Energy-Efficient Caching for Video-on-Demand in Fixed-Mobile Convergent Networks

Marco Savi*, Omran Ayoub*, Francesco Musumeci*, Zhe Li†, Giacomo Verticale*, Massimo Tornatore*
*Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Italy (name.surname@polimi.it)
†JCP-Connect, Rennes, France (zhe.li@jcp-connect.com)

***Abstract*—The success of novel bandwidth-consuming multimedia services such as Video-on-Demand (VoD) is leading to a tremendous growth of the Internet traffic. Content caching can help to mitigate such uncontrolled growth by storing video content closer to the users in core, metro and access network nodes. So far, metro and especially access networks supporting mobile and fixed users have evolved independently, leveraging logically (and often also physically) separate infrastructures; this means that mobile users cannot access caches placed in the fixed access network (and vice-versa), even if they are geographically close to them, and energy consumption implications of such undesired effect must be investigated. We define an optimization problem modeling an energy-efficient placement of caches in core, metro and fixed/mobile access nodes of the network. Then, we show how the evolution towards a Fixed-Mobile Converged metro/access network, where fixed and mobile users can share caches, can reduce the energy consumed for VoD content delivery.**

I. Introduction

The Internet has become one of the major energy consumers worldwide. According to [1], if the Internet were a country, it would be the 5th biggest energy-consuming country in the world. Moreover, both fixed and mobile Internet traffic is steadily increasing [2], and network operators are constantly upgrading their infrastructure by deploying innovative fixed and mobile network access technologies, such as fiber-to-the-home (FTTH), fiber-to-the-cabinet (FTTC) or Long Term Evolution (LTE), to provide the users with higher access rates.

Internet traffic increase is mainly driven by the adoption of the broadband video-streaming services, such as Video-on-Demand (VoD), that are recognized to be the most bandwidth-consuming Internet services. Usually, video contents are stored in centralized data-centers located in the core segment of the network and far from the users. At the current pace of multimedia traffic growth and without any upgrade in the network infrastructure, the core network will soon be flooded by multimedia traffic with a consequent risk of congestion and service disruption. Such undesirable scenario can be mitigated by pushing the contents towards the users through a distributed system of *caches* in the core, metro and access networks, so that multimedia traffic can be kept local and the traffic can be partly offloaded by the core network.

On the other hand, the proliferation of caches must be carefully handled from an energy consumption perspective. Several studies, e.g., [3][4][5] show that, in general, deploying a system of caches closer to the users results in two contrasting effects. On one side, the *caching energy consumption*, i.e., the energy needed to power the caches where contents are stored, consistently increases. On the other side, the *transport energy consumption*, needed to move data through the network (e.g., data forwarding and transmission over the links) is reduced, as the content is delivered to the users from a closer location.

However, when a multimedia service can be accessed by both fixed and mobile users, the effectiveness of content caching in the metro/access network segment is hindered by the fact that, traditionally, fixed and mobile metro/access networks have been evolved and deployed independently. Thus, a cache placed in the fixed metro/access network cannot be easily reached by a mobile network user (and vice-versa), even if the user is geographically close to the cache, as the streaming flow must traverse several additional nodes and links to reach the interconnection point between the two networks. This undesired effect is also known as *trombone effect*.

Recent research studies (as, e.g., the FP7 European COMBO project [6]) aim at defining new architectures for Fixed-Mobile Converged (FMC) networks, where the fixed and mobile metro/access networks are jointly designed and optimized both from a *functional convergence* perspective (i.e., by unifying network functionalities among heterogeneous network types) and a *structural convergence* perspective (i.e., by sharing equipment and infrastructures among heterogeneous network types). Clearly, a FMC architectural solution can help in reducing the aforementioned trombone effect.

A. Related works

Our previous work in Ref. [7] defines a strategy to efficiently switch on/off caches placed in the metro/access networks according to traffic load variations. In this paper we adopt similar energy modeling for network equipment and caches, but we focus on an energy-efficient placement of caches in the network considering a static scenario. Several other works deal with optimization problems for energy-efficient placement of caches in the network. Ref. [8] is one of the first works investigating the effect of energy-efficient caching by adding caching capabilities to the core nodes. Also Ref. [9] exploits caching in core nodes, focusing on the definition of efficient strategies to switch off caches and links. We adopt similar approaches, but focus especially on caching in metro and access nodes, and investigate on the impact of FMC in reducing the network energy consumption. Ref. [10] defines an energy-efficient strategy for cache location and content dissemination, considering a logical tree caching hierarchy

towards the users. With respect to [10], we also consider a caching hierarchy, which depends on the physical network topology and not on an overlay logical network. Finally, Refs. [11][12][13] focus on caching in a Content Centric Networking (CCN) scenario. Refs. [11][12] show how a CCN approach can be energy-efficient, especially when popular contents are delivered. Though we do not focus on CCN, these works are relevant as we gather the assumption that every network equipment can be equipped with storage capabilities. Ref. [13] evaluates the impact in terms of performance of shared caching in FMC networks with respect to non-convergent networks. Our assumptions are similar, but we focus on an energy consumption evaluation.

B. Paper Contribution

In this paper we model the problem of energy-minimized cache and content placement and VoD requests routing in FMC networks using an Integer Linear Program (ILP). Then we evaluate the benefits provided by fixed-mobile convergence in reducing the energy consumption of content delivery networks for the specific case of video streaming.

The paper is organized as follows. Section II describes the network and traffic models as well as the energy contributions considered in our evaluation. In Sec. III we introduce the ILP model which used to minimize the overall energy consumption of the VoD content delivery. In Sec. IV we discuss the numerical results. Finally, Sec. V concludes the paper.

II. Network, Traffic and Energy Models

A. Network model

We consider a hierarchical network topology spanning over three segments, as depicted in Fig. 1:

- The *core* segment, consisting of *core routers* interconnected in a mesh topology and connected to *data centers* hosting video servers.
- The *metro* segment, consisting of *aggregation switches* in a ring topology. Each metro segment is connected to the core segment through an *edge router*. We consider the edge routers as part of the metro segment.
- The *access* segment, consisting of *access nodes* connected to multiple users. Access nodes act as source of VoD requests. They can be OLTs or DSLAMs for fiber and copper fixed access networks, and eNodeBs for LTE mobile access networks. Note that some access nodes can be directly connected to edge routers.

All the network devices (routers, switches and access nodes) can be equipped with some storage capacity to perform *caching* of content. Specifically, Solid State Drives (SSDs) are used in core and edge routers, Dynamic Random Access Memory (DRAMs) are used in aggregation switches and Reduced Latency DRAMs (RLDRAMs) are used in OLTs and DSLAMs. These assumptions come from the fact that network nodes closer to the users host storage devices with smaller capacity and footprint with respect to nodes far from the users, since the space for the additional storage equipment is more limited (e.g., Cabinets, where DSLAMs or OLTs can be located, are smaller than Central Offices, where aggregation switches are usually located).

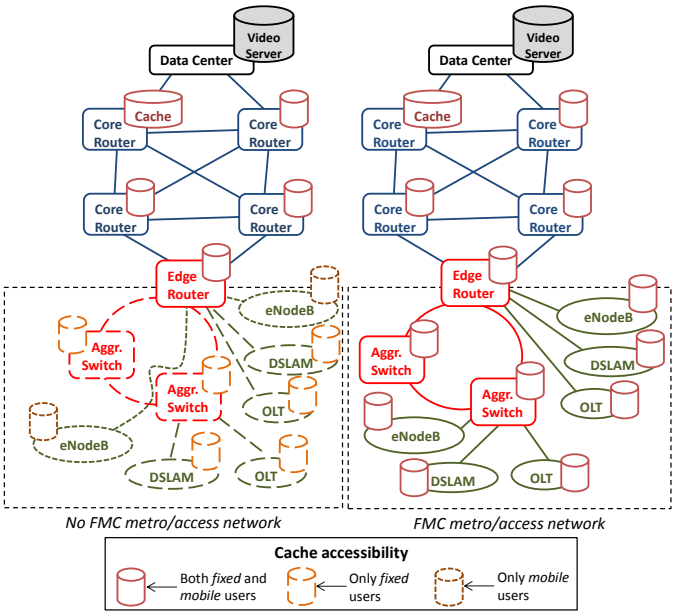


Fig. 1. Examples of *No FMC* network architecture (left) and of *FMC* network architecture (right) spanning over the access (green), metro (red) and core (blue) segments. All the transmission devices can be equipped with caches.

For the metro and access segments, we focus on the two following network architectures, as shown in Fig. 1.

1) No FMC metro/access network architecture (No FMC): This network architecture, as depicted in the left-hand side of Fig. 1, models the current mode of operation of the metro and access segments. The fixed and mobile metro/access networks are mostly functionally and structurally independent. Indeed, it can happen that the mobile network uses optical fibers of the fixed network for mobile backhauling and thus a certain degree of structural convergence exists, but the traffic is just tunnelled and cannot be accessed, since no functional convergence is provided. In this scenario, we assume that *edge routers* are the first interconnection point between the fixed and mobile networks. Therefore, caches placed in the fixed (mobile) metro and access network can only be accessed by fixed (mobile) network users to avoid trombone effect. Conversely, caches placed in the edge routers and in the core routers can be accessed by both fixed and mobile users.

2) FMC metro/access network architecture (FMC): This network architecture is depicted in the right-hand side of Fig. 1. In this case, we consider a fully structural and functional FMC network [6]. If such FMC solution is considered, multiple heterogeneous access network technologies are deployed only in the peripheral areas of the access network, while the metro/access segments are fully shared by the fixed and mobile users. This means that a cache placed in a mobile access node can be accessed by a fixed user by only traversing one or more *aggregation switches*, which are shared among the heterogeneous fixed and mobile networks, and vice-versa. Thus, in a FMC solution no trombone effect must be mitigated, since all the caches are shared by fixed and mobile users and both fixed and mobile VoD traffic are kept as local as possible.

TABLE I
VIDEO RESOLUTIONS AND BITRATES FOR THE VIDEO CONTENTS

Video Quality	Video Resolution	Bitrate R (Mbit/s)
SD	640x460	1.3
HQ	960x540	2.8
Mobile	960x640	3.3
HD	1280x720	4.9

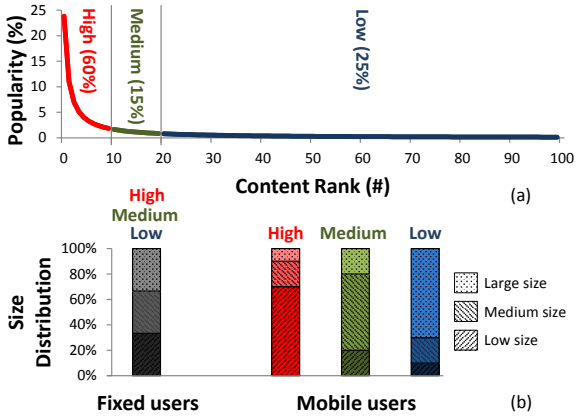


Fig. 2. (a) Content popularity distribution and (b) VoD content size distribution for fixed and mobile users.

B. Traffic model

Video contents and VoD requests are modeled as follows:

1) Video content model: Each content is described by *i)* its popularity and *ii)* its size (byte). Concerning the content popularity several studies, e.g. [14][15], show how video streaming contents have a unbalanced popularity distribution, where around 80% of content requests are for the 20% most popular contents. This behaviour can be modeled by a *Zipf distribution*. Considering a set M of contents, where $m = 1$ is the most popular content and $m = |M|$ is the least popular content, the probability that the content $1 \leq m \leq |M|$ is requested by a user is defined by the probability density function $h(m) = K/m^{\phi}$, where K is a normalization constant and ϕ is the Zipf distribution parameter. In Fig. 2(a) we show the popularity distribution assumed in this work: considering a total (i.e., content rank) of 100 contents[1], we assume that the 20 most-popular contents account for an overall 75% of popularity (60% of global popularity corresponds to the 10 most popular contents). On the other hand, the 80 least-popular contents are characterized by an overall 25% of popularity. As for the size of the contents, we consider three different categories: low, medium and large size video contents, with file-size of 200 MB, 1 GB and 2 GB respectively. We assume that fixed and mobile users content popularity behaves in a different way: mobile users require on average more low-size video contents than medium- or large-size ones, while fixed users tend to uniformly choose among the three categories. In Fig. 2(b) we show, for the three categories, the video size distribution for both fixed and mobile user requests[2].

[1]Investigations considering a larger set of VoD contents is left for future work, as we are developing heuristic algorithm to deal with scalability issues.

[2]Note that our assumption is reasonable as in general mobile users are more bandwidth-consumption aware than fixed users, since they usually must deal with very strict weekly or monthly caps on their Internet traffic. Moreover, large-size videos (e.g., long movies) are unlikely to be watched via mobile devices such as smartphones or tablets.

TABLE II
ENERGY CONSUMPTION OF NETWORK EQUIPMENT [8] AND ADOPTED STORAGE TECHNOLOGY

Transmission Device	Energy Consumption E^{sw} (J/bit)	Storage Technology
Core Router	$1.7 \cdot 10^{-8}$	SSD
Edge Router	$2.63 \cdot 10^{-8}$	SSD
Aggr. Switch	$8.21 \cdot 10^{-9}$	DRAM
OLT	$1.92 \cdot 10^{-8}$	RLDRAM
DSLAM	$1.4 \cdot 10^{-7}$	RLDRAM
eNodeB	$2 \cdot 10^{-6}$	RLDRAM

TABLE III
POWER CONSUMPTION OF STORAGE TECHNOLOGIES AND THEIR STORAGE CAPACITY [12]

Storage Technology	Baseline Power P^{base} (W)	Load-Dependent Power P^{st} (W/bit)	Capacity S (GB)
SSD	5	$6.25 \cdot 10^{-12}$	200
DRAM	3.5	$2.5 \cdot 10^{-9}$	50
RLDRAM	3.5	$3.75 \cdot 10^{-9}$	20

2) VoD request model: Every VoD request is characterized by *i)* the requested content, *ii)* the bitrate for the requested content (bit/s) and *iii)* the access node requesting the content. A fixed (mobile) user requests a specific content according to the fixed (mobile) popularity content distribution. We assume that a scalable coding technique is used to encode each video content and that different users can request the same content at different bitrates. One over four possible bitrates can be associated to a VoD request: Tab. I shows the possible bitrates and the respective video resolutions. We assume that mobile users always request a content at *mobile* resolution, as usually they access contents through a mobile device. Conversely, VoD requests of fixed users are assumed uniformly distributed among the four bitrates shown in Tab. I, since fixed users access the content through heterogeneous devices (e.g., mobile devices connected to WiFi networks, TVs, laptops etc.).

C. Energy models

We model the *transport* and the *caching* energy consumption contributions as follows [7].

- Transport energy consists of *Switching* energy contribution and *Link transmission* energy contribution. The former is due to the switching operations performed by the backplane of switches and routers to forward traffic from/to an incoming/outgoing network interface. We consider this contribution as proportional with respect to the amount of data processed by the network device, as reported in Tab. II, where the switching *energy per bit* E^{sw} of various network equipment is reported. The link transmission contribution is due to network interfaces of switches and routers that are active and transmit data. An active network interface consumes a fixed amount of energy independently from the actual transmitted traffic. We assume that all network devices are able to "pack" the traffic in the minimum number of needed interfaces. We assume the usage of Ethernet network interfaces of capacity $C = 1$ Gbit/s consuming $P^{int} = 1$ W.
- Caching energy is composed by a load-independent *baseline* contribution, P^{base}, needed to power-on one cache, and a traffic-dependent *Storage* contribution, P^{st}, which

is proportional to the amount of data stored in the cache. The values for P^{base} and P^{st} for various storage technologies are reported in Tab. III.

III. ILP MODEL FOR ENERGY-EFFICIENT CACHING

The ILP problem is stated as follows. **Given** a physical fixed-mobile network topology (as in Fig. 1) and a set of VoD requests, we **decide** the optimal cache and content placement as well as VoD request routing in order to **minimize** the overall network energy consumption. We refer this ILP-based optimization to as *Intelligent Caching* strategy.

A. Sets and parameters

- $\mathcal{G} = (N, A)$ is the graph used to model the physical network topology, where N represents the set of nodes and A the set of bidirectional links.
- The subsets $F, T \subseteq N$ ($T \cup F = N$) represent the set of nodes with forwarding and caching capabilities used to serve VoD requests (we will generically refer to such nodes as *caches*), and the set of terminal nodes, which act as destination for VoD traffic, respectively.
- The subsets $T_{mob}, T_{fix} \subseteq T$ ($T_{mob} \cup T_{fix} = T$, $T_{mob} \cap T_{fix} = \oslash$), are the sets of fixed (OLTs or DSLAMs) and mobile (eNodeBs) terminal nodes, respectively.
- The subsets $F_{fix}, F_{mob}, F_{fixmob} \subseteq F$ ($F_{fix} \cup F_{mob} \cup F_{fixmob} = F$), represent the set of caches which can be accessed by a fixed user, a mobile user, or from any user, respectively.
- The storage capacity of a generic cache $f \in F$ is denoted as S_f. As described in Sec. II P_f^{base} and P_f^{st} represent the baseline and the load-dependent cache power contributions, respectively. E_f^{sw} is the switching energy contribution.
- M is the set of contents m, each with size B_m and popularity $h(m)$, and $\mathbb{Z}$ is the set of content requests, where each $Z_t^m \in \mathbb{Z}$ represents the number of requests for content $m \in M$ originating in $t \in T$. Every VoD content requested by a terminal node $t \in T$ is associated to a bitrate R_t, chosen among the values in Tab. I.
- C is the capacity of one interfaces, consuming a fixed amount of power P^{int}.
- δ is a time-interval normalization factor used to evaluate the storage and interfaces energy consumption given the corresponding power consumption values (P_f^{base}, P_f^{st} and P^{int}).

B. Decision variables

- x_f (binary) is used to indicate whether the cache $f \in F$ is used ($x_f = 1$) or not ($x_f = 0$).
- l_{ij} (integer) represents the number of active interfaces used in link $(i, j) \in A$.
- $u_{ij}^{m,t}$ (binary) is equal to 1 iff the link $(i, j) \in A$ is used to transmit content $m \in M$ requested by terminal node $t \in T$.
- k_{ij} (integer) represents the amount of data transported on link $(i, j) \in A$. Note that this is an auxiliary variable used to improve the clarity of the ILP model description, and is defined as $k_{ij} = \sum_{\substack{t \in T \\ m \in M}} z_t^m R_t u_{ij}^{m,t}$.
- v_f^m (binary) is equal to 1 iff the cache $f \in F$ is storing content $m \in M$.
- $w_f^{m,t}$ (binary) is equal to 1 iff the cache $f \in F$ satisfies the VoD requests of terminal node $t \in T$ for the content $m \in M$.

C. Objective function

$$\min \sum_{f \in F} P_f^{base} \delta x_f + \sum_{\substack{f \in F \\ m \in M}} B_m P_f^{st} \delta v_f^m + \sum_{\substack{f \in F \\ i \in F:(i,f) \in A \\ j \in F:(f,j) \in A}} E_f^{sw}(k_{if} + k_{fj}) + \sum_{(i,j) \in A} P^{int} \delta l_{ij} \quad (1)$$

The first two terms of the objective function account for the *baseline* energy consumption and the *storage* energy consumption of the caches (i.e., the *caching* energy consumption). The third and the fourth terms refer to the *switching* and *link transmission* energy consumption (i.e., to the *transport* energy consumption). The objective of the optimization is to minimize the overall energy consumption. Note that the normalization factor δ is needed to have homogeneous contributions in the objective function. As the switching energy consumption contribution refers to the absolute amount of data switched by each switching device in a given time-interval (and not on the device power consumption), δ can be seen as the average holding time for each VoD request.

D. Constraints

$$\sum_{m \in M} v_f^m B_m \leq S_f x_f \quad f \in F \quad (2)$$

Eqn. 2 guarantees that the total amount of video content data stored by the cache f does not exceed its capacity and that no data is stored by the cache if the cache is powered-off.

$$w_f^{m,t} \leq v_f^m \leq x_f \quad f \in F, t \in T, m \in M \quad (3)$$

Eqn. 3 assures that a cache f cannot satisfy a VoD request from a terminal node t unless it is powered on and it stores the requested content m.

$$\sum_{f \in F} w_f^{m,t} = 1 \quad t \in T, m \in M \quad (4)$$

Eqn. 4 assures that every VoD request for a content m by a terminal node t is satisfied by exactly one cache f.

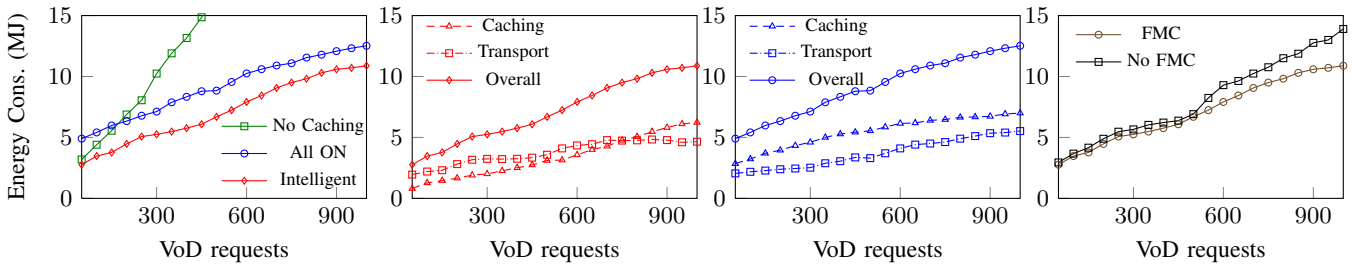
$$l_{ij} \geq k_{ij}/C \quad (i, j) \in A \quad (5)$$

Eqn. 5 is used to calculate the number of active network interfaces on link (i, j).

$$\sum_{\substack{i \in F: \\ (i,f) \in A}} k_{if} + \sum_{\substack{t \in T \\ m \in M}} Z_t^m R_t \delta w_f^{m,t} = \sum_{\substack{j \in F: \\ (f,j) \in A}} k_{fj} \quad f \in F \quad (6)$$

$$\sum_{i \in F:(i,t) \in A} k_{it} = \sum_{m \in M} Z_t^m R_t \delta \quad t \in T \quad (7)$$

We set the flow balancing constraints for the caches f (eqn. 6) and the terminal nodes t (eqn. 7). Eqn. 6 refers to the nodes f, that can both generate traffic and forward traffic coming from other nodes. Eqn. 7 refers to the terminal nodes t, that are the destination of video traffic generated by the caches to accommodate the VoD requests.



(a) *FMC*: *No caching* vs. *All ON* vs. *Intelligent caching* (b) *FMC*: *Intelligent caching* (c) *FMC*: *All ON* (d) *Intelligent caching*: *FMC* vs. *No FMC*

Fig. 3. Energy consumption as a function of the number of VoD requests per terminal node in case of different caching strategies (*No Caching*, *All ON*, *Intelligent Caching*) for the *FMC* architecture (a). (b) and (c) show the caching and transport energy consumption for the *Intelligent caching* and *All ON* strategies, respectively. (d) compares the *Intelligent caching* strategy for the *FMC* and *No FMC* network architectures.

From a methodological point of view, the fixed-mobile convergence is captured by considering a *unique* (integrated) physical topology, where all caches can be accessed by all terminal nodes. However, two additional sets of constraints are needed when *No FMC* architecture is considered.

$$w_f^{m,t} = 0 \quad f \in F_{fix}, t \in T_{mob} \tag{8}$$

$$w_f^{m,t} = 0 \quad f \in F_{mob}, t \in T_{fix} \tag{9}$$

Eqn. 8 and 9 assure that caches in the fixed network cannot be accessed by mobile terminal nodes and vice-versa.

IV. Numerical Results

A. Case study

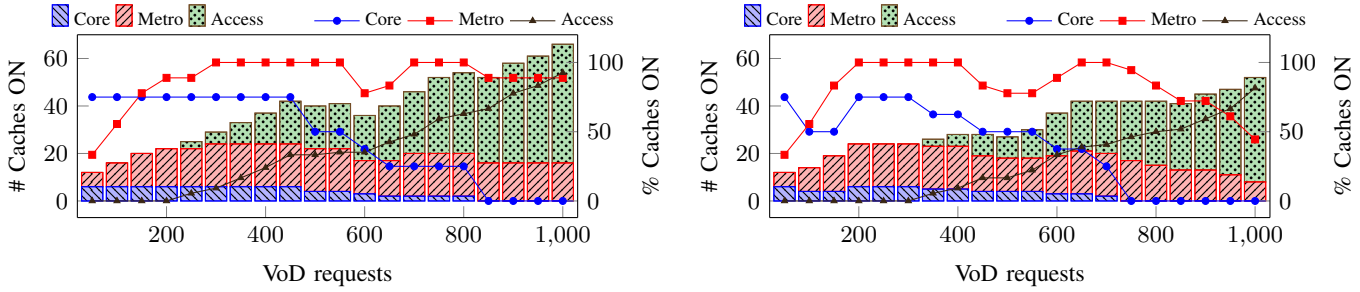
We consider a network topology similar to the one shown in Fig. 1. The core segment is a mesh network composed by 8 core routers. The metro segment consists of 6 rings, each with 3 metro nodes, where aggregation switches are interconnected to the core network through edge routers. The access segment consists of 54 access nodes, each connected to an aggregation switch or an edge router. In the *FMC* architecture both fixed and mobile access nodes can be connected to an aggregation switch, whereas in the *No FMC* architecture mobile access nodes are always directly connected to an edge router. We consider a content catalogue of $|M| = 100$ contents, whose popularity is Zipf-like distributed ($\phi = 1.1$) and behaves in a different way for fixed and mobile users, as shown in Fig. 2 and described in Section II. We focus on a time-frame δ of 1 hour. To obtain numerical results we have used ILOG CPLEX 12.4 on a workstation equipped with 8×2.00 GHz processors and with 32 GB of RAM. In most cases, optimal results are obtained in computational time in the order of minutes.

The ILP-based *Intelligent Caching* strategy is compared with two benchmark strategies: *i) No Caching*, where a centralized video server in the core network stores and delivers all the contents; *ii) All Caches ON* (*All ON*), where all the caches are assumed as powered-on.

B. Discussion

Our evaluation focuses on the energy consumption variation when the number of VoD requests increases. We first focus on the *FMC* scenario. Figure 3(a) shows the overall energy consumption of the three caching strategies. The *Intelligent caching* strategy always outperforms both the *No caching* and the *All ON* strategies, proving that powering-on only selected caches always allows to save energy. Also, note that the *All ON* strategy in general outperforms the *No Caching* one. However, under very low traffic conditions *No Caching* strategy is more energy-efficient than *All ON*, as the transport energy in the *No caching* scenario, spent to deliver few contents to a low number of users from the video-server in the core network, is lower than the penalty introduced in the *All ON* case to power-on caches in all network nodes. Figure 3(b) divides the contribution of caching and transport energy consumption on the overall energy consumption when the *Intelligent caching* strategy is considered. For a low number of VoD requests the transport energy is higher than the caching energy, but the caching energy curve has a much steeper increase, and in fact, for about 800 VoD requests the caching energy overcomes the transport energy. This is due to the fact that, for higher traffic loads, contents are mostly served by a high number of caches closer to the users (i.e., in the access) and many network interfaces in the higher segments (i.e., core and metro) can be switched off; conversely, for lower traffic, most of the contents are served by a small number of caches in the core, so caches closer to the users are not needed. Figure 3(c) shows the same energy contributions of Fig. 3(b) for the *All ON* strategy. In this case, the caching energy consumption is always dominant with respect to the transport energy consumption and, in average, slightly higher than in the *Intelligent Caching* strategy of Fig. 3(b), since all the caches are always powered-on. On the other hand, the transport energy consumption in the *All ON* strategy is lower than in the *Intelligent Caching* strategy. Note also that in the *No Caching* strategy the overall energy consumption is just equal to the transport energy consumption, as there is no caching energy consumption.

We now concentrate on the energy consumption comparison between the *FMC* and *No FMC* architectures (Fig. 3(d)), considering only the *Intelligent Caching* strategy. The *FMC* architecture is more energy-efficient with respect to the *No FMC* one especially for a higher load, i.e., when it is more convenient to power-on the caches located in the access nodes, as in the *FMC* architecture they can be accessed by both fixed and mobile users. Figures 4(a) and 4(b) show the number of caches that are powered-on in each network segment (core/metro/access) when *Intelligent Caching* is performed,



(a) No FMC: *Intelligent caching*

(b) FMC: *Intelligent caching*

Fig. 4. Number/percentage of caches that are powered-on per each segment (core/metro/access) as a function of the number of VoD requests per terminal node, considering the *Intelligent caching* strategy for both the *No FMC* (a) and *FMC* (b) network architectures.

in absolute terms (histograms) and as a percentage of the total number of caches per each segment (curves, referred to the secondary y-axes on the right), for the *No FMC* and *FMC* scenarios, respectively. As expected, the total number of powered-on caches increases with the number of VoD requests. In general, in the *FMC* case a lower number of caches is needed in comparison to the *No FMC* case, especially at higher traffic loads. This demonstrates the energy benefit provided by the opportunity of sharing caches for fixed and mobile users in the *FMC* scenario. In both cases, as the number of VoD requests increases, we observe a "migration" of the used caches from the core to the access segment. In fact, when more requests need to be provisioned the transport energy contribution becomes more relevant, so its effect is mitigated by pushing the contents closer to the users. Note that the percentage of powered-on caches in the metro segment behaves differently for the *No FMC* and *FMC* architectures. For low traffic load, in both the architectures the percentage of powered-on caches increases for increasing traffic, until 90-100% of metro caches are powered-on. Then, for higher loads and only in the *FMC* case, the number of powered-on caches in the metro segment starts decreasing. This is due to the fact that in the *FMC* scenario, energy benefits are obtained by serving both fixed and mobile users through the same shared caches, especially from those placed in the access segment.

V. Conclusion

We provided an energy-aware cache and content placement strategy that allows to energy-efficiently power-on and -off caches located in core, metro and access network equipment, as well as routing VoD requests, according to traffic load conditions. We evaluated the effectiveness of this strategy on two different network architectures: a *FMC* architecture, where fixed and mobile users can access all the caches deployed in the network, and a non converged network architecture (the current mode of operation, i.e., *No FMC*), where no cache sharing is allowed in the metro and access network segments. We confirmed that, in general, transport energy consumption has higher impact with respect to caching energy consumption as seen in previous works. Indeed, powering-on all the caches distributed in the network allows to save energy with respect to retrieving content from a centralized location, except for very low traffic load conditions. We demonstrated that the proposed strategy allows always to save energy in comparison to the cases where all the caches are always powered-on or all the contents are retrieved from a centralized video-server location, as it allows to better manage the trade-off between caching and transport energy consumption. With the provided strategy we are able to power-on only selected caches in the network, typically pushing the contents closer to the users when the traffic load is higher. Also, for the first time, we qualitatively showed the impact of *FMC* on network energy-efficiency, in comparison to *No FMC* scenarios, demonstrating that the structural and functional convergence provided by the *FMC* approach is also beneficial for VoD content delivery.

Acknowledgment

The research leading to these results has received funding from the European Community Seventh Framework Programme FP7/2013-2015 under grant agreement no. 317762 COMBO project.

References

[1] G. Cook *et al.*, "How dirty is your data? A look at the energy choices that power cloud computing," *Greenpeace International*, Mar. 2011.
[2] "Cisco visual networking index: Global mobile data traffic forecast update, 2014-2019," May 2015.
[3] K. Hinton *et al.*, "Power consumption and energy efficiency in the Internet," *IEEE Network*, vol. 25, no. 2, pp. 6–12, Mar. 2011.
[4] C. Jayasundara *et al.*, "Improving energy efficiency of Video on Demand services," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 3, no. 11, pp. 870–880, Nov. 2011.
[5] J. Baliga *et al.*, "Architectures for energy-efficient IPTV networks," in *Conference on Optical Fiber Communication (OFC)*, Mar. 2009.
[6] S. Gosselin *et al.*, "Fixed and mobile convergence: Needs and solutions," in *European Wireless Conference*, May 2014.
[7] M. Savi *et al.*, "Energy-efficient VoD content delivery and replication in integrated metro/access networks," in *IEEE Latincom*, Nov. 2014.
[8] U. Mandal *et al.*, "Energy-efficient content distribution over telecom network infrastructure," in *ICTON*, Jun. 2011.
[9] J. Araujo *et al.*, "Energy efficient content distribution," in *IEEE ICC*, Jun. 2013.
[10] S. Imai *et al.*, "Energy efficient content locations for in-network caching," in *IEEE APCC*, Oct. 2012.
[11] K. Guan *et al.*, "On the energy efficiency of content delivery architectures," in *IEEE ICC Communications Workshops*, Jun. 2011.
[12] N. Choi *et al.*, "In-network caching effect on optimal energy consumption in content-centric networking," in *IEEE ICC*, Jun. 2012.
[13] Z. Li *et al.*, "ICN based shared caching in future converged fixed and mobile network," in *IEEE HPSR*, Jul. 2015.
[14] W. Tang *et al.*, "Medisyn: A synthetic streaming media service workload generator," in *ACM NOSSDAV*, Jun. 2003.
[15] H. Li *et al.*, "Video requests from online social networks: Characterization, analysis and generation," in *IEEE INFOCOM*, Apr. 2013.