# Caching Placement Strategies for Dynamic Content Delivery in Metro Area Networks

Omran Ayoub*, Francesco Musumeci*, Christian Addeo†, Marco Mussini† and Massimo Tornatore*
*Politecnico di Milano, Department of Electronics, Information and Bioengineering, Milan, Italy
†SM-Optics, Vimercate (MB), Italy
*Email:{*firstname.lastname*}@polimi.it †Email:{*firstname.lastname*}@sm-optics.com

***Abstract*—Video-on-Demand (VoD) traffic explosion has been one of the main driving forces behind the recent Internet evolution from a traditional connection-centric architecture towards the new content-centric paradigm. To cope with this evolution, caching of VoD contents closer to the users in core, metro and even metro-access optical network equipment is regarded to be a prime solution that could help mitigating this traffic growth. However, the optimal caches placement and dimensioning is not univocal, especially in the context of a dynamic network, as it depends on various parameters, such as network topology, users behavior and content popularity. In this paper, we focus on a dynamic VoD content delivery scenario in a metropolitan network implementing different caching strategies. We evaluate the performance of the various caching strategies in terms of network-capacity occupation showing the savings in resource occupation in each of the network segments. We also evaluate the effect of the distribution of the storage capacity on the overall average number of hops of all requests. The obtained numerical results show that, in general, a significant amount of network resources can be saved by enabling content caching near to end-users. Moreover, we show that blindly providing caching capability in access nodes may result unnecessary, whereas a balanced storage distribution between access and metro network segments provides the best performance.**

***Index Terms*—Caching placement strategies; cache deployment; video-on-demand delivery; content caching.**

## I. Introduction

Fast data proliferation has been a main driving force behind the recent Internet evolution. According to the Cisco Visual Networking Index, global IP traffic will have an annual growth rate of 22% till 2020 [1]. Moreover, the recent success of novel bandwidth-hungry multimedia services such as Video-on-Demand (VoD) has caused further challenges for an efficient capacity utilization. As a matter of fact, Cisco predicts VoD to represent approximately 78% of the global consumer traffic by 2019 and as well 80% of global mobile data traffic by 2020 [1]. With such a growth, the Internet network architecture is shifting from its traditional host-centric (connection-centric) architecture, based on named hosts, to a content-centric (information-centric) architecture, based on named data objects, i.e., videos or, in general, contents [2]. Unfortunately, current and future networks mostly focus on increasing the capacity and improving the connection capability, whereas a new architectural solution is urgently needed to efficiently distribute the high amount of video contents over the network.

A promising solutions consists of equipping edge network nodes with storage and computing capabilities [3], and enabling them, through Network Function Virtualization (NFV) and cloud-computing paradigms, to terminate services locally and to offload traffic of the core network [2]. As a consequence, the opportunity of terminating services locally, e.g., from the metro-network segment, gained vast attention from service and content providers and network operators as well. In particular, VoD delivery, being one of the most bandwidth-demanding services, gained extra attention to be terminated from nodes hosting video contents (i.e., caches) close to end-users. However, the placement of caches in the network, their number and storage capacities remain ambiguous, as they heavily depend on many decisive factors, such as network topology, users behavior and contents characteristics.

### A. Overview and Related Work

A Content Delivery Network (CDN) is a network that duplicates, stores and distributes contents from a distributed set of storage units, i.e., the *caches*, typically located across the optical metro and access network nodes, to avoid that all users requests are provisioned through the origin servers, usually located at the Internet Service Provider Points of Presence [4]. This process results in many advantages such as a decreased origin server load, improved user experience, due to reduced latency, and lower network bandwidth usage. Recently, a new trend has been to deploy caches in the metro and access segments[1], thus pushing contents closer to user premises [6], [7]. A main advantage of this technique is the reduction of the overall capacity utilization of the network, as a high number of requests is being served from locations close to end-users. Another benefit is the improved user experience as latency decreases remarkably allowing for an optimal quality of experience.

Several studies have investigated the trade-offs of cache deployments in CDNs and as well the performance evaluation of content caching in CDNs. As an example, Ref. [8] investigates the cache deployment problem determining how much server, energy and bandwidth resources are needed to provision in each cache deployed with the aim of minimizing the total cost incurred by the CDN. In addition, Ref. [9] focused on decreasing the overall network energy consumption by deploying caches in the core network and switching

[1]clearly, such an approach has a high capital cost but it guarantees a return on investment after 2 years due to savings in transport operational expenditure [5].

off links. In Refs. [10] and [6], authors went further by defining an in-network caching models for energy-efficient content distribution in metro and access networks. Unlike the above mentioned works, in this paper we focus on the online problem, where dynamically arriving users VoD requests are provisioned, and consider a more realistic VoD scenario, in terms of content catalog size, number of requests, request bit-rates and duration. Moreover, Ref. [11] evaluates the impact in terms of performance of shared caching in fixed-mobile convergent networks with respect to non-convergent networks. In addition, Ref. [12] proposes a cache replacement algorithm in a hierarchical network to minimize the Internet bandwidth but without considering dynamic traffic. We follow a similar approach but in addition to the placement (i.e., the location) of caches in the access and the metro segments, we elaborate more on the number of caches and their size dimensioning while considering a maximum overall amount of storage capacity allowed under a dynamic VoD delivery scenario.

### B. Paper Contribution

With respect to previous literature, in this paper, we model a dynamic VoD content distribution scenario implementing different caches placement strategies, where cloud-enabled edge-nodes, i.e., nodes with computing and storage capabilities, host and deliver video contents, in a metro-area network. We evaluate the performance of the caching strategies in terms of network occupation. Numerical results quantify the savings in network resources due to different caches placement strategies in a metro-area network. Furthermore, we present a thorough evaluation of different caching scenarios considering a maximum overall amount of storage capacity to be utilized while varying the size and number of caches and the popularity distribution. Numerical results show that a balanced distribution of the storage capacity between caches of the access and metro segments achieves better performance than placing caches only in the access segment.

The rest of the paper is organized as follows. In Sec. II we describe our models for the network architecture and the VoD requests used in our work. Sec. III presents the problem statement and shows how the dynamic provisioning/deprovisioning of the VoD requests is performed. In Sec. IV we present the different caching scenarios considered, whereas in Sec. V we describe the settings of the simulations of a realistic VoD content distribution scenario and discuss the numerical results. In Sec. VI we conclude the paper.

## II. Network and VoD Models

### A. Network Model

In this study we consider a metro-area network consisting of different type of cloud-enabled edge-nodes in a topology spanning over four segments (as depicted in Fig. 1):

- The *core* segment, consisting of *Metro-Core Nodes (MCNs)* connected to *data centers* hosting video servers.
- The *metro-core* segment, consisting only of *Metro-Core Nodes (MCNs)* interconnected in a ring topology.

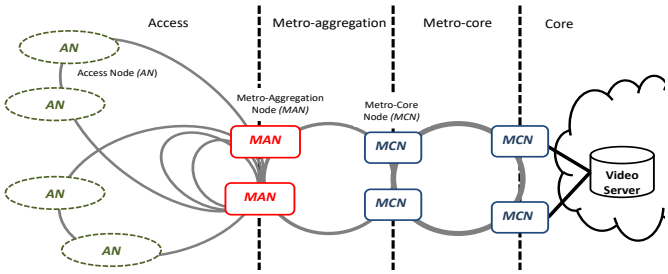



Fig. 1. The network topology considered in our study.

- The *metro-aggregation* segment, consisting of *Metro Aggregation Nodes (MAN)* and *MCNs* interconnected in a ring topology. Metro-access rings are connected to the metro ring through the *MAN*.
- The *metro-access* segment, consisting of *Access Nodes (ANs)* interconnected in a ring topology, where each *AN* represents aggregated users.

The cloud-enabled edge-nodes are the ANs, the MANs and the MCNs which could be equipped with computing capabilities and storage capacity to perform *caching* of content, depending on the caching scenario. The caching technology is assumed to be independent from cache location.

### B. Video-on-Demand Model

The video contents and the VoD requests are modeled as follows:

*1) VoD Content Model:* Each content is described by *i)* its popularity, *ii)* its size (byte) and *ii)* its duration. Concerning the VoD content popularity, several studies, such as [8] and [13], show how the popularity of video streaming follows a $Zipf$ distribution characterized by a long-tail and a small head, where around 80% of content requests are for the 20% most popular contents. The fact that VoD popularity distribution is characterized in this manner motivates caching popular contents near end-users, as storing a small amount of popular contents closer to users is sufficient to serve high amount of the VoD content requests. As an example, 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 $h(m) = K/m^{\alpha}$, where $K$ is a normalization constant and $\alpha$ is the *Zipf* distribution parameter set at 1. As for the size of the contents, it ranges from a 2 GB size video (e.g., a standard definition TV-series episode) to 14 GB size (e.g., a high definition movie) [14].

*2) VoD request Model:* Every VoD request is characterized by *i)* the requested content ($m_r$), *ii)* the bit-rate ($b_r$) for the requested content (Mbps), *iii)* the duration of the request ($t_r$) and *iv)* the destination node ($D_r$) of the content requested, which is the end-user. More specifically, bit-rate $b_r$ can be chosen around 3, 6 or 9 Mbps depending on the type of the video resolution requested. Note that the bit-rate requested does not affect the duration of the request while the overall

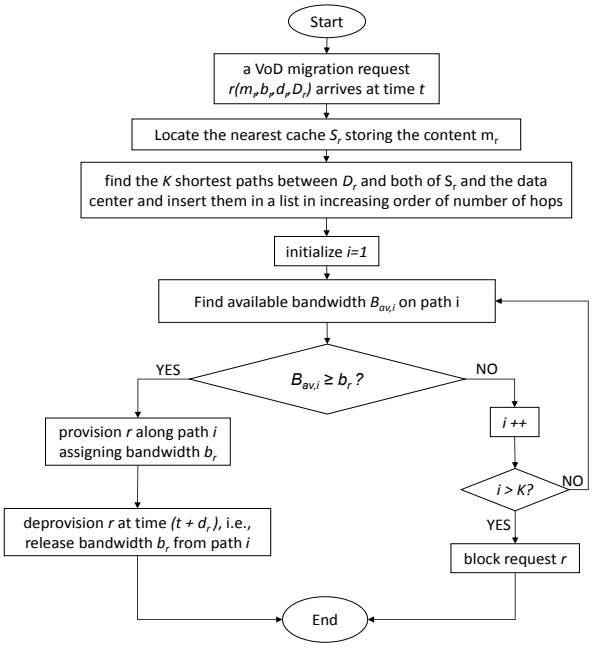



Fig. 2. Flow chart of the provisioning/deprovisioning of a VoD request

data transmitted may vary according to the bit-rate chosen[2]. As an example, a content of duration of 5400 seconds viewed with a Standard Definition (3 Mbps), or with Half High Definition (6 Mbps) or with Full HD (9 Mbps) results in a total amount of data transferred of approximately 2 GB, 4 GB and 6 GB respectively.

## III. Dynamic VoD Request Provisioning

In Fig. 2 we show the flow-chart of the VoD request provisioning/deprovisioning process. Upon the arrival of a VoD request $r : (m_r, b_r, d_r, D_r)$, from a user $D_r$ at time instant $t$, a cache $S_m$ storing the requested content $m_r$ and the data center are located. $b_r$ represents the requested bit-rate whereas $D_r$ represents the destination of the request (which is here the user) whereas $d_r$ is the duration of the content requested. Then, we apply anycast routing and find the $k$-shortest paths towards the nodes where the content is placed.

Starting from the shortest path, say path $i$, the available bandwidth $B_{av,i}$ is found and compared to the requested bit-rate $b_r$. If enough bandwidth is available on the path, request $r$ is provisioned on the path for the duration of the content requested, $d_r$. Finally, the VoD request is deprovisioning at time $t + d_r$ deallocating bandwidth $b_r$ from path $i$. If no path is found with enough available bandwidth, the VoD request is blocked.

## IV. Caching Scenarios

In our study, we implement and compare 5 different caching scenarios that differ in the placement of the caches in the network and their number. Each cache has a storage capacity of 8 TB, which makes up approximately 10% of the content catalog. In other words, each node equipped with a cache has the ability to store the most popular 10% of the content catalog. The 5 caching scenarios are modeled as follows:

- *No Caching.* The *No Caching* scenario serves as a benchmark, where no nodes in the network act as caching nodes. In this case, all VoD contents requested are served from the video server located in the core network so the VoD content spans all the network segments to reach the end-user.
- *Caching at MCNs.* In this scenario, the *MCNs* perform caching of the most popular 10% of contents. The *MCN* serves users requesting contents cached in its storage capacity. Otherwise, the video data center handles the request.
- *Caching at MANs.* This scenario is similar to the previous one, but in this case caches are located at the MANs.
- *Caching at ANs.* In this caching scenario the popular video contents are pushed even closer to end-users and are stored in caches located at the ANs. In this caching placement strategy, the number of caches deployed in the network is higher but the popular contents that are stored in the ANs are served directly and do not traverse the network.
- *Caching at ANs & MANs.* In this case, ANs and MANs are equipped with caches. The ANs store the most popular 10% of contents until the storage capacity is full, whereas the next most popular 10% of contents are stored in MANs.

Note that during the same simulation the caches position and contents placed in caches do not change.

## V. Numerical Results

### A. Simulation Settings and Performance Metrics

To evaluate the performance of each of the considered caching scenarios in Sec. IV, we developed a discrete event-driven C++ simulator. The topology considered in this study is similar to the topology in Fig. 1 and consists of 2 *MCNs*, 2 *MANs*, and 32 *ANs*. As for the technology of the links adopted, 20 GigabitEthernet (GE) technology is adopted in the *core* ring, whereas 10 GE is adopted in the *metro-core* ring and 2 GE is adopted in the *metro-access* segment. Moreover, each *metro-access* ring consists of 8 *ANs*, making up a total of 32 *ANs*.

We simulate the arrival of 500000 VoD requests, assumed as Poisson-distributed, with a fixed arrival rate guaranteeing no blocking of connections occurs, so that a fair comparative analysis between the different caching strategies is performed. $k$ is set to 3 and the content catalog size is 10000 contents, whose popularity is Zipf-like distributed as specified in Sec. II.

The performance of the caching algorithms is evaluated considering the following metrics:

- *Network Capacity Utilization*: it is the amount of network capacity utilized in a given caching scenario.

[2]We assume that contents are stored in their best resolution and if lower resolution is required, the content is encoded "on-the-fly" and transmitted with the proper bit-rate.

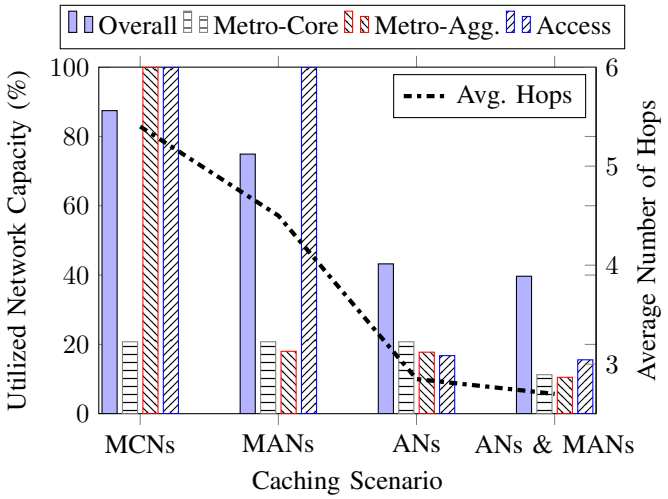



Fig. 3. Percentage of network capacity utilized in the overall network and in each of the network segments for each caching strategies with respect to the *No Caching* scenario and the average number of hops.

- *Average Number of Hops*: average number of hops of all provisioned VoD requests. We assume the average number of hops to be 8, 6, 4 and 1 from the video server, the *MCNs*, the *MANs* and the *ANs*, respectively.

### B. Discussion

*1) Effect of Caching Placement Strategy:* In this analysis we compare the performance of the caching strategies in terms of the overall *Capacity Utilization* in the whole network and in each of the network segments with respect to the *No Caching* scenario. For all the caching strategies, Fig. 3 shows the percentage of the network capacity utilized in the network and in each of the network segments with respect to the *No Caching* scenario and as well the average number of hops of all requests for each caching placement strategy. As expected, the network capacity utilized to serve all the VoD requests decreases as contents are cached in nodes closer to the end-users. More specifically, the *ANs* and *ANs & MANs* caching scenarios show the best performance as only 42% and 40% of the capacity utilized by *No Caching* are used, respectively. Note that, introducing caches at $MANs$ in addition to the caches at $ANs$ has small effect, as the difference between the capacity utilization of the mentioned scenarios is very small (2%). This is because caches at $MANs$ store unpopular contents that account for a relatively low number of requests.

Moreover, the results show that applying content caching and storing 10% of the popular contents at MCNs, reduces the traffic in the *metro-core* segment by around 80%. Similarly, by moving the caches to the MANs, the traffic in the *metro-aggregation* segment is also reduced. Applying content caching at $ANs$ reduces the traffic in the *access-rings* as all requests for contents stored at the $ANs$ are directly served without having to traverse the *access-rings*. Furthermore, additional savings in the *metro-core* and *metro-aggregation* segments of the network are still possible if content caching is enabled at both $MANs$ and $ANs$. However, in order to significantly reduce the amount of network capacity, an adequate distribution of the storage capacity should be performed such as to maximize the effect of the caches deployed at $MANs$. As for the average number of hops, it shows a behavior similar to that of the utilized network capacity as it decreases when caching is adopted at levels closer to end-users. We further notice that the average number of hops represents the efficiency of the caching placement strategy and thus it is adopted as a main metric in following simulations.

*2) Effect of Storage Distribution:* To investigate the convenience of utilizing caches at $MANs$ and $ANs$ simultaneously, we performed a new study considering different combinations of the number of caches and their placement for a fixed amount of total storage capacity in the network for two values of the popularity distribution parameter $\alpha$.

In these simulations settings, we allow a fixed total amount of storage of 64 TB, 128 TB and 160 TB in the network distributed among caches located at $MANs$ and at $ANs$ as shown in Tab. I. These total amounts of storage could allow caches at $ANs$, which sum up to 32, to store the most popular 2.5%, 5% or 6.25% of the contents (e.g., 2 TB, 4 TB or 5 TB).

Tab. I shows, for the same overall amount of storage capacity, 7 different combinations of the location and the storage capacity of caches in the network, going from a case where the storage capacity is located only in the metro segment (e.g., case #1 where 2 caches are located at $MANs$, each having 50% of the total storage capacity) to a case where all storage capacity is distributed in the access segment (e.g., case #7 where 32 caches are located at $ANs$, each with a capacity 3.125% of the total storage capacity). Thus, the main distinction between the different cases is the utilization of few caches of large capacity or more caches of lower capacity. The performance in the different cases is evaluated in terms of the average number of hops. In addition, we show the percentage of requests served from each segment of the network in each case study.

In Fig. 4 we compare the average number of hops in each of the cases for $\alpha$ = 0.8 and $\alpha$ = 1. The main difference between the popularity distributions for the mentioned values of $\alpha$ could be summarized as follows:

- for $\alpha$ = 1, the 10% most popular contents accounts for approximately 70% of the requests and the 20% most popular contents accounts for around 85% of the requests;
- for $\alpha$ = 0.8, the popularity 10% most popular contents accounts for approximately 55% of the requests and the 20% most popular contents accounts for around 75% of the requests.

Generally, for both values of $\alpha$ and for all different total storage amounts, the *Average Number of Hops* tends to decrease as we place more storage capacity in the access segment until a point when it becomes less convenient to use small caches even though they are deployed closer to end-users. As expected, case #1 in Tab. I, where contents

TABLE I
THE COMBINATIONS OF $MANs$ AND $ANs$ CACHES WITH THEIR RESPECTIVE STORAGE CAPACITY IN EACH OF THE CONSIDERED PLACEMENT CASES.

| | Total Storage 64 TB | | | | | | Total Storage 128 TB | | | | | | Total Storage 160 TB | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MANs | | | ANs | | | MANs | | | ANs | | | MANs | | | ANs | | |
| Case # | N. | Cap. | Total | N. | Cap. | Total | N. | Cap. | Total | N. | Cap. | Total | N. | Cap. | Total | N. | Cap. | Total |
| 1 | 2 | 32 | 64 | 0 | 0 | 0 | 2 | 64 | 128 | 0 | 0 | 0 | 2 | 80 | 160 | 0 | 0 | 0 |
| 2 | 2 | 16 | 32 | 8 | 4 | 32 | 2 | 32 | 64 | 8 | 8 | 64 | 2 | 40 | 80 | 8 | 10 | 80 |
| 3 | 2 | 16 | 32 | 16 | 2 | 32 | 2 | 32 | 64 | 16 | 4 | 64 | 2 | 40 | 80 | 16 | 5 | 80 |
| 4 | 2 | 16 | 32 | 32 | 1 | 32 | 2 | 32 | 64 | 32 | 2 | 64 | 2 | 40 | 80 | 32 | 2.5 | 80 |
| 5 | 2 | 8 | 16 | 32 | 1.5 | 48 | 2 | 16 | 32 | 32 | 3 | 96 | 2 | 20 | 40 | 32 | 3.75 | 120 |
| 6 | 0 | 0 | 0 | 16 | 4 | 64 | 0 | 0 | 0 | 16 | 8 | 128 | 0 | 0 | 0 | 16 | 10 | 160 |
| 7 | 0 | 0 | 0 | 32 | 2 | 64 | 0 | 0 | 0 | 32 | 4 | 128 | 0 | 0 | 0 | 32 | 5 | 160 |

are stored in $MANs$, show the worst performance. *Counter-intuitively, the cases where all storage capacity is deployed in the access segment, i.e., cases #6 and #7, do not show the best performance*. This is due to the fact that a lower number of contents is stored near end-users when a large number of caches is utilized. On the contrary, cases #4 and #5, where relatively small capacity caches are utilized at all $ANs$ and $MANs$, show the best performances for all given amounts of total storage and for both values of $\alpha$. This is mainly because the most popular contents are stored in caches located at $ANs$ and a large amount of different popular contents are stored in $MAN$ caches. Therefore, deploying caches in $ANs$ provides substantial benefits until a certain threshold. When this threshold is reached, deploying additional storage in all $ANs$ is less beneficial than deploying the same amount of storage in $MANs$ as it concentrates the storage of less popular, but more contents, different from those which could be possibly stored in $ANs$. This yields to more network capacity savings as a large number of requests have a reduced path length.

Comparing the network performance for varying values of $\alpha$, we notice that for the same overall amount of storage, the *Average Number of Hops* for $\alpha$ = 0.8 is higher than that for $\alpha$ = 1. This is due to the fact that when $\alpha$ = 0.8, the most popular contents account for a lower percentage of the requests if compared to the case where $\alpha$ = 1, and thus more requests

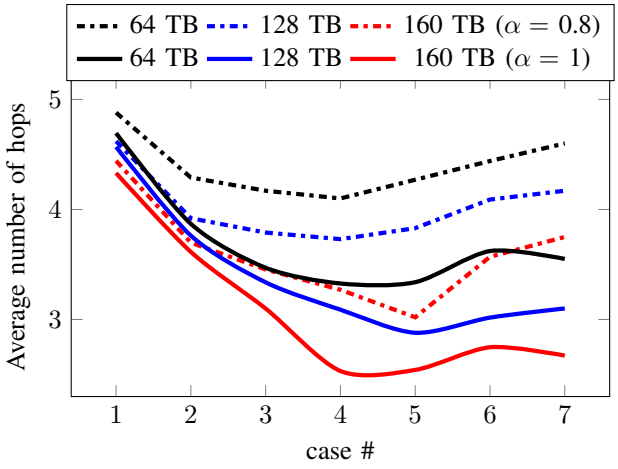
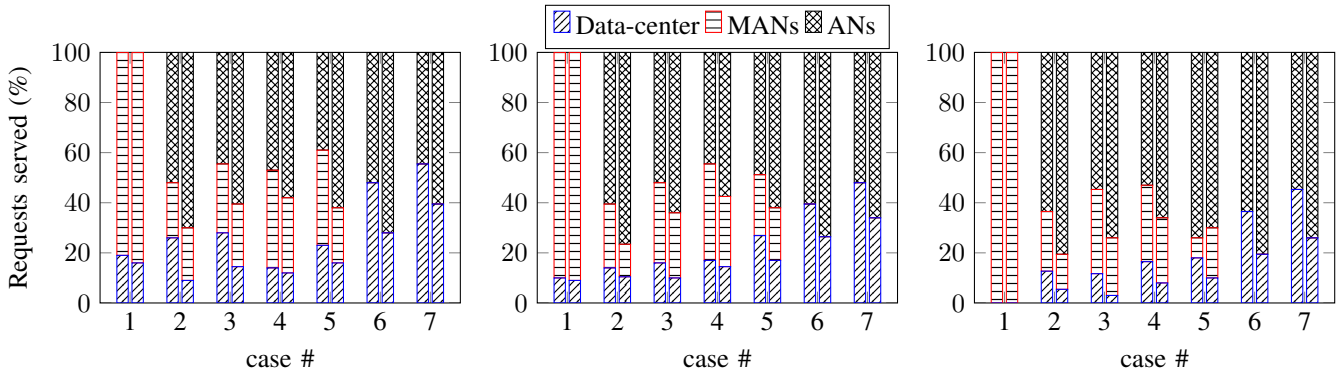



Fig. 4. Average number of hops of the various cases for different values of the distribution parameter $\alpha$.

are served from origin server, leading to an increase in the *Average Number of Hops*. Generally, for $\alpha$ = 0.8, cases #2, #3 and #4 show a better performance than other cases except the case for a total storage amount of 160 TB, which shows the best performance for case #5. This is due to the nature of the popularity distribution, as it is beneficial to store as many contents as possible closer to end-users and thus guaranteeing a significant amount of traffic is offloaded from the origin server, while for a total amount of 160 TB of storage, the available amount of storage is high enough to encourage more distribution in the $ANs$ and yet store a significant number of contents in the $MANs$.

In order to compare more in detail the different placement and storage combinations, we show the percentage of requests served from the *data-center*, and from caches located at $MANs$ and at $ANs$ for $\alpha$ = 0.8 and $\alpha$ = 1 in Fig. 5. First, we compare each considered case for the two values of $\alpha$, showing the effect of the popularity distribution parameter. Then, we compare the 7 cases for each total amount of storage. As expected, we notice that for $\alpha$ = 0.8, less requests are served from the $ANs$ and the $MANs$ even for the same total amount of storage. This is because of the nature of the popularity distribution for $\alpha$ = 0.8, which exhibits a lower percentage of requests for the most popular contents with respect to that for $\alpha$ = 1. Comparison the different cases in Tab. I, we notice that, although in cases #6 and #7 there is higher storage capacity in *ANs*, they do not result in the best performance. As a matter of fact, case #5 (where 75% of the total amount of storage is deployed in $ANs$) shows a better performance as contents can be retrieved from more locations. Indeed, cases #2, #3 and #6 have the highest percentage of requests served from $ANs$, but these cases do not show the best performance in terms of the average number of hops. This is because less $AN$ caches are utilized, and thus requests are routed over longer paths to reach end-users. In fact, case #5 shows a lower percentage of requests served from $ANs$ with respect to the mentioned cases, but since all $ANs$ were equipped with caches, requests are routed over shorter paths, thus saving more network resources. This implies that utilizing adequately the available storage capacity into caches at $MANs$ and $ANs$ leads to larger benefits from a network point of view, especially if an upper bound on the overall storage capacity is set. Additionally, we notice that different deployments of the available storage capacity across




(a) Total storage amount of 64 TB (b) Total storage amount of 128 TB (c) Total storage amount of 160 TB

Fig. 5. Percentage of requests served from the data-center, the caches located at $MANs$ and the caches located at $ANs$ for the caching strategies in Tab. I for $\alpha$ = 0.8 and $\alpha$ = 1 for a total storage amount of (a) 64 TB, (b) 128 TB and (c) 160 TB.

the access and metro segments, has a different impact on the percentage of requests served from the origin server. Indeed, the percentage of requests served from origin in cases #6 and #7, where all the storage capacity is deployed in the access segment, is higher than that in cases #3, #4 and #5, where the storage capacity is deployed across both the metro and access segments.

## VI. Conclusion

In this paper, we modeled a VoD content distribution scenario in a dynamic optical metro network. We presented a detailed comparison between different cache placement strategies. The results show that 70% of network resources can be saved by enabling content caching, especially at the access network segment. In addition, some of the results show how caches at a higher network level lose much of their impact when caches at lower level are utilized. To examine this issue, we performed a comprehensive analysis by considering a fixed amount of total storage capacity, to be distributed between $ANs$ and $MANs$. Results show that a blind placement of caching capability in access nodes may be inappropriate, and yields sub-optimal effects, whereas an adequate storage distribution between the access and the metro network segments exhibits an improved performance. Moreover, results show different deployments of caches in access and metro, has a direct impact on the amount of requests routed to the origin server.

## Acknowledgments

The work leading to these results has been supported by the European Community under grant agreement no. 761727 *Metro-Haul* project and the *Lombardy region* through *New Optical Horizon* project funding.

## References

[1] "Forecast and methodology, 2014-2019 white paper," *Cisco Visual Networking Index Technical Report*, 2015.

[2] J. Tang and T. Q. Quek, "The role of cloud computing in content-centric mobile networking," *IEEE Communications Magazine*, vol. 54, no. 8, pp. 52–59, 2016.

[3] L. Peterson, A. Al-Shabibi, T. Anshutz, S. Baker, A. Bavier, S. Das, J. Hart, G. Palukar, and W. Snow, "Central office re-architected as a data center," *IEEE Communications Magazine*, vol. 54, no. 10, pp. 96–101, 2016.

[4] G. Pallis and A. Vakali, "Insight and perspectives for content delivery networks," *Communications of the ACM*, vol. 49, no. 1, pp. 101–106, 2006.

[5] O. Ayoub, F. Musumeci, M. Tornatore, and A. Pattavina, "Techno-economic evaluation of cdn deployments in metropolitan area networks," in *International Conference on Networking and Network Applications*, 2017.

[6] 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 Online Conference on Green Communications (OnlineGreenComm)*, 2015.

[7] S. Dernbach, N. Taft, J. Kurose, U. Weinsberg, C. Diot, and A. Ashkan, "Cache content-selection policies for streaming video services," in *IEEE INFOCOM*, 2016.

[8] S. Hasan, S. Gorinsky, C. Dovrolis, and R. K. Sitaraman, "Trade-offs in optimizing the cache deployments of CDNs," in *INFOCOM, 2014 Proceedings IEEE*. IEEE, 2014.

[9] J. Araujo, F. Giroire, J. Moulierac, Y. Liu, and R. Modrzejewski, "Energy efficient content distribution," *The Computer Journal*, 2015.

[10] J. Llorca, A. M. Tulino, K. Guan, J. Esteban, M. Varvello, N. Choi, and D. C. Kilper, "Dynamic in-network caching for energy efficient content delivery," in *IEEE INFOCOM*, 2013.

[11] Z. Li *et al.*, "ICN based shared caching in future converged fixed and mobile network," in *IEEE HPSR*, Jul. 2015.

[12] C. Fricker, P. Robert, J. Roberts, and N. Sbihi, "Impact of traffic mix on caching performance in a content-centric network," in *IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2012.

[13] D. Kim, Y.-B. Ko, and S.-H. Lim, "Comprehensive analysis of caching performance under probabilistic traffic patterns for content centric networking," *China Communications*, vol. 13, no. 3, pp. 127–136, 2016.

[14] V. K. Adhikari, Y. Guo, F. Hao, M. Varvello, V. Hilt, M. Steiner, and Z.-L. Zhang, "Unreeling netflix: Understanding and improving multi-cdn movie delivery," in *IEEE INFOCOM*, 2012.