacting the optimization problem (and not the content ids) are necessary and sufficient. This approach to manage high and variable volumes of very diverse videos calls for data analytics to extract the only information necessary for the problem at hand, as exposed below. It proves highly efficient to find QoE-power trade-offs within a few seconds to minutes as detailed in Sec. VI.

The size of the MILP depends on the number of ems {end node id,content id,resolution}. Before each optimization round, a clustering is performed to collapse the requests into groups meaningful to the optimization by revealing organization of the requests into patterns [26]. Their considered features are those impacting the resource allocation: content type, duration, size, resolution indicator (0 for up to 1080p, 1 for 4K), number of parallel requests for each end node. As an ordinal variable, type is normalized as a rank, duration, size and parallel requests as z-scores, while the 4K flag is left as binary [26, Chap. 4]. A K-Means clustering with Euclidean distance is then invoked on the normalized observation matrix made of all the requested content. The maximum number of centroids can be set to control the MILP solving complexity, which scales as (number of paths)×(number of contents). We set this max product to 10^6 in the results below and deduce the maximum number of clusters in each case. The obtained centroids are then de-normalized. The size of each centroid is replaced with the sum of sizes of cluster's members, which guarantees that all individual contents in the same cluster can be stored in the intended cache. The 4K flag of each centroid is set to the majority. The hence obtained synthetic contents are representative of the actual demand. We do not consider prediction of the video demands, though this can be included to make the synthetic content even more accurate, as in [5].

The second data analytics tool is provided within iProxy [9]: each cache stores a content-based hash (from DCT coefficients) of each video, i.e. a video dictionary. It prevents from storing replicates requested from different URLs.

C. Decomposition and dynamic orchestration

A primal decomposition is typically used in resource allocation problems (as the ViRCA problem) where a *master problem* allocates the existing resources by directly giving each subproblem the amount of resources that it can use [27]. We identify here the virtualized infrastructure allocation variables (z, g, r, k, w, t) as the *coupling* variables, leading to a two-level structure where ViRCA corresponds to the *master problem* and the subproblem, called Routing and Caching Assignment (RCA), solves routing and content cache placement (iProxy instance selection in our reactive framework). Since the physical and virtualized network infrastructure (VMs deployment and active routers) is an input parameter of the RCA problem, the computational and switching budgets are the resources RCA allocates to maximize QoE only:

$$\max_{\{x,y,f,vf,h\}} QoE, \quad \text{s.t. (2), (3), (5)} \tag{7}$$

where the integer variables z, g, r, k, w, t are now inputs set to the last ViRCA solution's values. Another benefit is that, since all the integer variables in the ViRCA problem are *coupling* variables, the RCA subproblem becomes a simpler Linear Programming (LP) problem solvable in polynomial time.

Algo. 1 shows how the network and computational resources dynamically reconfigure by means of ViRCA and RCA. ViRCA is not meant to be solved as often as RCA, as it implies activating and deactivating DC nodes introducing non negligible control overheads, while RCA simply consists of lighter routing and iProxy selection updates. Before running ViRCA or RCA, to reduce the unnecessary executions of the solver (typically, CPLEX), two re-optimization conditions are identified. For ViRCA, the actual number of used vCPUs $\#vCPUs$ in all the micro-DC nodes is simply verified to decide whether one of the nodes can be switched off (meaning VMs can be better consolidated). For the RCA case, the re-optimization condition is based on the validity of the last RCA optimal solution with respect to the current network state. In a few words, we compare if the video demand variation in terms of traffic units between two consecutive RCA re-optimizations can be accommodated in the budgets of bandwidth (x_i^{ems} variables) and transcoding resources (t_i variables) found at the last RCA or ViRCA run, respectively. To do so, the slackness of the constraints (2) and (5d) are saved. These slackness (referred to as Σ^*) represent the spare bandwidth and transcoding resources for the optimal allocation. Then, from the monitoring described in the next paragraph, we compute the variations $\Delta\{v_{ems} f_i^{ems}\}$ and $\Delta\{v f_i^{eus}\}$ between two consecutive RCA re-optimizations (discrepancy between the planned and actual number of parallel requests of type

ems served from i). We obtain the approximate slackness $\widehat{\Sigma}$. If $\widehat{\Sigma}$ does not exceed Σ^*, the last optimized bandwidth and transcoding allocations can still hold the actual video demand despite the Δ deviations. Otherwise, RCA is triggered again. If this RCA re-optimization provokes a significant drop in QoE (larger than 20% of the maximum QoE), the ViRCA routine is triggered again. This enables the resources to be scaled up when only re-configuration with RCA is mo more sufficient.

Algorithm 1: Routine to trigger re-optimizations

if *time* $\geq$ *lastViRCAreopt_time* + *ViRCA_reoptimPeriod* **then**
  lastViRCAreopt_time = *time* ;
  if *#vCPUs* $\leq \sum_{\in \mathcal{N}} (w_i + t_i) - KT$ **then**
    Trigger power-QoE optim. by running **ViRCA**;
else if *time* $\geq$ *lastRCAreopt_time* + *RCA_reoptimPeriod* **then**
  lastRCAreopt_time = *time* ;
  if $\widehat{\Sigma} \geq \Sigma^*$ **then**
    Trigger QoE optim. by running **RCA**;
    if *RCA_sol* $\leq$ *0.8 maxQoE* **then**
      Trigger power-QoE optim. by running **ViRCA**;

D. Implementation

We consider a SDN management, where each client incoming request is intercepted and sent to the controller. As described in Algo. 2, based on the x_p^{ems} budgets, the controller decides which iProxy is going to serve the client through which path (injecting the appropriate rules into the switches), unlike in greedy or time-slot routing [4]. Another key-component is the monitoring process. The average number $vf_i^{ems}(k)$ of parallel requests of type ems (defined over synthetic content used for the optimization) served and rejected by node i is monitored each period k of duration T_{sample}. The RCA re-optimization is hence triggered depending on the value $vf_i^{ems} = \max_{k=1,\dots,K} vf_i^{ems}(k)$, where K is the number of samples since last optimization. Prior to running RCA or ViRCA, the content clustering is performed with the new video set (which may have changed) since last optimization.

VI. Numerical results

In order to get first assessments, we create a Matlab discrete-event simulator (with CPLEX as solver). By lack of space, this choice is more thoroughly motivated in [28]. However to consider reproducible research standards, we make our simulator publicly available at [28] and are planning a full deployment within ns-3.

A. Simulation settings

The FMC and backhaul target scenarios are represented with 2 topologies: 'FMC tree' from [23] and 'Mobile backhaul' from [10] depicted in Fig. 2. Dashed links are redundant links only used in case of failure in today's configuration, but meant to be activated in FIA, such as SDN or ICN for 5G; we thus consider them activated permanently. FMC tree is a four-stage tree with 1 PoP, 2 level-1 and 4 level-2 aggregation nodes. The latter connect the end nodes (5 fixed, 10 mobile).

Algorithm 2: Routine to handle requests

Data: request for video with attributes [type,duration,size,4K flag,originating end node e], $B(p)$: available bandwidth on each path p, $CPU(i)$: available CPUs at i, cluster centroids used for the last optim., budgets x_p^{ems} from last optim.
Result: server index i and path p
Find ems by classifying request into the appropriate cluster based on attribute vector;
Find p with $x_p^{ems} \geq b_{min}^{ms}$, $B(p) \geq b_{min}^{ms}$ and $CPU(src(p)) \geq d^{ms}$;
if p *empty* **then**
  Find p with $\sum_{ms} x_p^{ems} \geq b_{min}^{ms}$, $B(p) \geq b_{min}^{ms}$ and $CPU(src(p)) \geq d^{ms}$;
else if p *empty* **then**
  Find p with dest. e (if fixed) or $BBU(e)$ (if mobile) s.t. $B(p) \geq b_{min}^{ms}$, and $CPU(src(p)) \geq d^{ms}$;
else
  Reject request;
if p *not empty* **then**
  return p, $i = src(p)$;

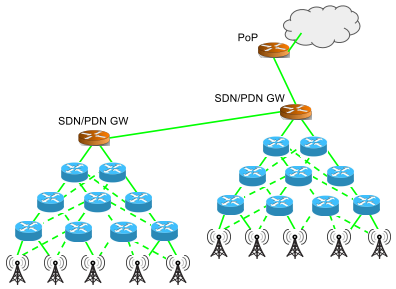

Fig. 2: The Mobile backhaul topology from [10]

The load and catalog assumptions are the same as in the literature [10], [4], [5], [24]. Users demand follows a Poisson process of rate $\lambda = 0.4$ (requests per end node per second). We later investigate the impact of the load on performance, through a load factor applied on λ. A catalog is made of 10000 videos, with a Zipf-distributed popularity with parameter 0.8. Let us re-state that the catalog size does not impact the online operation of our optimization (the catalog is never assumed to be known a priori), but only serves to generate the event trace. To be representative of YouTube-like services where short videos prevail [29] and other OTT services like Netflix, we consider 3 possible durations of 4, 15 and 60 min. As well, shorter videos prevail on mobile accesses, and the distributions of video durations are set to $[0.5, 0.3, 0.2]$ on fixed accesses, on mobile accesses to $[0.66, 0.24, 0.1]$ and $[0.5, 0.3, 0.2]$ for high and low popularity content, respectively. As most of YouTube videos are abandoned before the end, we consider the actual watching duration of each request

represent a random fraction of the content duration, with three possible modes $[0.72, 0.65, 1.00]$ taken from [30] for the first two. The last 1.0 accounts for longer videos such as series or movies that people tend to watch entirely. The probabilities of each resolution $(360p, 720p, 1080p, 4K)$ to be requested are set to $[0.3, 0.3, 0.3, 0.1]$ for fixed accesses and $[0.4, 0.4, 0.2, 0]$ for mobile, based on [31]. According to [32], [33], the last-hop bandwidths are picked within $[5, 15, 30, 40, 60, 80]$ Mbps and $[2, 5, 10, 15, 20, 25]$ Mbps for fixed and mobile accesses, respectively. The simulation results are obtained with $T_{sample} = 5$min (as in MPLS-TE), $RCA_reoptimPeriod = 5$min, $VIRCA_reoptimPeriod = 30$min, $\kappa = 0.5$. The per-node storage is $500GB$.

To objectively assess the gains of each component of ViRCA (flexible caching, transcoding and BBU consolidation), we define the following competitors:

- ViRCA2: ViRCA with no possible deportation of BBU (entirely flexible caching with no FMC);
- ViRCA3: ViRCA with caching only at end node (or BBU if mobile node) and PoP as considered in the PDNCache/ENodeBCache solution in [10]; corresponds to constrained caching compliant with today's 4G functioning (also with BBU deportation here).

Finally the QoE metrics retain only the impacts of elements ViRCA controls: (i) rejection ratio, (ii) average relative rate: for each served request, (served video rate)$/\min(b^{ms}_{max}$, last hop bw) and (iii) startup delay to fetch the first 15s of video (non-zero, when a iProxy is forced to fetch the content through the PoP; and negligible, otherwise, as the cache-to-client delay is maintained almost fixed by the iProxy video rate adaptation to bandwidth).

B. Pareto analysis of optimal solutions

We first consider the results of the optimization alone and analyze the QoE-power tradeoff. For conciseness, we cannot show all results of both topologies. They however yield qualitatively similar results analyzed thereon. Fig. 3 represents the Pareto curves where each point is obtained for a certain value of γ (see Eq. 1a), which denotes below the normalized value once the difference of units between the QoE and power component has been corrected. The first asset is that our ViRCA plane allows to find the minimum power to reach the highest QoE, by considering $\gamma > 1$. The gains in power range from 10% for FMC tree to 30% for Mobile backhaul, compared to ViRCA2 which does not allow BBU deportation. While ViRCA3 exhibits gains in power too, when the load increases it is unable to maintain QoE. Indeed, as the number of cacheable content is included in the QoE formulation, ViRCA3 is unable to spend more power to cache (and transcode) more content beyond the last blue point, as it allows for caching only at BBU and PoP. ViRCA, with all degrees of freedom both in terms of BBU deportation and flexible caching, obtains the best of both limited solutions ViRCA2 (no convergence) and ViRCA3 (no ubiquitous caching).

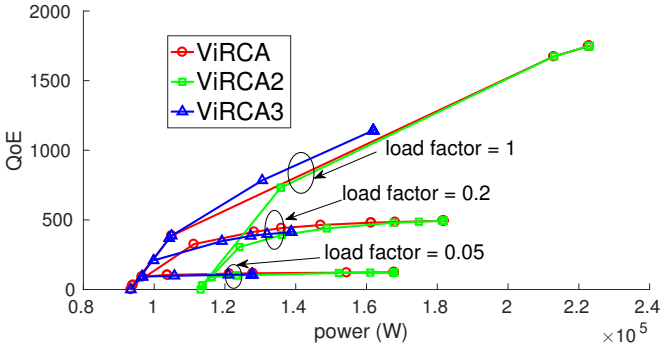

Fig. 3: Pareto frontier, FMC tree

Fig. 4 depicts the breakdown of consumed power between CPU (for transcoding and BBU operations), switching (for mini-DCs hosting a BBU or cache) and IP routing (for mini-DCs adding, dropping or simply switching IP traffic via a virtual link). When the load factor grows from 0.05 to 0.2, the CPU consumption almost doubles (more requests must be transcoded) while the routing and switching remain almost constant as no more nodes are used. The increase in active nodes is seen when reaching a load factor of 1. Proper consolidation is therefore crucial for power gains.

Finally, Fig. 5 shows the geographic breakdown of each power item. Under low load (0.01 and 0.05), most of the power is located high in the network, showing a high level of consolidation: co-location of the various computations and higher number of uninterrupted lightpaths. The load increase makes more lower-level nodes and end nodes to be used, to exploit the computational and caching abilities.

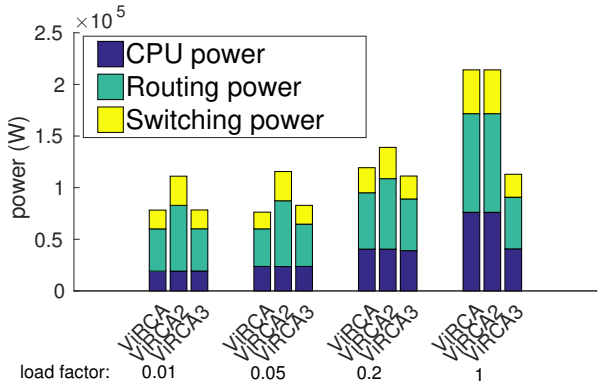

Fig. 4: Breakdown of power, Mobile backhaul, $\gamma = 1$

C. Simulation results: fixed load

The following results are obtained for Mobile backhaul. Fig. 6 to 7 are obtained for a constant load factor of 1 and $\gamma = 1$. While the ViRCA model reduces the QoE to a single metric (the linearized VQM) for the sake of tractability, it is remarkable that ViRCA outperforms its competitors on other metrics (startup-delay and rejection ratio). Fig. 6 indeed shows that ViRCA achieves a number of rejected requests lower than those of ViRCA2 and ViRCA3, an intermediate startup delay for a power consumption close to that of ViRCA3. The relative

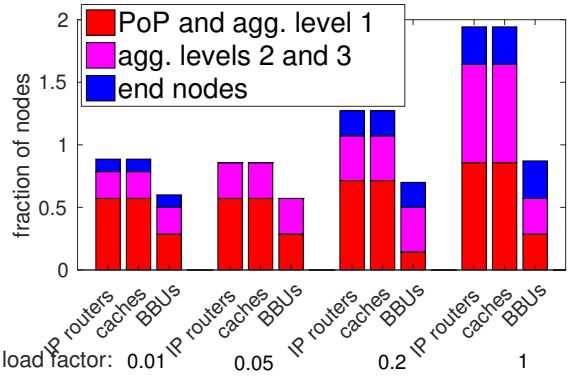

Fig. 5: Breakdown of power, Mobile backhaul, $\gamma = 0.4$

video rate (for the accepted requests) saturating at 1 reveals that the limiting resource in this configuration are the available serving iProxies. As exposed before, the delay is zero when the selected iProxy does not need to fetch content via the PoP, and ViRCA3 is more often in such situation, since iProxies are more constrained to be at the PoP, yielding to lower startup delays.

Fig. 7 shows the startup delay (top figure) per class of popularity (from high to low: 5% most popular, next 15%, next 80%). Let us first specify that the other two QoE metrics, fraction of rejections and relative rate, are not correlated with the popularity. Indeed, there is intentionally no (costly) cache lookup at the time of assigning a request to a cache and no resource reservation (only planning as in MPLS-TE).

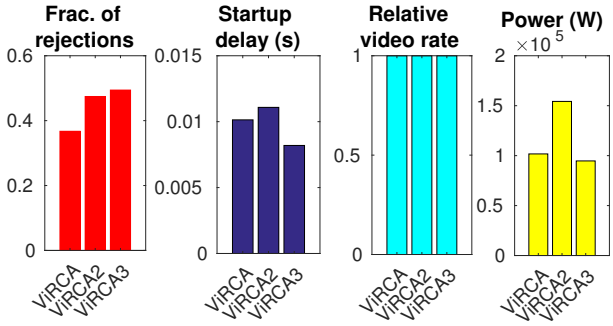

Fig. 6: Global performance metrics

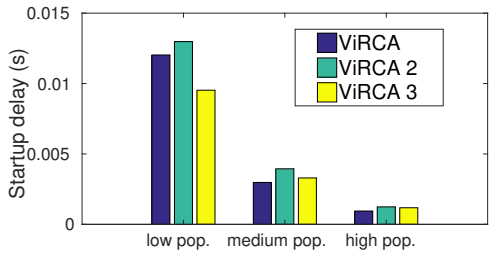

Fig. 7: Metrics per content popularity

D. Simulation results: varying load and catalog

We then consider the load varies over time and space, corresponding to flash crowds. To make sure to provide maximum QoE under minimum corresponding power, γ is set to 10. The end nodes are divided in halves. The 4 successive quarters of requests are generated with a load factor of $[0.05, 1, 0.05, 0.05]$ respectively for the first half of end nodes, $[0.05, 0.05, 1, 0.05]$ for the second half. The plotted fraction of nodes are with respect to the maximum value over time. Fig. 8 shows that the fraction of (virtual) CPUs dedicated to BBU operations remains constant as the vCPUs demand d_e is independent from load [19]. The red dots denote the reason for a rejection (1: not enough CPU, 2 and 4: not enough bandwidth cache-end node and PoP-cache, resp.). When the load suddenly increases, the drop in QoE (relative rate) triggers a ViRCA optimization right after the more frequent RCA. The number of VCPUs for transcoding gets higher, and the QoE goes back to maximum. Another drop is experienced when the load shifts to the second half of end nodes, and proper re-optimization is again performed to re-locate the resources. After the flash crowds, resources are scaled back down again while maintaining maximum QoE.

Our proposal lies within the FIA trends, in particular for centralized control and MP ability. If MP to serve a single request is not considered as in ICN, optimizing budget over bundles of similar (ems) requests is meant to leverage MP at the bundles level. To verify if this ability is indeed exploited by the system, Fig. 9 shows that during the period of flash-crowd about 15% to 25% of concurrent same-type requests follow two different paths from the same cache to the same end nodes. This fraction may be increased in case the main limiting resource is not the cache's CPUs (as shown in Fig. 8) but the bandwidth (if links capacity are lowered to current values where optic fibers are not in the whole backhaul yet). Indeed, let us change the number T of CPUs per machine from 12 to 48 to compare with an ICN solution such as that presented in [10] for this same backhaul topology. Fig. 9 shows that the gains in rejections between ViRCA (meant to encompass the ICN abilities) and ViRCA3 (restricted solution similar to the PDNCache/ENodeBCache solutions of [10]) are about 50%. This is the same order of magnitude as the delay performance shown in [10, Fig. 3]. We compare to delay performance as our simulator does not involve retroactive bandwidth sharing with congestion control and merely rejects excess requests. This comparison thereby demonstrates that our scheme is able to leverage the ICN principle, while being more complete by incorporating formally video QoE and power objectives through the transcoding and FMC capabilities.

Finally, we consider that the catalog varies over time and space. The load factor is set to 0.05. The event trace is generated with 2 distinct catalogs. The share of the first catalog over the 4 quarters is $[1, 1, 0.8, 0]$ for the first half of end nodes, and $[1, 0.8, 0, 0]$ for the second half. Fig. 10 shows that the system is able to keep up with the content features changes.

The above analysis therefore demonstrates the ability of the proposed system to handle highly dynamic environments.

VII. Conclusion

We have designed a control plane for telcos to address the massive increase of OTT video demand. Bringing to-

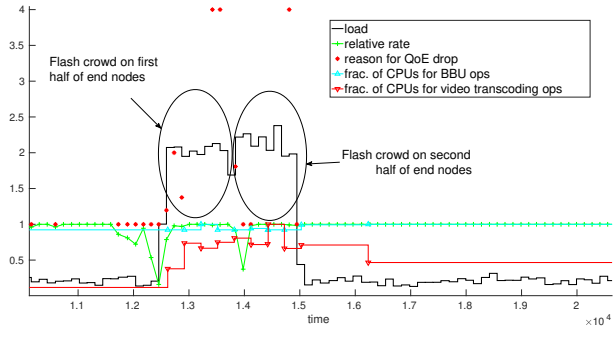

Fig. 8: Load varying in time and space

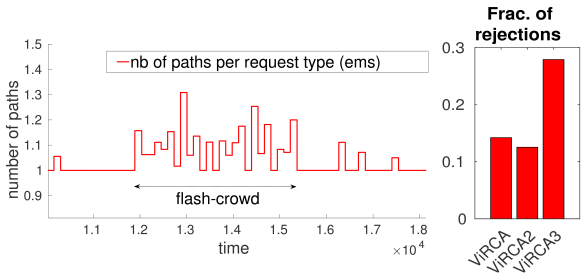

Fig. 9: Left: Instantaneous number of paths per request type. Right: Performance with $T = 48$ instead of 12 (static load).

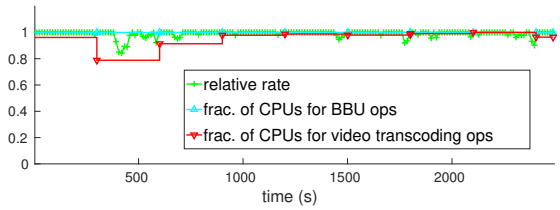

Fig. 10: Catalog varying in time and space

gether BBU deportation (FMC), micro-DC with virtualization, transcoding, reactive caching and data analytics, the most power-efficient configuration of active nodes, routing and caching is found to get the highest QoE. A dynamic orchestration with infrastructure-level and operation-level re-optimizations is devised from a primal decomposition to track load variations in time, space and content features. Simulations show power gains of up to 30% while the scores on different QoE metrics are maximized. Elastic consolidation of caches/transcoders and radio base-band operations is hence crucial for power gains while maintaining highest QoE. Next works involve employing column-generation to better scale the optimization in number of paths, and deploying our control plane on an SDN testbed.

REFERENCES

[1] Cisco, "VNI Global IP Traffic Forecast, 2015 - 2020," 2016.

[2] A. Gudipati, D. Perry, L. E. Li, and S. Katti, "SoftRAN: Software defined radio access network," in *ACM SIGCOMM HotSDN*, 2013.

[3] "COMBO." [Online]. Available: http://www.ict-combo.eu/

[4] S. Ren, T. Lin, W. An, G. Zhang, D. Wu, L. N. Bhuyan, and Z. Xu, "Design and analysis of collaborative EPC and RAN caching for LTE mobile networks," *Computer Networks*, vol. 93, Part 1, 2015.

[5] M. Ruiz, M. German, L. Contreras, and L. Velasco, "Big data-backed video distribution in the telecom cloud," *Comput. Commun.*, vol. 84, Jun. 2016.

[6] H. A. Pedersen and S. Dey, "Enhancing mobile video capacity and quality using rate adaptation, RAN caching and processing," *IEEE/ACM Trans. on Netw.*, vol. 24, no. 2, pp. 996–1010, Apr 2016.

[7] A. Araldo, F. Martignon, and D. Rossi, "Representation selection problem: Optimizing video delivery through caching," in *IFIP Networking*, May 2016.

[8] Y.-T. Yu, F. Bronzino, R. Fan, C. Westphal, and M. Gerla, "Congestion-aware edge caching for adaptive video streaming in information-centric networks," in *IEEE CCNC*, Jan 2015.

[9] S.-H. Shen and A. Akella, "An information-aware QoE-Centric mobile video cache," in *ACM Mobicom*, 2013.

[10] G. Carofiglio, M. Gallo, L. Muscariello, and D. Perino, "Scalable mobile backhauling via information-centric networking," in *IEEE Int. Workshop on LAN and MAN*, Apr 2015.

[11] H. Nam, D. Calin, and H. Schulzrinne, "Towards dynamic mptcp path control using sdn," in *IEEE NetSoft*, Jun. 2016.

[12] X. Corbillon, R. Aparicio-Pardo, N. Kuhn, G. Texier, and G. Simon, "Cross-layer scheduler for video streaming over MPTCP," in *ACM MMSyS*, 2016.

[13] B. Rainer, D. Posch, and H. Hellwagner, "Investigating the performance of pull-based dynamic adaptive streaming in NDN," *IEEE JSAC*, vol. 34, no. 8, pp. 2130–2140, Aug 2016.

[14] N. Choi, K. Guan, D. C. Kilper, and G. Atkinson, "In-network caching effect on optimal energy consumption in content-centric networking," in *IEEE ICC*, Jun. 2012.

[15] J. Araujo, F. Giroire, Y. Liu, R. Modrzejewski, and J. Moulierac, "Energy efficient content distribution," in *IEEE ICC*, Jun. 2013.

[16] J. Llorca, A. M. Tulino, M. Varvello, J. Esteban, and D. Perino, "Energy efficient dynamic content distribution," *IEEE JSAC*, vol. 33, no. 12, pp. 2826–2836, Dec 2015.

[17] R. Aparicio-Pardo and L. Sassatelli, "Adaptive video streaming and fixed-mobile convergence: A good team to reduce power consumption and improve users' QoE," in *IEEE ICTON*, Jul. 2016.

[18] M. Sheng, W. Han, C. Huang, J. Li, and S. Cui, "Video delivery in heterogenous CRANs: architectures and strategies," *IEEE Wireless Comm. Mag.*, vol. 22, no. 3, pp. 14–21, June 2015.

[19] R. Aparicio-Pardo and L. Sassatelli, "A cost model for green fog computing and networking," in *IEEE ICTON*, Jul. 2017. [Online]. Available: http://www.i3s.unice.fr/~raparicio/costmodel2017.pdf

[20] "VQM software." [Online]. Available: https://tinyurl.com/kdu93oc

[21] E. Baik, A. Pande, Z. Zheng, and P. Mohapatra, "VSync: Cloud based video streaming service for mobile devices," in *IEEE Infocom*, 2016.

[22] A. Devlic, P. Kamaraju, P. Lungaro, Z. Segall, and K. Tollmar, "QoE-aware optimization for video delivery and storage," in *IEEE WOWMOM*, Jun. 2015.

[23] N. Carapellese, M. Tornatore, and A. Pattavina, "Energy-efficient baseband unit placement in a fixed/mobile converged WDM aggregation network," *IEEE JSAC*, vol. 32, no. 8, pp. 1542–1551, Aug 2014.

[24] M. Savi, O. Ayoub, F. Musumeci, Z. Li, G. Verticale, and M. Tornatore, "Energy-efficient caching for video-on-demand in fixed-mobile convergent networks," in *IEEE OnlineGreenComm*, Nov 2015.

[25] N. C. Fofack, M. Dehghan, D. Towsley, M. Badov, and D. L. Goeckel, "On the performance of general cache networks," in *Valuetools*, 2014.

[26] G. Gan, C. Ma, and J. Wu, *Data Clustering*. SIAM, 2007.

[27] D. P. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE JSAC*, vol. 24, pp. 1439–1451, 2006.

[28] R. Aparicio and L. Sassatelli, "Simulator." [Online]. Available: https://github.com/sassatelli/Green_Video_Control_Plane

[29] J. Li, A. Aurelius, M. Du, H. Wang, A. Arvidsson, and M. Kihl, "Youtube traffic content analysis in the perspective of clip category and duration," in *IEEE NoF*, Oct 2013.

[30] L. Maggi, L. Gkatzikis, G. S. Paschos, and J. Leguay, "Adapting caching to audience retention rate: Which video chunk to store?" *CoRR*, vol. abs/1512.03274, 2015.

[31] Adobe, "Digital Index Q3 Digital Video Report," 2015.

[32] FCC, "Measuring Fixed Broadband Report," 2016.

[33] OpenSignal, "Global State of Mobile Networks," 2016. [Online]. Available: https://opensignal.com/reports/2016/08/global-state-of-the-mobile-network/