Techno-Economic Evaluation of CDN Deployments in Metropolitan Area Networks

Omran Ayoub, Francesco Musumeci, Massimo Tornatore and Achille Pattavina
Politecnico di Milano, Milan, Italy

Abstract—The explosion of Video-on-Demand (VoD) traffic has been a main driving force behind the Internet's evolution from a traditional connection-centric network architecture towards the new content-centric network architecture. To support this evolution, operators are deploying caches of VoD contents closer to users across network equipments deployed in core, metro and even access network segments to mitigate the traffic growth and improve the VoD quality of service. The deployment of storage elements (caches) across the network to deliver contents to end-users is known as a Content Delivery Network (CDN). For a CDN operator, it is important to minimize the cache-deployment cost while satisfying end users performance requirements. On one hand, deploying a high number of large caches closer to users improves the performance of a CDN (e.g., decreasing latency) but introduces huge capital and operational costs. On the other hand, deploying fewer caches in higher network segments introduces high operational costs due to high data traffic and might not satisfy future traffic demands, thus failing to meet users requirements. In this paper, we aim to identify the most cost-efficient cache deployment in CDN and to study the trade-off between the CDN performance and cost. We propose a CDN cost model which takes into consideration the capital and operational expenditures of CDN devices (e.g., caches and video interfaces) and of traffic required to serve the end users. We examine the effect of the content popularity on the cost of *CDN deployment* strategies, showing that there are different optimal cache deployment strategies for different popularity distributions. Results show that deploying a huge number of large caches in the access segment optimizes the quality of service for end users but increases the operational expenditure. Instead, a *CDN deployment* which utilizes caches across both the access and metro segments achieves a more balanced solution in terms of both the overall performance and cost.

I. INTRODUCTION

The growth of video-streaming traffic is motivating the shift of the Internet from a traditional "connection-centric" to a "content-centric" communication platform. According to [1], video streaming is expected to represent approximately 80% of global mobile data traffic by 2020. In addition to high bandwidth, video streaming features strict Quality of Service (QoS) requirements, e.g., in terms of low latency. This calls for new architectural solutions, based on Content Delivery Networks (CDN), where video contents are replicated in storage units (caches) to avoid that all user requests are served by the origin server [2] [3]. The video caches can be placed deep in the network (e.g., at cell sites or metropolitan central offices) to provide benefits in terms of network-capacity utilization and users' Quality of Experience (QoE), as most of the video traffic will only traverse network segments close to users. However, deploying a large number of caches in the access and metro networks requires excessive capital expenditure (CapEx). How to deploy a cost-effective caching solution for massive content distribution is not trivial. On one hand, deploying large caches near end-users improves QoS but increases CapEx and operational expenditure (OpEx). On the other hand, insufficient caching might not allow to save on network resources thus leading for an increased OpEx due to high data transit. In other words, the performance of a CDN (e.g., latency and bandwidth) and its cost are subject to trade-off. Moreover, it is decisive to examine this trade-off while considering different popularity distributions. In this paper, we present a techo-economic study of various CDN deployments with the aim of identifying the most cost-efficient CDN deployment. The study shows the effect of the performance of the CDN deployment on its cost and quantifies the trade-off between CapEx and OpEx under different content popularity distributions. Results show that utilizing caches across both the access and the metro segments yields a better solution than other CDN deployments, both in terms of performance and cost.

A. Related Works

Previous works have investigated the CDN deployment, yet with little focus on the problem of cost-effective cache deployment in the metro and access, or with focus limited to the core segment of the network. For example, Ref. [4] presented an optimization model with the objective to minimize the total cost of a CDN considering energy, bandwidth and server costs constricted by a performance constraint, but focusing only on core networks. Refs. [5] and [6] investigated the optimal location of the caches in the network showing the reduction of traffic and energy due to content caching, but without accounting the cost of the CDN deployment. In our work, we also aim to find an optimal CDN deployment, but considering CDN costs. Also, In Refs. [7]-[8], authors presented an in-network caching model in access and metro showing the impact of video caching on the network but neglecting the cost of the CDN deployment. Likewise, Ref. [9] shows the impact of caching VoD contents in access routers on the Internet bandwidth through a cache replacement algorithms.

The rest of the paper is organized as follows. In Sec. II we describe the models for the network architecture and the video content used in our work. Sec. III we present the CapEx/OpEx model and the main CDN parameters. In Sec. IV we describe the evaluation settings and discuss the numerical results. Sec. V concludes the paper.
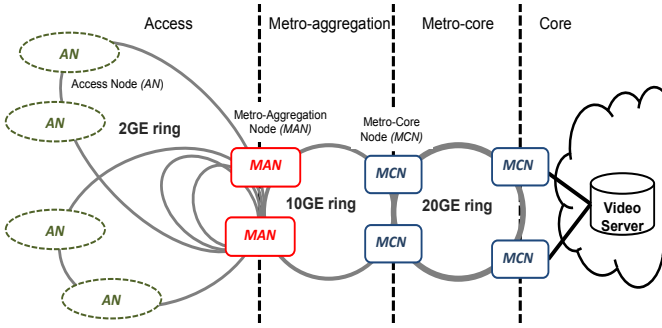
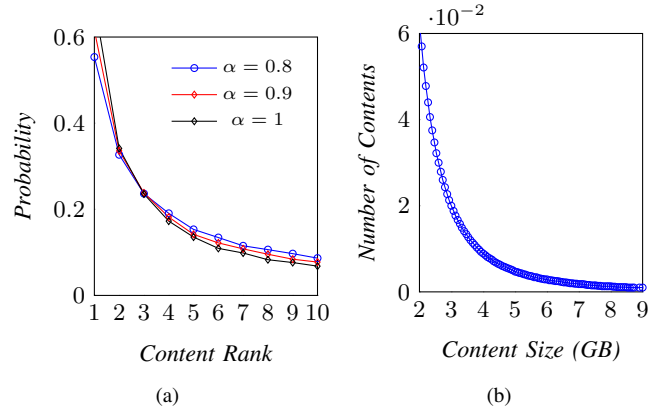Fig. 1. The network topology considered in our study.

Fig. 2. (a) An example of the effect of the popularity skew parameter α on the popularity distribution of 10 contents, (b) The distribution of the size of contents according to Pareto distribution ($\beta = 1.8$).

II. NETWORK AND VIDEO-ON-DEMAND MODELING

A. Network Model

We consider a metropolitan network topology spanning over four hierarchical levels ($L = 4$), as depicted in Fig. 1.

- *core* level, consisting *Metro-Core Nodes (MCNs)* connected to data centers hosting *video server*.
- *metro-core* level, consisting of *Metro-Core Nodes (MCNs)* interconnected in a ring topology.
- *metro-aggregation* level, consisting of *Metro Aggregation Nodes (MANs)* and *MCNs* interconnected in a ring topology. Metro-access rings are connected to the metro ring through the *MANs*.
- *access* segment, consisting of *Access Nodes (ANs)* interconnected in a ring topology, where each *AN* represents aggregated users.

According to this considered topology, we can define an average hop distance for each network level. The average hop distance of users from from ANs, MANs, MCNs and the video server are 1, 4, 6.5 and 8 respectively. Network nodes such as ANs, MANs and MCNs can be equipped with storage capacity to perform *caching* depending on the *CDN deployment*. Hence, we can define different *CDN deployments* based on the location of the caches and their storage capacity. For sake of simplicity, the caching technology is assumed to be independent from cache location. Note that the network devices are also equipped with video interfaces to perform content streaming.

B. Video-on-Demand Model

The video content catalog is characterized by its size (number of contents) and its popularity distribution, whereas a video content is described by its size (byte) and its popularity. We consider a content catalog consisting of 10,000 videos following a *Zipf* popularity distribution characterized by the skew parameter (exponent) α [10]. 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 α is the Zipf popularity distribution parameter. Some studies [11] [12] have shown that the value of α varies around 0.8, whereas other studies

suggested that α could have a value = 1 [10]. The value of α is a critical parameter. For example, for $\alpha = 1$, the 15% most popular account contents accounts for 80% of the video requests, whereas for $\alpha = 0.8$, the 40% most popular contents account for 80% of video requests. Figure 2(a) plots the popularity distribution curves of the first most popular 10 contents for $\alpha = 0.8$, 0.9 and 1. The curves show that for lower values of α, popular contents becomes slightly less popular, whereas contents belonging to the long tail gain more popularity. Contents sizes vary between 2 GB (e.g., short duration episodes) and 9 GB (e.g., long duration movies) following a *Pareto* distribution model having a distribution parameter $\beta = 1.8$ [10] [13], thus achieving an average video content size of 4.5 GB, as shown in Fig. 2(b). Finally, we consider that each video consists of a number of chunks, each with a size c and a duration of r seconds. This consideration will allow us to precisely estimate the number of simultaneous requests per second.

III. CDN DEPLOYMENT AND COST MODEL

A. CapEx/OpEx Model

In this section, we present the model for evaluating the CapEx and OpEx associated to the various *CDN deployment*. For clarity and ease of discussion, we summarize the important notations in Tab. I.

The CapEx accounts for the cost of the caches, which depends on their storage capacities and their video-delivery interfaces. We calculate the CapEx using a 3-step-procedure:

- 1) We evaluate the number of caches N_i to be deployed at each level.
- 2) We calculate the storage capacity of the cache, S_i, in level i, in function of the hit-ratio of the cache and the popularity distribution.
- 3) We determine the minimum number of 1 Gbps video interfaces needed to accommodate the peak traffic, represented by V_i.

V_i is calculated as shown in Eqn. 1, where U represents the maximum simultaneous number of users, and b_r represents the average bit-rate. For example, assuming that b_r = 6 Mbps and U = 5000, the maximum amount of traffic per second in the network is 30000 Mbps ($b_r \cdot U$). If the caches located in the access segment have a hit-ratio H_1 = 0.8, they will serve 24000 Mbps ($b_r \cdot U \cdot 0.8$). If the number of caches deployed in access are 32, then each cache serves 750 Mbps ($\frac{24000}{32}$), thus each cache requires a 1 Gbps video interface.

$$V_i = \frac{U \cdot b_r \cdot H_i}{N_i} \qquad (1)$$

C_i represents the cost of a cache deployed at level i calculated as shown in Eqn. 2, where Sc and Vc represent the cost of 1 GB of storage and the cost of 1 Gbps video interface, respectively. The storage cost [15] and video delivery interface cost [16] in terms of cost units (CU) are reported in Tab. III.

$$C_i = S_i \cdot Sc + V_i \cdot Vc \qquad (2)$$

Finally, the total CapEx of the CDN is calculated according to Eqn. 3 by summing up the total costs of caches deployed in access, metro and core networks.

$$CapEx = \sum_{i=1}^{L-1} N_i \cdot C_i \qquad (3)$$

To calculate OpEx, the total amount of traffic per hour B_{tot} and the total amount of traffic delivered from the i-th network level B_i are calculated. The average amount of traffic per hour B_{tot} is calculated as shown in Eqn. 4, where U_{avg}, c and r represent the average simultaneous number of users, the chunk's size and chunk's duration, respectively.

$$B_{tot} = \frac{3600 \cdot c \cdot U_{avg}}{r} \qquad (4)$$

The average amount of traffic per hour served by the caches at network level i is thus calculated as shown in Eqn. 5.

$$B_i = H_i \cdot B_{tot} \qquad (5)$$

The total OpEx for a given time interval τ is calculated as shown in Eqn. 6. Specifically, the first contribution accounts for caches powering, where E represents the operational cost per hour for each storage element of 250 GB and is set to 0.2 CU per hour and D_i represents the number of 250 GB storage elements in every network level. The second contribution represents the transit costs where T_i is the cost of delivering 1 GB of video content. The third contribution accounts for the amount of traffic sent from the origin video server to the caches to perform content replacement operations. We consider a reasonable cache overwrite value per day (20% of the cache size, represented by w) [14], as less popular contents are substituted with more popular ones. The values for the Internet transit costs[1] [17] are reported in Tab. III.

[1]Note that the costs from the metro cache and ANs caches are assumed to be equal to half and one quarter of the transit cost from the core level.

Symbol	Definition
P	CDN performance parameter
L	Number of network levels
k_i	Mean hop distance between level i and users
h_i	Hit-ratio of caches deployed at level i
H_i	Overall hit-ratio of caches deployed at level i
S_i	Storage capacity deployed at cache of level i
V_i	Number of video interfaces deployed at level i
C_i	Cost of cache deployed at network level i
U	Maximum simultaneous number of users
U_{avg}	Average simultaneous number of users
B_{tot}	Total amount of traffic served per hour
B_i	Amount of traffic served from caches at level i
w	Percentage of cache capacity daily replaced
c	Chunk size
r	Chunk duration
b_r	Average bit-rate of all video requests

$$OpEx = \sum_{i=1}^{L-1} ((N_i \cdot D_i \cdot E) + (B_i \cdot T_i) + (w \cdot S_i \cdot T_4)) \cdot \tau \quad (6)$$

B. CDN Performance Parameter

In this section, we introduce the overall hit-ratio parameter, the CDN performance parameter and we clarify how the *CDN deployments* (cache hit-ratio and storage capacity) are set up. First, we introduce the overall hit-ratio H_i of a cache deployed at level i, which represents the number of requests delivered by the caches at network level i with respect to the total number of requests. As previously mentioned, the *CDN deployments* differ in the combination of number of caches deployed, their locations, hit-ratios and storage capacities. For each of these combinations, i.e., *CDN deployments*, a parameter is assigned to quantify the CDN's performance.

Similar to [4], we characterize the performance of a *CDN deployment* by a parameter P, which is linked to the average number of hops traversed by all requests to reach the end users, where a lower P indicates a higher overall QoE for the users. We extend the approach in [4] by also considering metro and access caches in addition to the core caches. P is calculated according to Eqn. 7 where k_i represents the average hop distance from level i to end users.

$$P = \sum_{i=1}^{L} k_i \cdot H_i \qquad (7)$$

The set up of the *CDN deployment* is performed in a 3-step-procedure:

- 1) We insert the value of P as a constraint on the performance of the CDN.
- 2) We find the possible combinations of the hit-ratios of the caches of different levels that guarantee the performance constraint. The combination which results in the lowest CapEx is chosen.

TABLE II

THE OVERALL HIT-RATIO AND RESPECTIVE AMOUNT OF STORAGE IN EACH OF THE CACHES IN THE NETWORK LEVELS FOR $\alpha = 0.8$

Overall Hit-ratio			Storage (GB)			P
AN Caches	Metro Cache	Core Cache	AN Caches	Metro Cache	Core Cache	
1	0	0	45000	0	0	1
0.86	0.14	0	23706	21294	0	1.4
0.86	0.04	0	12708	285574	3735	1.8
0.66	0.34	0	8000	37000	0	2
0.5	0.5	0	2713	42286	0	2.5
0.38	0.58	0	1003.5	36778	0	3
0.3	0.46	0.02	450	13729.5	1602	4
0.18	0.42	0.02	100	5377	756	5
0	0	1	0	0	45000	6.5
0	0	0	0	0	0	8

- 3) We calculate the amount of storage capacity required at caches of every level (according to popularity distribution).

Table II shows the hit-ratio and the storage capacity of caches for the chosen *CDN deployments* for various values P. The *CDN deployments* corresponding to $P = 1$ (caches in access store all contents) and $P = 8$ (no caches, all contents served from video server) serve as benchmark *CDN deployments*.

TABLE III

THE COST VALUES OF THE DATA STORAGE, VIDEO INTERFACES AND INTERNET TRANSIT.

Parameter	Value (CU)
Sc	0.17/GB
Vc	50
T_1	0.09/GB
T_2	0.08/GB
T_3	0.04/GB
T_4	0.02/GB

IV. ILLUSTRATIVE NUMERICAL RESULTS

A. Evaluation Settings

We consider a maximum simultaneous number of users U = 5000 and average simultaneous number of users U_{avg} = 2500. The chunk size c is set at 1.8 MB with a duration r = 1.5 seconds [18]. In the topology considered, caches could be possibly deployed at 32 *ANs*, 2 *MANs* and 1 *MCN*.

In addition to the mentioned evaluation settings, we assume the operational and the transit costs to decrease by 5% yearly, whereas we consider a traffic increase of 10% per year. This increase can be associated to an increase in the number of users and/or in the required bit-rate, corresponding to improved quality of video delivery. Note that in case the deployed interfaces are no more sufficient to accommodate peak traffic, we assume that additional video interfaces are deployed and their cost is also included in our evaluation.

Moreover, for the CapEx/OpEx evaluations, we focus on 6 different *CDN deployments* in which 2 are benchmark *CDN deployments* as listed below:

- *No Caching*, a benchmark CDN deployment where all requests are served from the origin with $P = 8$.

- *Metro Cache* CDN deployment with $P = 4$, where the storage capacity is mainly centralized in the metro cache deployed at the MAN In this CDN deployment.
- *Balanced* CDN deployment with $P = 2.5$, where the overall hit-ratio is equally divided between caches in ANs and the metro cache.
- *Distributed* CDN deployment with $P = 2$, where the storage capacity is distributed in caches deployed at ANs havin a hit-ratio of 0.66 (refer to Tab. II).
- *Widely Distributed* CDN deployment with $P = 1.4$, where the storage capacity is deployed at ANs having a hit-ratio of 0.86 (as shown in Tab. II).
- *Only ANs* benchmark CDN deployment. This CDN deployment is the complete opposite of the *No Caching* CDN deployment where caches of storing all content library is deployed at ANs, thus all requests are served from ANs leading to $P = 1$.

B. Discussion

In Fig. 3, we examine the CapEx and OpEx of 6 *CDN deployments*. For all values of α, the *Only ANs CDN deployment* ($P = 1$), has highest cost. This is not only due to high CapEx (huge number of large-dimensioned caches), but also to the fact that such a number of large caches consumes high amounts of energy and require high operational costs. For the *No Caching CDN deployment* ($P = 8$), the CDN cost consists of only a transit contribution, yet still exposes much higher cost than that of other *CDN deployments*, where caches are utilized across both the access and metro segments. This is because the transit cost is very high as all requests are served from the origin server. As for the *Metro Cache CDN deployment* ($P = 4$) and the *Widely Distributed CDN deployment* ($P = 1.4$), results show that such deployments are not the most economically-efficient. This is because the *Metro Cache CDN deployment* is not sufficient to limit the transit costs due to the high traffic cost, whereas the *Widely Distributed CDN deployment* fails due to the high amount of OpEx required by the largely-dimensioned caches. On the contrary, the *Balanced* ($P = 2.5$) and the *Distributed CDN deployments* ($P = 2$) exhibit a good overall CDN quality and the lowest *CDN deployment* cost. Note that there are different optimal *CDN deployment* strategies for different popularity
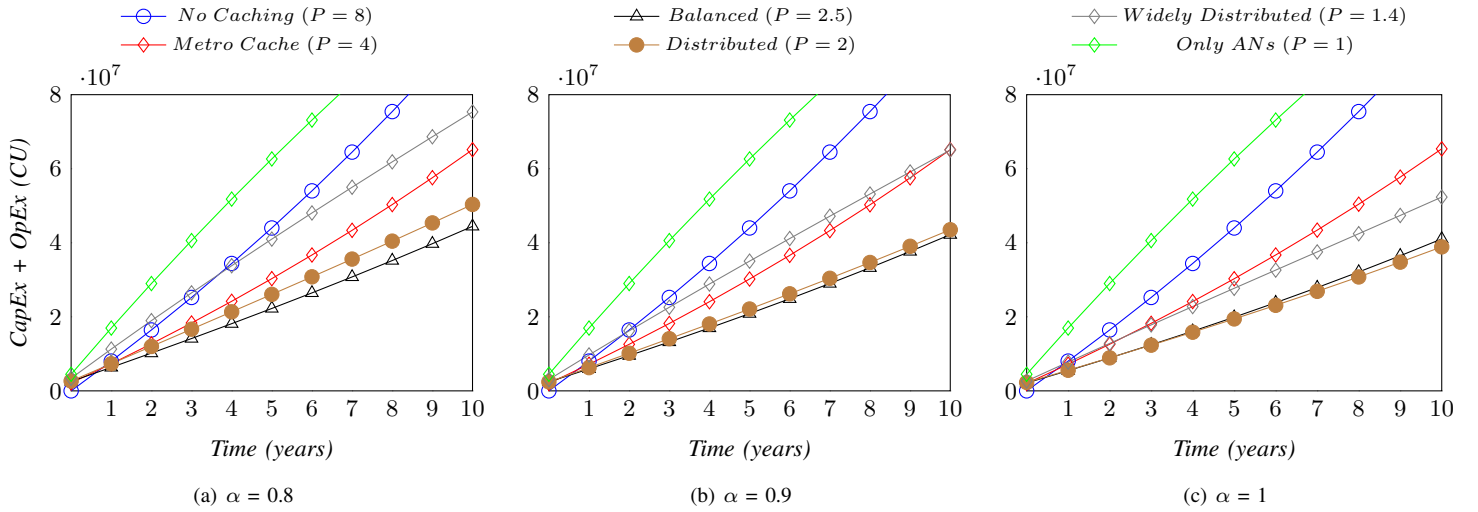
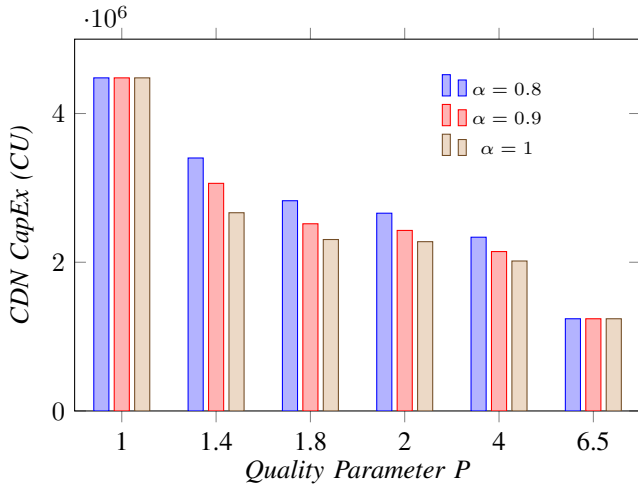Fig. 3. Total CDN cost as a function of time (years) for (a) $\alpha = 0.8$, (b) $\alpha = 0.9$ and (c) $\alpha = 1$.

Fig. 4. CDN $CapEx$ vs P for $\alpha = 0.8$, $\alpha = 0.9$ and $\alpha = 1$.

distributions. For instance, for the cases where $\alpha = 0.8$ and 0.9, the best *CDN deployment* is achieved by *Balanced CDN deployment* with $P = 2.5$, which exhibits a balanced storage distribution between caches deployed in ANs and cache located in metro (as shown in Tab. II). However, for $\alpha = 1$, the *Distributed CDN deployment* with $P = 2$ is preferred. This is because such a popularity distribution allows achieving higher hit-ratios with smaller cache dimensioning, thus obtaining an overall better CDN quality with lower lower OpEx.

Now, we focus solely on CapEx to investigate the effect of the *CDN deployment's* performance on its CApEx. Figure 4 plots the CDN CapEx as a function of the quality parameter P for $\alpha = 0.8$, 0.9 and 1. Note that, for $P = 8$, there is no CDN CapEx since in this scenario there is no *CDN deployment* (as shown in Tab. II) and all requests are served from the video server. For $P = 6.5$, the CDN CapEx is relatively low with respect to other *CDN deployments*, because only

one large cache storing all contents is placed in the core network level. For the same reason, the CDN CapEx is the same for different values of α. A significant cost increase is then noticed for a *CDN deployment* of $P = 4$, $P = 2$ and $P = 1.8$, that is when caches in the ANs are deployed. Although the quality parameter P differs significantly between these three *CDN deployments*, the CDN CapEx only shows a slight increase, up to 8%. This is mainly due to the popularity distribution, that allows enhancing the overall performance of the CDN (from $P = 4$ to $P = 1.8$) through the addition of a relatively small amount of storage capacity. Another significant CapEx increase is noticed for a *CDN deployment* of $P = 1.4$, as to achieve such a quality, caches in the ANs are largely-dimensioned to account for a high hit-ratio (around 0.85, as shown in Tab. II). A main drawback of such a *CDN deployment* is that it requires storing unpopular contents at ANs. Moreover, we notice that the benchmark *CDN deployment* where all contents are stored in caches at ANs ($P = 1$), exhibits significantly higher CapEx with respect to other *CDN deployments*. This is expected as caches are expanded to store unpopular contents, thus resulting in unnecessary CapEx. As far as the effect of α is concerned, we notice that, for a given value of P, the CDN cost for a larger value of α is lower, as less amount of storage is required to achieve high hit-ratios. However, for high values of P, i.e., lower quality, this difference is negligible as only few caches are deployed. For example, for $\alpha = 1$, a *CDN deployment* with $P = 1.4$ costs less than *CDN deployment* with $P = 2$ for $\alpha = 0.8$. Figure 5 shows the fraction of the CDN cost for the caches deployed in the ANs, MANs and MCNs for different values of α. For all values of α, numerical results show that, when caches are placed in the ANs (values of P less than 5), their cost makes up a large fraction of the total cost. As increasing the hit ratio of caches deployed at ANs until achieving a $P = 1.8$ only introduces a slight increase in the CapEx, we affirm that there exists a threshold on the cache dimensioning that should not
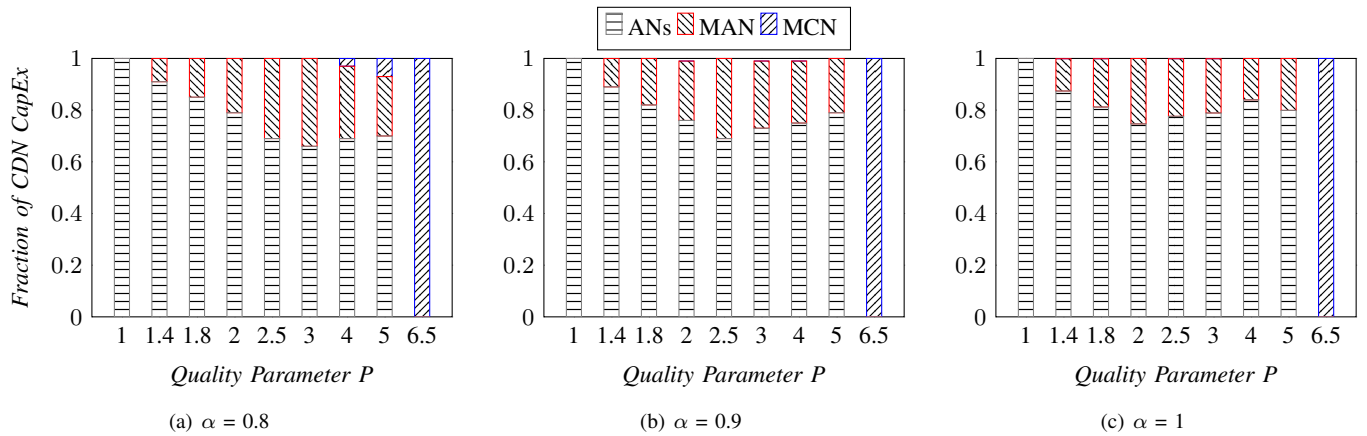
Fig. 5. Fraction of the CDN $CapEx$ per network level vs. the performance value P for (a) $\alpha = 0.8$, (b) $\alpha = 0.9$, (c) $\alpha = 1$.

be exceeded to avoid unnecessary CapEx.

V. CONCLUSION

This paper presents a techo-economical study of different *CDN deployments*. The trade-off between the CDN's CapEx/OpEx and its performance (represented by the average number of hops) under different content popularity distributions is investigated. Moreover, results show that for each specific popularity distribution a unique caches deployment strategy across the network levels is the most appropriate. Furthermore, the study shows how deploying a large-dimensioned caches in the access segment striving for an optimal quality is not the best economic solution as the OpEx of the caches increases exponentially, where instead, a deployment of caches in both access and metro segments achieves a near-optimal overall quality of video delivery and is a better economic solution.

REFERENCES

[1] Cisco Visual Networking Index. "Global Mobile Data Traffic Forecast Update, 2015−2020 White Paper, 2016."
[2] Nygren, Erik, Ramesh K. Sitaraman, and Jennifer Sun. "The Akamai network: a platform for high-performance Internet applications," ACM SIGOPS Operating Systems Review 44.3 (2010): 2-19.
[3] Dilley, John, et al. "Globally distributed content delivery," IEEE Internet Computing 6.5 (2002): 50-58.
[4] Hasan, Syed, et al. "Trade-offs in optimizing the cache deployments of CDNs," INFOCOM, Proceedings IEEE, 2014.
[5] Applegate, David, et al. "Optimal content placement for a large-scale VoD system," Proceedings of the 6th International COnference. ACM, 2010.
[6] Borst, Sem, Varun Gupta, and Anwar Walid. "Distributed caching algorithms for content distribution networks." INFOCOM, Proceedings IEEE, 2010.
[7] Llorca, Jaime, et al. "Dynamic in-network caching for energy efficient content delivery," INFOCOM, Proceedings IEEE, 2013.
[8] Savi, Marco, et al. "Energy-efficient caching for Video-on-Demand in Fixed-Mobile Convergent networks," Green Communications (Online-GreenComm), IEEE Online Conference on, 2015.
[9] Fricker, Christine, et al. "Impact of traffic mix on caching performance in a content-centric network," Computer Communications Workshops (INFOCOM Workshops), IEEE Conference on, 2012.
[10] Yu, Hongliang, et al. "Understanding user behavior in large-scale video-on-demand systems," ACM SIGOPS Operating Systems Review. Vol. 40. No. 4. ACM, 2006.
[11] Chang, Hyunseok, et al. "An empirical approach to modeling inter-AS traffic matrices," Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement. USENIX Association, 2005.
[12] Feldmann, Anja, et al. "A methodology for estimating interdomain web traffic demand," Proceedings of the 4th ACM SIGCOMM conference on Internet measurement. ACM, 2004.
[13] Li, Haitao, et al., "Video requests from online social networks: Characterization, analysis and generation," INFOCOM, Proceedings IEEE, 2013.
[14] Carbunar, Bogdan, et al. "A framework for network aware caching for video on demand systems," ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 9.4 (2013): 30.
[15] cloud.google.com/compute/pricing#internet_egress
[16] cisco.com/c/en/us/products/collateral/interfaces-modules/transceiver-modules/data_sheet_c78-455693.html
[17] azure.microsoft.com/en-us/pricing/details/cdn/
[18] Casas, Pedro, et al. "When YouTube does not workAnalysis of QoE-relevant degradation in Google CDN traffic," IEEE Transactions on Network and Service Management 11.4 (2014): 441-457.