Hybrid Multi-tenant Cache Management for Virtualized ISP Networks

Maxim Claeys[a,c,*], Daphné Tuncer[b], Jeroen Famaey[c], Marinos Charalambides[b], Steven Latré[c], George Pavlou[b], Filip De Turck[a]

[a] *Department of Information Technology, Ghent University - iMinds, Ghent, Belgium*
[b] *Department of Electronic & Electrical Engineering, University College London, London, United Kingdom*
[c] *Department of Mathematics and Computer Science, University of Antwerp - iMinds, Antwerp, Belgium*

Abstract

In recent years, Internet Service Providers (ISPs) have started to deploy Telco Content Delivery Networks (Telco CDNs) to reduce the pressure on their network resources. The deployment of Telco CDNs results in reduced ISP bandwidth utilization and improved service quality by bringing the content closer to the end-users. Furthermore, virtualization of storage and networking resources can open up new business models by enabling the ISP to simultaneously lease its Telco CDN infrastructure to multiple third parties. Previous work has shown that multi-tenant proactive resource allocation and content placement can significantly reduce the load on the ISP network. However, the performance of this approach strongly depends on the prediction accuracy for future content requests. In this paper, a hybrid cache management approach is proposed where proactive content placement and traditional reactive caching strategies are combined. In this way, content placement and server selection can be optimized across tenants and users, based on predicted content popularity and the geographical distribution of requests, while simultaneously providing reactivity to unexpected changes in the request pattern. Based on a Video-on-Demand (VoD) production request trace, it is shown that the total hit ratio can be increased by 43% while using 5% less bandwidth compared to the traditional Least Recently Used (LRU) caching strategy. Furthermore, the proposed approach requires 39% less migration overhead compared to the proactive placement approach we previously proposed in Claeys et al. (2014b) and achieves a hit ratio increase of 19% and bandwidth usage reduction of 7% in the evaluated VoD scenarios and topology.

Keywords: Cache management, Hybrid management, Content placement

1. Introduction

Over the last decade, video streaming services have become the principal consumers of Internet traffic. In 2014, video over Internet Protocol (IP) has been reported to account for 67% of all IP traffic while this share is predicted to grow to 80% by 2018 (Cisco, 2015). Furthermore, with the advent of 4K resolution and 3D video, the quality requirements for these services are becoming more stringent. Today, the prevalent method to deliver the video to the end-users relies on Content Delivery Networks (CDNs). In order to meet the growing quality requirements, CDNs aim at bringing the content closer to the clients to reduce both the latency and the bandwidth consumption. Currently, a common way for traditional CDNs, such as Akamai or Netflix, to bring their content to the edge of the network is to physically place part of their distributed storage infrastructure inside the Internet Service Provider (ISP) network or to connect it to a nearby Internet exchange point through manually-negotiated contractual agreements.

However, given that large CDNs control both the placement of the content and the server selection strategy (i.e., select from which location to satisfy each request) across their geographically dispersed storage infrastructure with only limited knowledge about the underlying network, they can put an immense strain on the resources of ISP networks (Jiang et al., 2009). This has resulted in increasing operational costs and decreasing revenues for the ISPs. Therefore, they have started to explore alternative business models and service offerings, leading to the deployment of Telco CDNs. As ISPs have global knowledge about the utilization of their network resources, controlling the storage of content deep inside their network allows them to reduce the bandwidth demand on their backbone infrastructure while significantly improving the service quality for the end-users.

Moreover, the advent of cloud computing, Software-Defined Networking (SDN) and Network Function Virtualization (NFV) technologies have enabled ISPs to virtualize their Telco CDN infrastructures. This allows them to dynamically offer virtual storage and content delivery services at the edge of the network, redeeming traditional

[*] Corresponding author

Email addresses: maxim.claeys@intec.ugent.be (Maxim Claeys), d.tuncer@ee.ucl.ac.uk (Daphné Tuncer), jeroen.famaey@uantwerpen.be (Jeroen Famaey), m.charalambides@ucl.ac.uk (Marinos Charalambides), steven.latre@uantwerpen.be (Steven Latré), g.pavlou@ucl.ac.uk (George Pavlou), filip.deturck@intec.ugent.be (Filip De Turck)

CDN providers from installing additional hardware. In previous work, we proposed a proactive cache management system for ISP-operated multi-tenant Telco CDNs to optimize the content placement and server selection across tenants and users (Claeys et al., 2014a,b). The tenants specify the amount of storage capacity they want to lease, while the management framework decides on the allocation of the leased capacity across the storage infrastructure, the content placement and the server selection. Such a scenario allows the tenants to bring their content closer to the end user without having to physically deploy dedicated storage infrastructure inside the ISP network. This reduces the installation and operational costs for the content provider, while the ISP has better control over its network resources. All of these decisions are based on the prediction of content popularity and the geographical distribution of requests, making the performance strongly dependent on the accuracy of the request prediction.

Proactively placing content inside the ISP network has multiple advantages over reactive cache replacement, for example reduced bandwidth consumption by performing content migrations during off-peak hours and delivering cache hits on the first request of popular content. However, given the dynamic nature of content popularity in a Video-on-Demand (VoD) scenario, the performance of the proactive cache management presented in previous work can significantly fluctuate over time. To deal with the uncertainties in the future request pattern, this paper proposes a hybrid cache management system, which combines the benefits of both proactive and reactive content placement strategies. In the proposed approach, the total caching space available in the network is divided in such a way that part of it is reserved to proactively push content in the different caching locations, while the rest is used to implement reactive caching. As such, content placement and server selection can be optimized across tenants and users, based on the predicted content popularity and the geographical distribution of requests, while simultaneously providing reactivity to unexpected changes in the request pattern. Furthermore, frequently migrating content to reconfigure the proactive placement can significantly influence the total bandwidth usage. Therefore, in this paper, the migration overhead is taken into account in the placement decisions.

The main contributions of this work are as follows. We redefine the Integer Linear Program (ILP) model of the multi-tenant content placement and server selection problem proposed in previous work (Claeys et al., 2014a,b) to take into account the overhead introduced by the frequent content migrations and to reduce the level of detail required for the request predictions. In contrast to our previous work, which required predictions of accurate request timestamps, the updated ILP model only requires predictions of request aggregates. We also investigate the benefits that can be obtained in terms of performance improvement by augmenting the results of the content popularity prediction algorithm with exogenous information about the future request pattern, e.g., based on trends in social media (Asur and Huberman, 2010). In addition, we show how the caching space can be divided between proactive and reactive placement and investigate the added value when considering adaptivity in the proposed hybrid caching approach. We evaluate the performance in terms of both network and caching metrics based on a real VoD request trace from a major European ISP. The results show that the proposed hybrid cache management approach can outperform both a purely proactive and a purely reactive approach while significantly reducing the migration overhead compared to our previous work.

The remainder of this paper is structured as follows. Section 2 highlights related work on cache management in distributed storage infrastructure. Section 3 presents the proposed management architecture and describes the scenario under study. Next, the problem is formally modeled as an ILP in Section 4 and the hybrid cache division strategy is described. Section 5 introduces the setup used to evaluate the proposed approach while the evaluation results are detailed in Section 6. Finally, the main conclusions are presented in Section 7.

2. Related work

Hybrid cache management approaches that partition the storage space available at each caching location in order to implement different caching strategies were proposed by Applegate et al. (2010) and Sharma (2013). However, the effect of partitioning was investigated on few performance metrics only (network utilization and latency) based on arbitrary fixed ratios. Furthermore, in these approaches, the partitioning is performed on each cache individually. In contrast, this paper proposes a network-level partitioning and provides a more systematic analysis as it relies on a wide range of ratios and highlights the effect based on both network and caching metrics. In addition, the benefits of updating the partitioning ratio in an adaptive fashion are also quantitatively evaluated.

In the last few years, there have been significant research efforts towards the development of proactive cache management approaches. While some of the proposals have focused on content placement strategy only (Sourlas et al., 2012; Laoutaris et al., 2006; Borst et al., 2010; Dai et al., 2012; Wauters et al., 2006; Korupolu and Dahlin, 2002), others have proposed new mechanisms to manage the redirection of user requests (Valancius et al., 2013; Frank et al., 2012). Optimal solution structures for the combined problem of content placement and server selection have also been developed (Baev and Rajaraman, 2008; Bekta et al., 2008; Applegate et al., 2010). However, all of these consider a single provider scenario only, which is a subset of the problem we investigate.

One of the most relevant approaches for the problem investigated here is the work of Laoutaris et al. (2004, 2005), in which algorithms for the joint optimization of capacity allocation and object placement decisions under known topological and user demand information were developed. However, despite the similarity in terms of problem objectives, the proposed models cannot apply to our scenario as they disregard per node capacity constraints and assume hierarchical caching infrastructures only.

In parallel to these management strategies, research efforts have also focused on the development of new models and frameworks to support the interaction between ISPs and CDNs. These range from ISP-centric caching approaches (Kamiyama et al., 2009; Cho et al., 2011), which exclude CDNs from the delivery chain, to collaborative solutions (Jiang et al., 2009; Frank et al., 2012; Wichtlhuber et al., 2015), defining new models of cooperation between ISPs and CDNs. Another relevant initiative concerns the Content Delivery Network Interconnection (CDNI) working group of the Internet Engineering Task Force (IETF) which focuses on standardizing the communications between CDNs to allow interoperability between different vendors[1]. All of these approaches are targeting to improve content delivery performance. Intermediate solutions, such as the one proposed by co-authors of this paper (Tuncer et al., 2013), have also been considered and rely on providing a limited capacity CDN service within ISP networks by deploying caches at network edges.

Finally, while this paper focuses on a VoD service, proactive content placement has recently been investigated in the context of live video streaming (Mukerjee et al., 2015), which focuses on the end-to-end optimization of the stream delivery. In contrast to our approach, this targets very short re-configuration intervals (in the order of milliseconds) and involves changing different parameters (e.g., video stream placement and bitrate encoding selection), which are features specific to this type of service.

To the best of the authors' knowledge, this paper is the first to propose a hybrid cache management approach with a network-level cache division and support for multitenant scenarios.

3. Experiment description

In this section, the considered experiment is described. Section 3.1 introduces the ISP-based CDN service where capacity is leased to one or more content providers. In Section 3.2, the characteristics of the investigated VoD use-case are presented, based on which the request prediction strategy is introduced in Section 3.3. Finally, Section 3.4 discusses the limitations of popularity prediction algorithms in the considered use-case.

3.1. Caching scenario

In this paper, a scenario is considered where a large-scale ISP operates a limited capacity CDN service by deploying caching points within its network, as depicted in Figure 1. However, it is important to note that the applicability of the proposed approach is not limited to this scenario. For example, the standardization effort in the CDNI working group of the IETF allows the proposed approach to be used on a larger scale and to communicate between different CDN vendors.

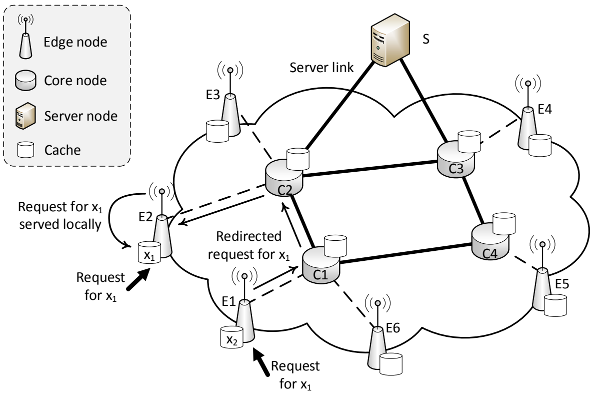


Figure 1: Overview of the Telco CDN service operated by the ISP.

In the considered scenario, the set of network nodes consists of edge nodes, which represent access networks connecting multiple users in the same region, and core nodes, which interconnect the different access networks. Each network node is equipped with caching capabilities, enabling a set of content items to be stored locally. These local caches could be external storage modules attached to routers or, with the advent of flash drive technology, integrated within routers.

All content requests are received at the edge nodes. If a requested content item is available in the local cache of the corresponding edge node, it is served locally. Otherwise, the request is redirected to one of the caches in the network where the requested item is stored. In case a copy of the item is not available in the ISP network, the request is served from the origin server outside of the network, hosted by the content provider. As depicted in Figure 1, the request for content x_1 received at edge node $E2$ is served locally, whereas the request for content x_1 received at edge node $E1$ is redirected to and served by node $E2$. In line with previous related research by Applegate et al. (2010), we assume that traffic is routed over the shortest path, which was shown to be more realistic than arbitrary routing (Garg and Konemann, 2007).

In this scenario, the ISP leases the caching space in its network to one or more content providers. Each content provider specifies the amount of caching capacity it wishes to lease for storing part of its content catalog, while the ISP decides which content items will be stored and where. The optimal placement of content in terms of network resource utilization depends on the geographical distribution of requests, the content popularity and the network topology. To minimize the resource utilization while simultaneously maximizing the total hit ratio (i.e., reducing the

[1] https://datatracker.ietf.org/wg/cdni

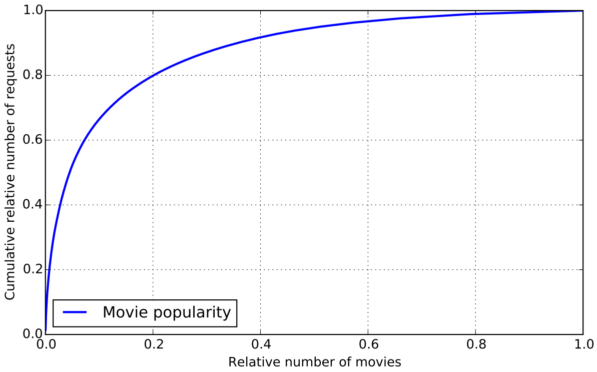


Figure 2: Popularity curve of the considered VoD trace.

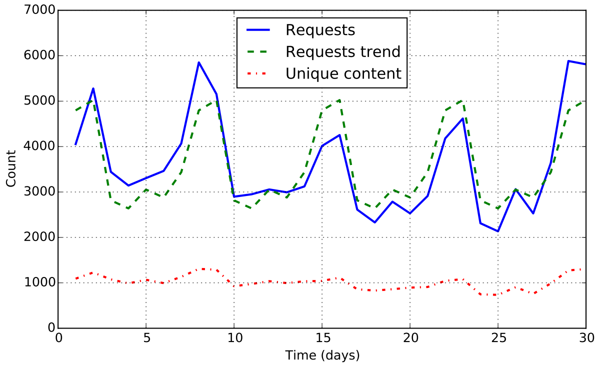


Figure 3: Request pattern and number of daily unique content items in the considered VoD trace.

number of requests that have to be served from outside the network), this paper proposes a hybrid multi-tenant cache management approach where the ISP controls the partitioning of the available storage space between the content providers. In the proposed hybrid approach, part of the available caching capacity is allocated and managed according to a proactive cache management strategy, while the rest is controlled through a reactive approach. The proactive placement algorithm is executed periodically by a central manager to allocate the proactive part of the capacity across the network based on the predicted value of the content popularity and its geographical distribution. In contrast, the reactive approach is applied at each cache independently and serves as a buffer to react locally to unpredicted popularity changes.

3.2. VoD trace characteristics

To evaluate the proposed approach, a request trace of the VoD service of a leading European telecom operator has been used. The trace was collected between Saturday February 6, 2010 and Sunday March 7, 2010 and contains monitoring information for a period of 30 days. Due to a failure of the probing nodes, a couple of hours of monitoring data was missing for both February 12, 2010 and February 19, 2010. These gaps have been filled by considering the request pattern of the same period in the previous week, mapped on the content popularity of the same period in the last day. The resulting trace consists of 108,392 requests for 5644 unique videos, originating from 8825 unique users spread across 12 cities. Figure 2 shows the popularity curve of the VoD trace. In this paper, all movies are considered to have an equal duration of 90 minutes and a bit rate of 1Mbit/s, resulting in a size of 675Mbyte for each video. The entire video catalog size thus amounts to about 3.64Tbyte. Each movie is requested in a segmented way, as is often the case in modern streaming technologies (e.g., Apple HLS, MPEG DASH), with a fixed duration of 1 second each. When multi-tenancy is considered, the set of movies in the VoD trace is uniformly split between the different content providers.

The daily number of requests and unique requested videos in the considered VoD trace are depicted in Figure 3. In this graph, a clear weekly pattern can be observed. The five peaks in the request pattern correspond to the five weekends, with increased activity on Friday, Saturday and Sunday. As only per-day aggregates are shown for the sake of visibility, the underlying diurnal trend cannot be observed. For Wednesdays and Sundays, the activity peak is situated between 4:30pm and 6:30pm, while for the other days of the week, the largest number of requests is reported between 8pm and 10pm. On average, 3613 requests for 1012 unique movies, i.e., about 18% of the total movie catalog, are monitored per day.

3.3. Request prediction

To be able to take proactive decisions concerning content placement, predictions about the content popularity in the considered period have to be made. Content popularity prediction is a complex issue that is outside the scope of this paper. Therefore, in this paper we apply a simple request prediction strategy based on the characteristics of the VoD trace described in the previous section.

The request pattern consists of multiple types of information: (i) the request intensity, i.e., the total number of requests in the network, over time, (ii) the geographical distribution of the requests, and (iii) the relative content popularity, i.e., the distribution of the requests amongst all the content items. As shown in Figure 3, a clear weekly pattern can be identified for the request intensity of the VoD trace. Based on this observation, the request intensity and the geographical distribution of requests of the same period of the previous week are used for predicting the request pattern in a specific period. However, given the highly dynamic nature of video popularity, such an approach cannot be used to predict content popularity. Although some request intensity trends can be identified for each day of the week, it is more likely that the change in popularity of a given content item will be more significant over a week than between two consecutive days. Based on the results reported in our previous work (Claeys et al., 2014b), we use the content popularity of the last 3 days to predict the popularity of specific content items as it was shown to result in the highest prediction accuracy.

To estimate the quality of the prediction, we define an accuracy metric as follows. When the total amount of leased capacity is equal to the combined size of y videos, the accuracy of the y most popular content items in the request prediction is crucial for the performance of the system, as these items are most likely to be cached. Therefore, we define the prediction accuracy in a specific period as the ratio between the relative number of requests for the predicted y most popular videos, denoted as r_{pred}, and for the actual y most popular videos, denoted as r_{act}, over that period. For example, we consider a total amount of leased capacity of y =100 videos. If the set of 100 videos that are predicted to be the most popular accounts for r_{pred}=50% of all requests in the considered period while the actual 100 most popular videos during that period amount for r_{act}=65% of the requests, the accuracy of the popularity prediction is said to be 76.92% $(= \frac{r_{pred}}{r_{act}} = \frac{50\%}{65\%})$.

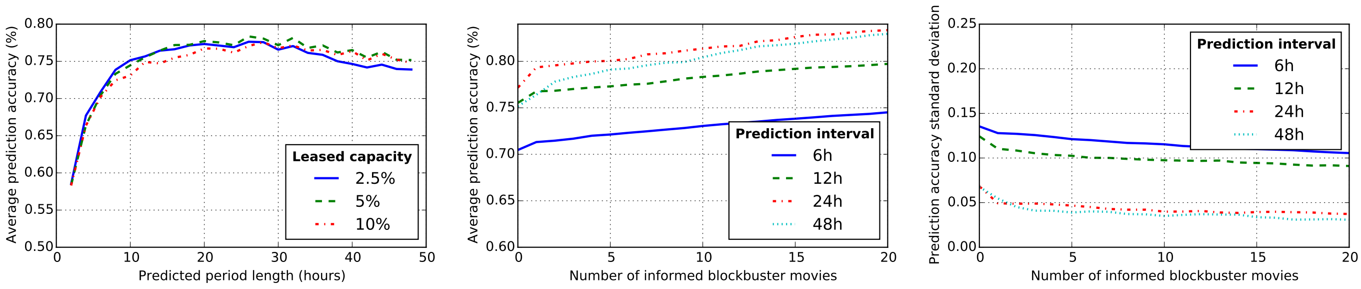
Figure 4a shows the average prediction accuracy for different amounts of leased capacity (expressed relatively to the total catalog size), in terms of the length of the predicted period. It can be observed that the prediction accuracy increases with the predicted period length, up to a length of 24h. This can be explained by the characteristics of the content popularity. Typically, requests for popular content are spread over the day while less popular content is only requested a few times a day at specific points in time. As the considered period of time decreases, the number of unique requested videos decreases and the steepness of the popularity curve significantly reduces. Therefore, the shorter the time period, the harder the content popularity prediction. In addition, due to the highly dynamic nature of the content popularity and the relatively short popularity lifetime of video content, the prediction accuracy degrades when the predicted period length exceeds 24h. Furthermore, it can be seen that the prediction accuracy is rather insensitive to the amount of leased capacity. However, this does not mean that the popularity of less popular content is as easy to predict as for more popular content. Because of the steep popularity curve of the considered VoD trace, shown in Figure 2, the vast majority of cache hits is for the most popular content, making these content items the decisive factor in the accuracy metric (which is based on cache hits). Therefore, the decreasing accuracy of predicting less popular content does not have a significant influence on the total prediction accuracy.

In the considered VoD request trace, a significant part of daily requests (17.92% on average) is for videos that were not requested in the recent history of 3 days. Even when longer history windows of up to one week are used, about 15% of the requests remains for unseen content. None of these requests can be predicted using pure history-based prediction techniques. The prediction of requests for new content is an active research area and previous research efforts have proposed to extract information about future blockbuster movies, for example based on activity trends on social media (Asur and Huberman, 2010), using collaborative filtering models (Koren, 2008) or relying on life span patterns of video popularity (Zhou et al., 2015; Yu et al., 2006). In this work, we assume that we are informed about a short list of the most popular blockbuster movies on a daily basis, regardless of the technique used to extract this information (i.e., a limited number of the most popular movies are assumed to be known by the system in advance, based on external information). Figure 4b and 4c show the influence of the number of informed blockbuster movies on the average prediction accuracy and its standard deviation, respectively. It can be seen that, for example in a scenario where predictions are made every 24h, even with a limited amount of 10 informed blockbuster movies (about 1% of the requested movies per day on average), the average prediction accuracy is increased by 5% while its standard deviation is decreased by 41% compared to the pure history-based prediction. This shows that besides increasing the average prediction accuracy, using knowledge about blockbuster movies significantly stabilizes the prediction accuracy over time. For example, information about 10 blockbuster movies lowers the difference between the highest and the lowest prediction accuracy over time from 28% to 18%.

3.4. Popularity prediction limitations

As described in the previous section, exogenous information can be used to predict blockbuster movies. However, this can only be done for a small fraction of popular content. A significant part of new content remains unpredicted. In addition, Figure 5 shows the average shifting probabilities in the VoD trace. The plot shows the percentage of content items that shifts from a given popularity category to another category on average. It is important to note that while the different popularity categories are unevenly sized in terms of number of contents, each category accounts for a similar number of requests. For example, the 2.5% most popular content items on a given day account for 29% of the requests on average, while the 80% least popular content items (i.e., category 100% in Figure 5) account for 33% of the requests on average. It can be observed that the content popularity in the considered trace is very volatile. For example, on average, 24.60% of the 2.5% to 5% most popular movies on a given day are not requested at all the next day. Furthermore, it can be deduced from Figure 5 that on average more than 13% of the top 2.5% most popular content was not requested the day before[2]. These findings strongly limit the possibilities of popularity prediction techniques. When considering time periods shorter than 24h, the fluctuation

[2] On average, 82.07% of all content is not requested on a single day, 0.40% of which is in the 2.5% most popular content the next day. This amounts to 0.33% of the total content catalog, or 13.2% of the 2.5% most popular content.

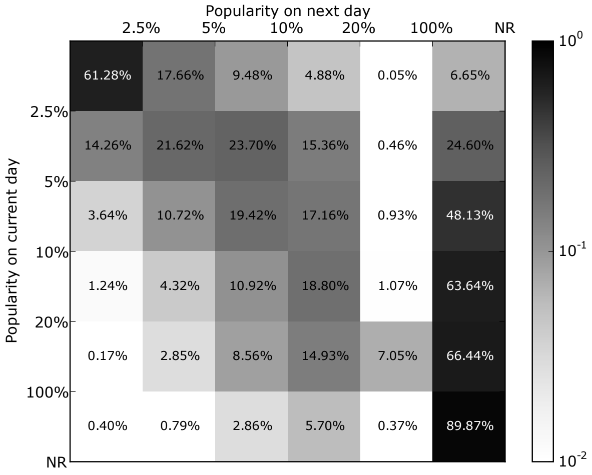


(a) Influence of the predicted period length on the average prediction accuracy with a history length of 3 days for different amounts of leased capacity.

(b) Influence of the knowledge about blockbuster movies on the average prediction accuracy with a history length of 3 days and a leased capacity of 5%.

(c) Influence of the knowledge about blockbuster movies on the standard deviation of the prediction accuracy with a history length of 3 days and a leased capacity of 5%.

Figure 4: Analysis of the prediction accuracy for the VoD trace.



	2.5%	5%	10%	20%	100%	NR
2.5%	61.28%	17.66%	9.48%	4.88%	0.05%	6.65%
5%	14.26%	21.62%	23.70%	15.36%	0.46%	24.60%
10%	3.64%	10.72%	19.42%	17.16%	0.93%	48.13%
20%	1.24%	4.32%	10.92%	18.80%	1.07%	63.64%
100%	0.17%	2.85%	8.56%	14.93%	7.05%	66.44%
NR	0.40%	0.79%	2.86%	5.70%	0.37%	89.87%

Figure 5: Average popularity shifting probabilities in the considered VoD trace (*NR: Not Requested).

in content popularity becomes even more significant, resulting in a lower prediction accuracy. Additionally, it can be observed from Figure 5 that the fluctuation in popularity grows for less popular content. Therefore, when less popular content is cached, it is more likely to be replaced in the next reconfiguration phase, resulting in additional migration overhead.

4. Hybrid cache management

As shown in previous work, proactive content placement can result in more efficient resource usage, compared to reactive cache management (Claeys et al., 2014b). However, it was shown that the achievable performance gain strongly depends on the accuracy of the popularity prediction. As described in Section 3.4, the characteristics of the VoD trace limit the accuracy of the request prediction. To deal with these limitations, a hybrid caching approach is proposed, applying both proactive and reactive caching techniques. A proactive content placement is performed periodically, based on a prediction of the content popularity, while part of the capacity is used for reactive caching in order to react to unexpected popularity fluctuations. It is important to note that even though the performance of both the proactive approach and the hybrid approach will be influenced by the quality of the request prediction, the hybrid approach will always benefit from the reactive caching part to deal with new popular content that will inevitably be missed by the request prediction strategy, as discussed in Section 3.4. Therefore, even though the optimal cache division might change, the hybrid approach will always outperform both the proactive and reactive caching approaches, or at least perform equally as good as the best of the two approaches.

First, some general notations are introduced in Section 4.1. Next, the hybrid cache division strategy is presented in Section 4.2. Finally, Section 4.3 describes the proactive placement algorithm.

4.1. General notations

To support the description of the algorithm, we first introduce some general notations used to represent the characteristics of the considered scenario. The notations are summarized in Table 1. The network topology is modeled as a directed graph $G = (N, L)$ with N and L representing the set of nodes and links, respectively. The set of nodes contains both the nodes N_{ISP}, belonging to the ISP network, and an external server node S, logically representing the Internet, containing all content of all providers. N_{ISP} can further be divided in a set of core nodes N_C and edge nodes N_E. The links L can be divided into a set of links, L_S, connected to the external server (i.e., the ingress links), and ISP-managed links L_{ISP}, connecting core and edge nodes. For each node $n \in N$, we define a caching capacity $c_n \in \mathbb{N}^+$ and a set of incoming and outgoing links, denoted by $I_n \subseteq L$ and $O_n \subseteq L$, respectively. For every link $l \in L$, the available bandwidth capacity is denoted by $c_l \in \mathbb{N}^+$. The routing strategy applied in the network is represented by a forwarding path $R_{s,d} \subseteq L$, for every

Table 1: Summary of the general notations.

Notation	Description
N	Set of network nodes.
$S \in N$	Server node.
$N_{ISP} \subset N$	Set of ISP-managed nodes.
$N_C \subset N_{ISP}$	Set of core nodes.
$N_E \subset N_{ISP}$	Set of edge nodes.
L	Set of network links.
$L_S \subset L$	Set of ingress links.
$L_{ISP} \subset L$	Set of ISP-managed links.
$c_n \in \mathbb{N}^+$	Storage capacity of node $n \in N$.
$I_n \subseteq L$	Set of incoming links of node $n \in N$.
$O_n \subseteq L$	Set of outgoing links of node $n \in N$.
$c_l \in \mathbb{N}^+$	Bandwidth capacity of link $l \in L$.
$R_{s,d} \subseteq L$	Routing path from $s \in N$ to $d \in N$.
P	Set of content providers.
$d_p \in \mathbb{N}^+$	Caching space leased by $p \in P$.
O^p	Content items offered by $p \in P$.
$O = \bigcup_{p \in P} O^p$	Total set of offered content items.
$s_o \in \mathbb{N}^+$	Size of content item $o \in O$.
$b_o \in \mathbb{N}^+$	Bit rate of content item $o \in O$.

source-destination pair $(s,d) \in N \times N$. The forwarding path can be divided into a set of server links $R^S_{s,d} \subseteq L_S$, containing the links in the forwarding path connected to the external server node S, and a set $R^{ISP}_{s,d} \subseteq L_{ISP}$ containing the other links in the forwarding path, inside the ISP network.

A set of content providers P lease caching space from the ISP. For each content provider $p \in P$, the leased amount of caching space and the set of offered content items are denoted by $d_p \in \mathbb{N}^+$ and O^p, respectively. $O = \bigcup_{p \in P} O^p$ represents the entire set of offered content. Every content item $o \in O$ has an associated size $s_o \in \mathbb{N}^+$ and bit rate $b_o \in \mathbb{N}^+$.

4.2. Cache division

In the proposed hybrid caching approach, a relative part $\lambda \in [0;1]$ of the leased capacity for each provider is used as a reactive cache, for example applying the Least Recently Used (LRU) replacement strategy. The remaining part $(1-\lambda)$ of the leased capacity is used for proactive placement. The value of λ is called the reactive ratio of the hybrid caching system. When a provider $p \in P$ leases a capacity of d_p bytes, then $\lambda \times d_p$ bytes are used for reactive caching, while the remaining $(1-\lambda) \times d_p$ bytes are used for proactive placement.

Instead of using a fixed reactive ratio on every node, the reactive part of the leased capacity is uniformly distributed across the entire topology, proportional to the storage capacity of the nodes. In this way, the entire network is provided with reactive caching capacity to deal with unexpected popularity fluctuations, independently of the geographical distribution of the predicted request pattern. When $\lambda \times d_p$ bytes are used for reactive caching for provider $p \in P$, on every node $n \in N_{ISP}$, the number of bytes allocated for reactive caching for provider p can be calculated using Equation (1).

$$c_n \times \frac{\lambda \times d_p}{\sum_{n' \in N_{ISP}} c_{n'}} \tag{1}$$

The remaining leased capacity of $(1-\lambda) \times d_p$ bytes are used for proactive placement. The allocation of this capacity, spread across the network, is based on the ILP model, described in Section 4.3. For each node $n \in N_{ISP}$, the total capacity available for proactive placement c^{pro}_n can be calculated using Equation (2).

$$c^{pro}_n = c_n - \frac{c_n \times \lambda \times \sum_{p \in P} d_p}{\sum_{n' \in N_{ISP}} c_{n'}} \tag{2}$$

Using this hybrid approach, the geographical distribution of the allocation of proactive caching capacity is based on the predicted request pattern, allocating more capacity where more requests are expected. However, by uniformly distributing the reactive capacity, every area in the network is provided with some backup capacity to deal with possible errors in the request prediction or rapid popularity fluctuations.

4.3. Proactive placement

The ILP, used to model the considered problem, is a modified version of the ILP described in our previous work (Claeys et al., 2014b). The modifications handle the difficulty in accurately predicting the exact time points at which a request will be sent. Therefore, while the ILP model in our previous work required the prediction of specific time points of requests, the modified ILP requires more realistic aggregated request predictions (i.e., the total number of predicted requests over the considered period instead of exact timing information of each request). Furthermore, the current placement configuration is taken into account in order to limit the migration overhead.

In the remainder of this section, the input values, the decision variables, the objective functions and the constraints of the ILP are presented.

4.3.1. Input values

The objective of the proposed approach is to periodically compute a new caching configuration based on the estimation of content popularity and geographical distribution of requests for the next provisioning interval. As such, besides the characteristics of the network topology, the content catalog and the leased capacity for each provider, a prediction of the request pattern for the considered time interval is required by the algorithm at each reconfiguration step to determine a new content placement and server selection strategy. In addition to the general notations introduced in the previous section, we note $r_{o,d} \in \mathbb{N}$ as the predicted number of requests for content $o \in O$, originating from edge node $d \in N_E$ in the considered provisioning interval. $N^o \subseteq N_E$ represents the set of edge nodes requesting content $o \in O$ in the considered interval. To be able to take into account the current placement configuration, an additional notation $X_n \subseteq O$ is introduced for each node $n \in N$, representing the set of content items that are currently stored at node n.

4.3.2. Decision variables

A solution to the content placement problem is translated into binary decision variables $x_{n,o} \in \{0, 1\}$ defining if a node $n \in N$ is used to store content $o \in O$. In addition, auxiliary decision variables $z_{n,o,d} \in \{0, 1\}$ are introduced to represent the server selection strategy. These variables define if a node $n \in N$ is used to store content $o \in O$ to be delivered to edge node $d \in N_E$.

4.3.3. Objective function

Different optimization criteria have been considered in the literature (Sharma, 2013; Laoutaris et al., 2006; Applegate et al., 2010). Even though other metrics such as delivery delay minimization or link load minimization can easily be integrated in the problem formulation, in this paper we focus on reducing the ISP network resource usage. As such, we define the optimal solution to the problem as the one minimizing the bandwidth usage inside the ISP network. While calculating the bandwidth usage, a weighting factor $\alpha \in [0; 1]$ can be used to define the importance of ingress link usage. In this way, the objective function can be tuned to purely optimize bandwidth usage or to focus on optimizing the hit ratio. Low values of α steer the ILP to minimize the bandwidth usage within the ISP network. In contrast, higher values of α result in minimizing the bandwidth usage on the ingress link, yielding a maximization of the hit ratio. In this work, we focus on optimizing the total bandwidth usage, considering both the server link usage and the bandwidth usage inside the ISP network. Therefore, a value of $\alpha = 0.5$ is used, unless otherwise stated. The link weight ω_l of a link $l \in L$ is shown in Equation (3).

$$\omega_l = \begin{cases} \alpha & \text{if } l \in L_{ISP} \\ 1 - \alpha & \text{if } l \in L_S \end{cases} \tag{3}$$

In order to minimize the total bandwidth usage in the ISP network, the bandwidth usage incurred by the video streaming sessions is modeled in the basic objective function (Equation (4)). However, besides the video streaming sessions, the periodic content migrations significantly influence the total bandwidth usage. Therefore, this migration overhead can be taken into consideration in the objective function as well, based on the current storage configuration as shown in Equation (5). Content that has to be placed at a specific node is fetched from the server node S. In the overhead-aware objective function, shown in Equation (6), both the video streaming bandwidth and the migration overhead are taken into account. It can be seen that only the streaming bandwidth objective function (Equation (4)) depends on predicted values (i.e., the predicted request intensities $r_{o,d}$). All remaining variables have known values. In the remainder of this paper, we will refer to Equation (4) and Equation (6) as the basic objective function and the overhead-aware objective function, respectively, both of which are subject to minimization in the considered ILP.

$$BW_{str} = \sum_{n \in N} \sum_{o \in O} \sum_{d \in N^o} \sum_{l \in R_{n,d}} \omega_l \times r_{o,d} \times s_o \times z_{n,o,d} \tag{4}$$

$$BW_{mig} = \sum_{n \in N} \sum_{o \in O \setminus X_n} \sum_{l \in R_{S,n}} \omega_l \times s_o \times x_{n,o} \tag{5}$$

$$BW = BW_{str} + BW_{mig} \tag{6}$$

4.3.4. Constraints

Multiple constraints are considered to define the set of valid solutions to the considered optimization problem. First of all, auxiliary constraints are introduced to formalize the relationship between the x and z decision variables. The constraints presented in Equation (7) and Equation (8) specify that content $o \in O$ is stored at node $n \in N$ if and only if at least one edge node $d \in N_E$ requests o from n.

$$\forall n \in N, \forall o \in O, \forall d \in N_E : \qquad z_{n,o,d} \leq x_{n,o} \tag{7}$$

$$\forall n \in N, \forall o \in O : \quad x_{n,o} \leq \sum_{d \in N_E} z_{n,o,d} \tag{8}$$

A valid solution to the optimization problem is so that the caching space reserved for each content provider $p \in P$ is at most equal to the part of the leased capacity assigned for proactive content placement, while satisfying the storage capacity limitations. These constraints are modeled in Equation (9) and Equation (10), respectively.

$$\forall p \in P : \quad \sum_{n \in N_{ISP}} \sum_{o \in O_p} s_o \times x_{n,o} \leq (1 - \lambda) d_p \tag{9}$$

$$\forall n \in N_{ISP} : \quad \sum_{o \in O} s_o \times x_{n,o} \leq c_n^{pro} \tag{10}$$

Finally, constraint Equation (11) ensures that every request is served from exactly one location.

$$\forall o \in O, \forall d \in N^o : \sum_{n \in N} z_{n,o,d} = 1 \tag{11}$$

Periodically solving the ILP results in a storage profile represented by the values of $x_{n,o}$ and a server selection strategy represented by the values of $z_{n,o,d}$, which minimizes the objective function in Equation (4) or Equation (6), while satisfying the constraints in Equation (9) – Equation (11). Every request from edge node $d \in N_E$ for content $o \in O$ is served from node $n \in N$ where $z_{n,o,d} = 1$, using the shortest path $R_{n,d}$.

For the reader's reference, the algorithm-specific notations are summarized in Table 2.

Table 2: Summary of the algorithm-specific notations.

Notation	Description
$\lambda \in [0;1]$	Reactive ratio.
$c_n^{pro} \in \mathbb{N}^+$	Proactive capacity of node $n \in N_{ISP}$.
$r_{o,d} \in \mathbb{N}$	Nr. of requests for $o \in O$ from $d \in N_E$.
$N^o \subseteq N_E$	Edge nodes requesting $o \in O$.
$X_n \subseteq O$	Content items stored at node $n \in N$.
$x_{n,o} \in \{0,1\}$	$o \in O$ placed on $n \in N$.
$z_{n,o,d} \in \{0,1\}$	$o \in O$ placed on $n \in N$ for $d \in N_E$.
$w_l \in [0;1]$	Weight of link $l \in L$.

5. Evaluation setup

To thoroughly evaluate the performance of the proposed approach, a topology based on the GÉANT network[3] is used, consisting of 23 nodes. As described in Section 3.2, the considered VoD request trace contains 12 cities, which are mapped on 12 edge nodes ($N_E = \{E1, ..., E12\}$). One node is selected as the external server node S, storing all content of all content providers. The 10 remaining nodes are modeled as core nodes ($N_C = \{C1, ..., C10\}$). The resulting topology is shown in Figure 6. In this topology, shortest path routing based on hop count is applied.

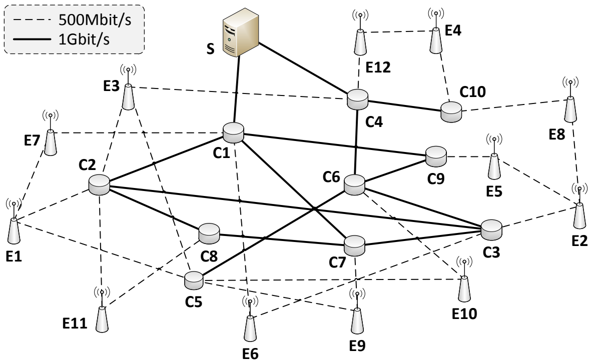


Figure 6: Evaluated GÉANT-based topology.

As preliminary evaluations have shown that the node capacities have limited influence on the performance of the approach, unless otherwise stated, the storage capacity of each core node was set high enough to be able to accommodate the leased capacity of all tenants throughout the evaluations. Concretely, for every core node $n \in V_C$, the capacity is defined as $c_n = \sum_{p \in P} d_p$. The capacity of the edge nodes is fixed to half of the capacity of the core nodes, i.e., $c_n = 0.5 \times \sum_{p \in P} d_p, \forall n \in V_E$. The bandwidth capacity of the links interconnecting core nodes and links connected to the server node is set to 1Gbit/s while all other links have a bandwidth capacity of 500Mbit/s.

Throughout the evaluations, the performance of the proposed approach is compared to a purely reactive approach. For the reactive approach, the LRU replacement strategy is applied, while the leased capacity is uniformly spread across the network, proportional to the node capacities. All requests are sent to the origin server S, applying reactive caching. Consequently, the reactive approach can result in a configuration where only parts of a movie are available at a given node. In contrast, the proactive approach either places an entire movie at a specific node or does not store it there at all.

Experiments have been performed for all of the 30 days in the VoD trace, while only evaluating the last 23 days (from February 13, 2010 to March 7, 2010). The first 7 days of the trace were used for obtaining the request prediction used in the proactive approach and serve as a cache-warming phase for the reactive approach. To periodically determine the proactive placement configuration, the ILP is solved using the *IBM ILOG CPLEX Optimization Studio 12.4* solver.

6. Evaluation results

To characterize the performance of the proposed approach, multiple performance indicators are evaluated:

- **Hit ratio:** the relative amount of segment requests that could be served from within the ISP network (in %).
- **Average bandwidth usage:** the average bandwidth usage in the entire ISP network (in Mbit/s), including both the video streaming bandwidth and the content migration bandwidth induced by the proactive placement.
- **Migration overhead:** the total amount of data transfer induced by the proactive content placement (in Gbyte).
- **Average hop count:** the average number of links a segment crosses between its storage location and the requesting client. This metric indicates how close the relevant movies are stored to the end-users.

First, the different aspects of the proposed hybrid cache management approach are evaluated in Section 6.1. Next, the performance of our approach is compared to a purely proactive and a purely reactive cache management approach in Section 6.2.

6.1. Influence of the system parameters

In this section, the optimal configuration of the proposed approach is determined, starting from a purely proactive approach ($\lambda = 0$) using the basic objective function (Equation (4)) without blockbuster movie knowledge and gradually evaluating the added value of more complexity. Unless stated differently, in all of the evaluations, the content catalog has been uniformly distributed between two content providers, each leasing storage capacity to accommodate 2.5%, 5% or 10% of their total movie catalog.

[3] GÉANT Project - http://www.geant.net

6.1.1. Proactive placement frequency

The frequency at which a new proactive placement configuration is computed defines a trade-off between optimality and overhead. While more frequent reconfigurations allow the system to be more reactive with respect to changes in the request pattern, this comes at the cost of more frequent content migrations. Figure 7 shows the influence of the length of the time interval between subsequent content placements on multiple performance indicators. For these evaluations, a purely proactive approach has been applied using the basic objective function without blockbuster movie knowledge. As could be expected, Figure 7a shows that the content migration overhead strongly increases when more frequent reconfigurations are performed, independently of the leased capacity.

Remarkably however, less frequent reconfigurations also result in a higher performance in terms of hit ratio and average bandwidth usage, as can be seen in Figure 7b and Figure 7c, respectively. This counter-intuitive result can be explained by the characteristics of the VoD request trace and the applied prediction strategy, discussed in Section 3.3. As was shown in Figure 4a, the prediction accuracy significantly decreases when predictions are made for shorter periods of time. In the case of more frequent reconfigurations, the low prediction accuracy results in more requests that have to be fetched from the origin server, resulting in longer routing paths, which lead to higher bandwidth usage and lower hit ratio. Similar observations can be made for the average hop count, which is on average 4.87% lower when placements are performed every 24h compared to every 6h (graphs omitted due to space limitations). Given the performance degradation in terms of prediction accuracy for periods longer than 24h (as described in Section 3.3) and the strong diurnal pattern in the request trace, reconfiguration frequencies lower than once every 24h have not been considered in the analysis.

The influence of the proactive placement frequency on the average time needed by the CPLEX solver to solve the ILP is presented in Figure 8. The error bars show the standard deviation. It can be seen that the execution time significantly increases with the content placement interval, as an increasing number of requests results both in a higher number of decision variables and constraints and an increased complexity for the objective function of the ILP. The high standard deviation values are due to the high fluctuation of the request intensity over time as shown in Figure 3. However, it can be seen that in the considered scenario, the average time needed to solve the ILP is reasonable compared to the placement interval (e.g., 15s every 24h on average for a leased capacity of 10%).

In the remainder of this paper, we assume that the proactive placement is performed every 24h given that this frequency results in the best performance. Furthermore, this allows the ISP to perform content migrations in off-peak hours (e.g., at night).

Table 3: Stability of the content popularity.

Amount[*]	**Stability**
2.5%	61.28%
5%	57.41%
10%	53.89%
20%	51.41%
100%	37.30%

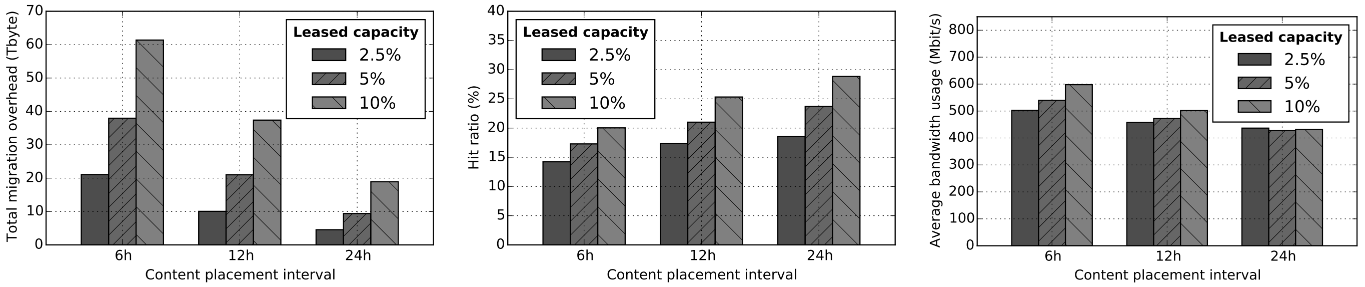
[*]Relative to the total number of requested contents on given day.

6.1.2. Overhead-aware placement

The benefits of adding overhead-awareness to the proactive placement are evaluated in a purely proactive scenario without blockbuster movie knowledge. Figure 9 shows the relative performance of the approach using the overhead-aware objective function compared to the performance of the approach using the basic objective function. It can be seen that taking into account the migration overhead in the placement decisions drastically reduces the migration overhead by 37.49% on average. Furthermore, the reduction in terms of migration overhead results in 3.33% less bandwidth usage on average. Given that, on average, the bandwidth introduced by the content migration only amounts to about 10% of the total bandwidth usage, this bandwidth reduction can be fully attributed to the reduced migration bandwidth. The deviation of the bandwidth usage introduced by the streaming sessions is limited to less than 1%. With a relative gain of 1.37% and 0.3% respectively, the influence on the performance in terms of hit ratio and average hop count is negligible. Furthermore, due to its higher complexity, using the overhead-aware objective function results in a (limited) increase of 6.99% in terms of the average time needed to solve the ILP.

As can be observed in Figure 9, the benefits of introducing overhead-awareness to the system increase when more capacity is leased. The explanation for this observation can be found in the shifting characteristics of the VoD request trace. Based on Figure 5, we can calculate the average percentage of the x most popular content on a given day that still belongs to the x most popular content on the next day with x being equal to 2.5%, 5%, 10%, 20% or 100% of the total number of content requested on a given day. The resulting values, referred to as the *stability* of the content popularity, are shown in Table 3. It can be seen that the stability of the popularity decreases when a higher amount of content is considered.

When more capacity is leased, content items with a lower popularity can also be cached proactively. However, as demonstrated, these items change more frequently than the popular ones. By adding overhead-awareness to the system, the placement algorithm can balance the bandwidth gain of placing the less popular content in the network against the cost of migrating the content. For popular content, the streaming bandwidth reduction exceeds



(a) Influence of the proactive placement frequency on the total migration overhead.

(b) Influence of the proactive placement frequency on the hit ratio.

(c) Influence of the proactive placement frequency on the average bandwidth usage in the network.

Figure 7: Influence of the proactive placement frequency in a purely proactive scenario using the basic objective function without blockbuster movie knowledge.

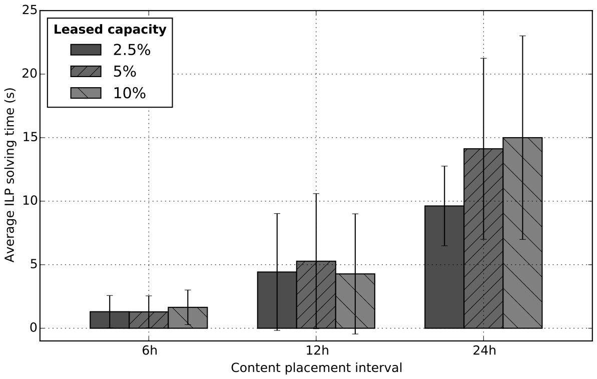


Figure 8: Influence of the proactive placement frequency on the average time needed to solve the ILP.

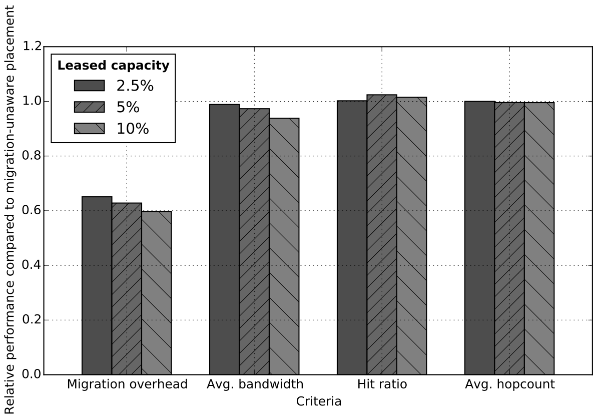


Figure 9: Influence of overhead-awareness in proactive content placement on different evaluation criteria.

the migration cost, while this may no longer be the case for less popular content. Therefore, overhead-awareness adds more value when less popular content can be placed in the network.

In the remainder of the evaluations, the overhead-aware objective function is used to calculate the proactive placement configuration.

6.1.3. Blockbuster movie knowledge

As shown in Section 3.3, the accuracy of the popularity prediction can be increased based on the availability of exogenous information about blockbuster movies. The influence of this information on the performance of the proposed approach in terms of hit ratio is shown in Figure 10. As expected, the perceived hit ratio increases with the number of known blockbuster movies. It is also interesting to see that the performance increase in terms of hit ratio exceeds the increased prediction accuracy. For example, when the leased capacity is equal to 5% of the total content catalog, knowledge about 5 blockbuster movies leads to a relative increase of 3.97% in terms of prediction accuracy (Figure 4b), while the relative improvement in terms of perceived hit ratio amounts to 17.22% (i.e., an absolute increase of 4.18% compared to the original hit ratio of 24.27%, see Figure 10).

To clarify this observation, it is important to note that the goal of the proactive placement is to minimize the bandwidth usage. Therefore, storing duplicates of popular content can be preferred over storing as much unique content items as possible, causing the perceived hit ratio to be lower than what could be achieved based on the prediction accuracy. This means that some of the less popular content items are not selected by the placement algorithm, even though they could be cached given the value of the prediction accuracy metric. The information about blockbuster movies can be used to derive the real popularity of some of these content items, resulting in a higher hit ratio and this without affecting the prediction accuracy.

In terms of average bandwidth usage and average hop count, a relative performance increase of 5.79% and 5.65%

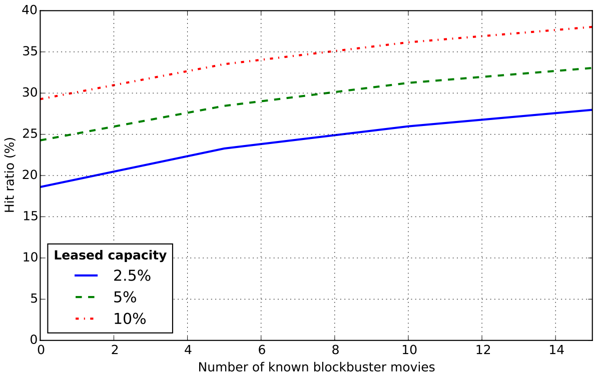


Figure 10: Influence of blockbuster movie knowledge on the performance in terms of hit ratio.

is achieved respectively, while the migration overhead is rather unaffected (decrease of 0.12%). In what follows, unless otherwise stated, it is assumed that the system is informed about the 5 most popular blockbuster movies by an external source, as described in Section 3.3.

6.1.4. Hybrid cache division

Up to now, a purely proactive approach has been followed. In Section 4, it was argued that applying a hybrid caching approach can significantly improve the performance of the system. To evaluate the influence of the cache division, the reactive ratio λ has been varied between $\lambda = 0.0$ (i.e., a purely proactive approach) and $\lambda = 1.0$ (i.e., a purely reactive approach) in a scenario where knowledge about 5 blockbuster movies is available for the request prediction and the overhead-aware objective function is used. The proactive placement configuration is recalculated every 24h.

Figure 11 shows the influence of the reactive ratio λ on the hit ratio, the average bandwidth usage and the average hop count of the proposed approach. It can be seen that, in the majority of the cases, both the purely proactive and the purely reactive approach are outperformed by the hybrid scheme in terms of hit ratio, average bandwidth usage and average hop count (i.e., for 79%, 94% and 78% of the λ values, respectively). Given that the lease constraint (Equation (9)) of the placement ILP limits the amount of placed content to the size of $(1-\lambda)d_p$ for each provider $p \in P$, the migration overhead decreases linearly with increasing reactive ratio λ (graph omitted due to space limitations).

The reactive ratio that gives the optimal performance in the considered scenario depends on the considered metric. However, as can be seen in Figure 11, a wide range of ratios has a performance close to the performance of the optimal ratio for each of the metrics. To be able to select a single ratio, the different metrics should be optimized simultaneously. Therefore, for each evaluated metric M and each reactive ratio λ, we define the deviation from the optimum δ_λ^M as shown in Equation (12). It is calculated as the ratio between the performance of reactive ratio λ in terms of metric M, denoted as ϕ_λ^M, and the optimal performance for that metric, ω^M. ω^M is defined as the maximum or minimum value $\phi_{\lambda'}^M, \forall \lambda' \in [0;1]$, depending on whether metric M should be maximized or minimized respectively.

$$\delta_\lambda^M = \left| 1 - \frac{\phi_\lambda^M}{\omega^M} \right| \tag{12}$$

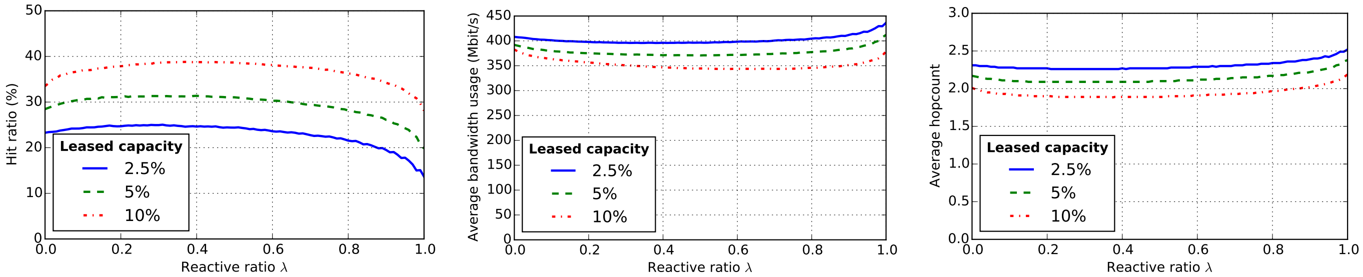
For each reactive ratio λ the average value of δ_λ^M for the considered metrics (M_1: hit ratio, M_2: average bandwidth usage and M_3: average hop count) can be calculated as $\Delta_\lambda = \frac{1}{3}\left(\delta_\lambda^{M_1} + \delta_\lambda^{M_2} + \delta_\lambda^{M_3}\right)$, representing how close the performance is to the optimal performance on average. Figure 12 shows the average deviation from the optimum for different amounts of leased capacity. It can be seen that the optimal ratio slightly increases with an increasing amount of leased capacity. This can be explained by the growing performance of LRU for bigger caches and the reduced stability of the content popularity for less popular content. On average, the optimal performance is achieved with a reactive ratio $\lambda = 0.41$.

6.1.5. Reactive ratio adaptation

As described in Section 1, the performance of the proactive placement approach strongly depends on the quality of the popularity prediction. Given that the prediction accuracy fluctuates over time, the optimal reactive ratio λ is also subject to change. Therefore, changing the reactive ratio over time might be considered. For example, Figure 13 shows the optimal reactive ratio λ over time at a granularity of 6h for different amounts of leased capacity in order to maximize the hit ratio. It can be seen that the optimal ratio indeed strongly fluctuates over time without following any well-defined pattern.

Figure 14 shows the relative hit ratio when using a fixed reactive ratio $\lambda = 0.41$ compared to using the optimal ratios shown in Figure 13 at a granularity of 6h. It can be seen that the hit ratio using a fixed reactive ratio is close to the optimal adaptive performance, with an average deviation of only 2.90%. Furthermore, it is important to note that the optimal performance is a theoretical optimum that would require to be able to determine the optimal reactive ratio at any point in time.

Adaptively changing the reactive ratio λ would introduce a lot of added complexity to the system, as it requires to be able to monitor the accuracy of the request prediction. Furthermore, given the noisy pattern shown in Figure 13, the accuracy of predicting the optimal ratio would be limited, resulting in a lower performance gain than theoretically possible. In addition, changing the reactive ratio demands a reconfiguration of both the reactive and the proactive cache at every single node in the network, again introducing additional migration overhead. Given this high level of complexity and the limited theoretical performance increase that could be achieved in the



(a) Influence of the reactive ratio λ on the hit ratio.

(b) Influence of the reactive ratio λ on the average bandwidth usage.

(c) Influence of the reactive ratio λ on the average hop count.

Figure 11: Influence of the reactive ratio λ on the performance of the hybrid caching approach.

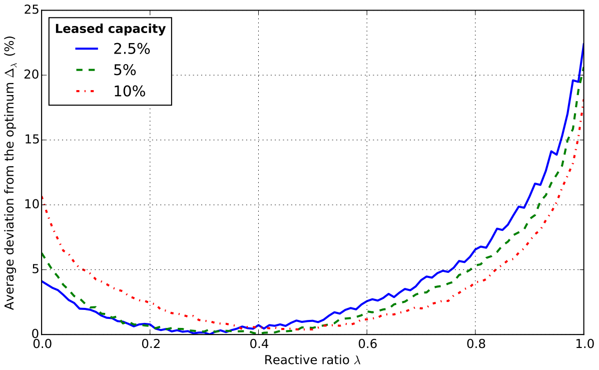


Figure 12: Influence of the reactive ratio λ on the average deviation from the optimum.

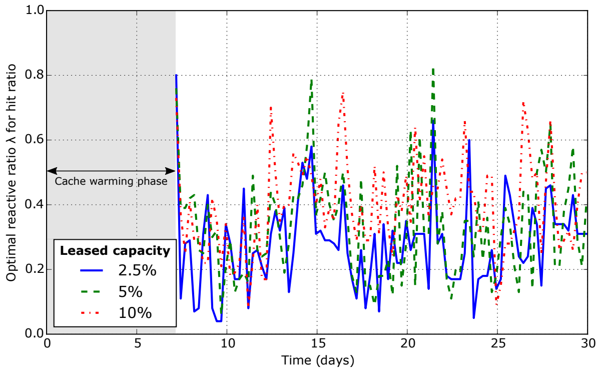


Figure 13: Optimal reactive ratio λ over time.

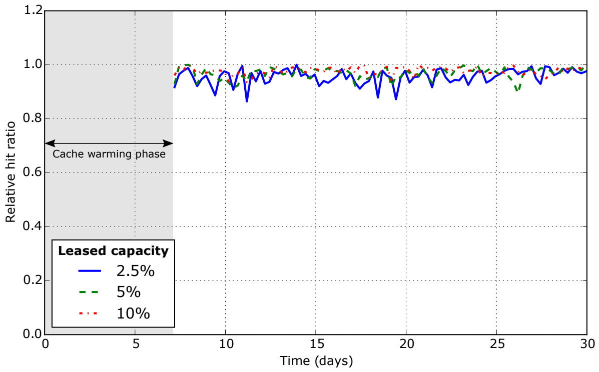


Figure 14: Relative hit ratio with $\lambda = 0.41$ compared to an adaptive reactive ratio.

considered scenario, a fixed reactive ratio is proposed in practice.

6.1.6. Number of tenants

Up to now, all the evaluations have been performed for two content providers. To show the applicability of the proposed approach in the general case, evaluations have been performed for multiple tenants leasing storage capacity from the ISP. For this purpose, the VoD movie catalog of 5644 movies has been uniformly split amongst the tenants (i.e., each movie in the catalog has been randomly assigned to one of the tenants with equal probability). Even though the characteristics of the request trace and the content catalog are unchanged and the total amount of leased capacity across the tenants remains the same in all the evaluations (i.e., independent of the number of tenants), the results of the proactive placement approach are influenced by the multi-tenancy. Given the leased capacity constraint (Equation (9)) for each provider, a fixed amount of content items has to be stored for each provider to meet their capacity requirements, even though other providers might have more popular content. In contrast, in a scenario with a single content provider, only the globally most popular content will be proactively placed. Figure 15 shows the influence of the number of tenants on the hit ratio, the average bandwidth usage and the average hop count. It can be seen that the influence of increasing the number of tenants from 1 to 10 is limited to a relative performance degradation of 4.54%, 2.78% and 2.72%, respectively. These values are in line with the degree of performance degradation obtained in both the purely proactive and reactive cases. The results show that the proposed approach can be applied with a fixed parameter configuration which is independent of the number of tenants, without incurring any significant performance degradation.

6.1.7. Server link weight

In all of the above evaluations, the server link weight α has been fixed to 0.5 in order to optimize the bandwidth usage. As explained in Section 4.3, the value of α can be used to steer the placement decisions between optimizing the bandwidth usage and optimizing the hit ratio. For the sake of completeness, the influence of the server link weight α on the performance of the proposed approach is evaluated.

Figure 16 shows the influence of the server link weight α on the performance of the hybrid caching approach with a reactive ratio of $\lambda = 0.41$, using the overhead-aware objective function and a request prediction without knowledge of blockbuster movies. The proactive placement is performed every 24h. Using a value of $\alpha > 0.5$ steers the ILP to minimize the bandwidth usage on the server ingress link. Minimizing the server link usage directly results in a higher hit ratio, as can be seen in Figure 16a. In terms of bandwidth usage, Figure 16b shows a clear increase for the values of $\alpha > 0.5$. When the focus is given to the ingress link bandwidth, the algorithm prefers storing more unique movies in the network instead of duplicating the most popular ones. This results in longer routing paths for the popular content, introducing additional bandwidth usage. Finally, Figure 16c shows that for large values of α, the migration overhead is significantly reduced. This behavior is due to the fact that the overhead-aware objective function is used. As all migrated content originates from the server node, all migration overhead is routed over a server ingress link, having a strong impact on the objective of the ILP. Placing content inside the ISP network can decrease the streaming bandwidth usage, but for large values of α, the bandwidth inside the ISP network is not decisive. Consequently, large values of α will steer the algorithm to migrate less content items.

6.1.8. Capacity limitations

Up to now, all evaluations have been performed using overprovisioned node capacities, i.e., sufficiently high to allocate all of the leased capacity on a single core node. This gave the algorithm complete freedom on how to distribute the capacity. To demonstrate the general applicability of the proposed approach, simulations have been performed in a scenario where the total storage capacity available in the network is equal to the capacity leased by all of the tenants. As with the previous evaluations, the capacity of the edge nodes is set to half of the capacity of the core nodes. In this configuration, the algorithm has less freedom to decide how to place the content given that the space available at each node is limited.

Figure 17 shows the performance of the proposed approach when two tenants lease storage capacity in a scenario with limited node capacities, relative to the scenario with overprovisioned nodes. It can be seen that, due to the restricted freedom of the algorithm, the average bandwidth usage is slightly increased by 2.67%. Furthermore, the limited capacity causes the content to be stored further away from the end user, resulting in a slight increase of 3.85% in terms of average hop count. However, it is interesting to note that the number of duplicated content items is reduced, resulting in the average hit ratio being increased by 2.85%.

6.2. Performance comparison

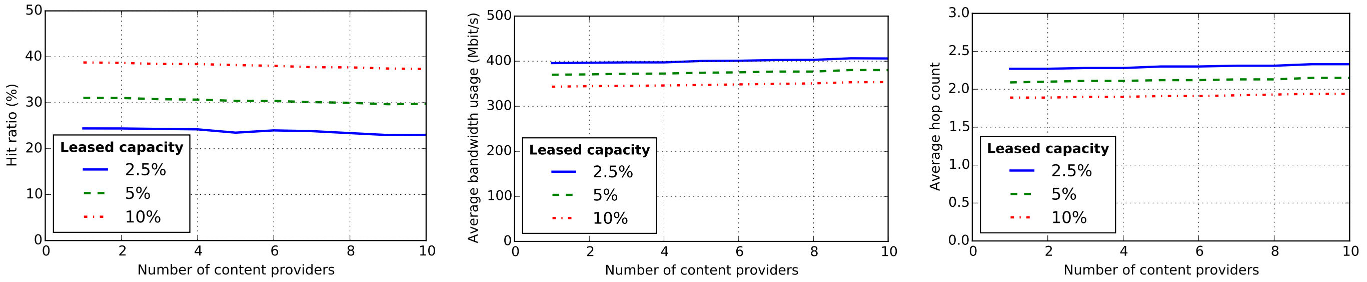
In this section, we compare the performance of the proposed approach to a purely proactive and reactive scheme based on the preferred parameter configuration (i.e., the one that optimizes the performance of the hybrid scheme) derived from the analysis in Section 6.1:

- Reconfiguration interval of 24h
- Overhead-aware objective function
- Fixed reactive ratio $\lambda = 0.41$

In this section, we consider a scenario where no knowledge about blockbuster movies is available and a scenario where 5 blockbuster movies are known *a priori*. In these evaluations, two content providers lease some caching capacity from the ISP with overprovisioned node capacities.

Figure 18 shows the performance of the proposed approach, relative to a purely reactive approach in a scenario where no information about blockbuster movies is available. It can be seen that the benefits of the proposed approach are higher for lower amounts of leased capacity. This can be explained by the combination of the increasing performance of the LRU cache replacement strategy for larger cache sizes and the decreasing stability of the popularity for higher amounts of leased capacity, as described in Section 6.1.2. For the evaluated VoD scenario and topology, the relative performance increase on average amounts to 5.35%, 42.96% and 8.15% in terms of average bandwidth usage, hit ratio and average hop count respectively.

In the same scenario, the performance of the proposed approach is compared to a purely proactive approach in Figure 19. As opposed to the comparison with a reactive approach, the performance gain compared to a proactive approach increases with the amount of leased capacity. Again, this can be explained by the increasing performance of the LRU replacement strategy and the decreasing stability of the content popularity: the reactive part of the hybrid cache covers the decreasing performance of

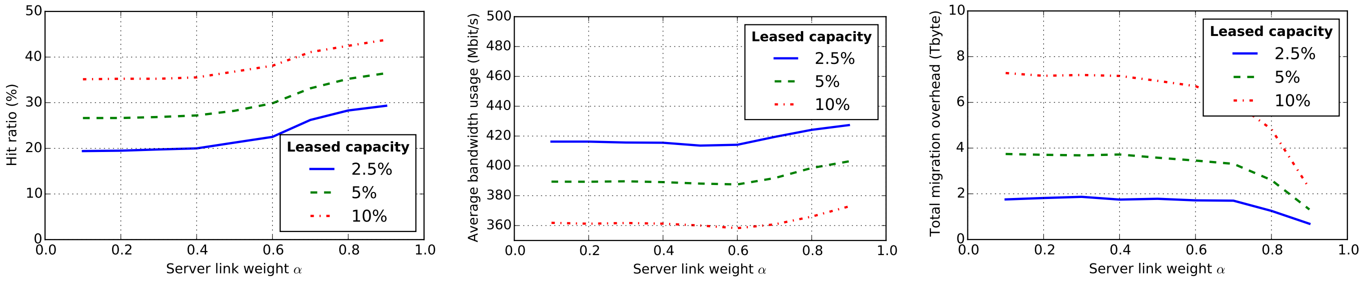


(a) Influence of the number of content providers on the hit ratio.

(b) Influence of the number of content providers on the average bandwidth usage.

(c) Influence of the number of content providers on the average hop count.

Figure 15: Influence of the number of content providers on the performance of the hybrid caching approach.



(a) Influence of the server link weight α on the hit ratio.

(b) Influence of the server link weight α on the average bandwidth usage.

(c) Influence of the server link weight α on the migration overhead.

Figure 16: Influence of the server link weight α on the performance of the hybrid caching approach.

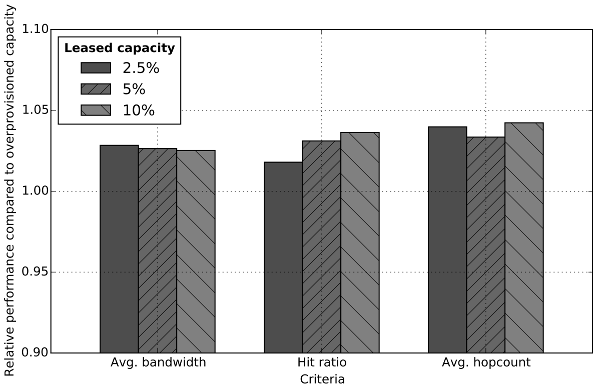


Figure 17: Relative performance of the proposed approach in a scenario with limited capacity compared to the scenario with overprovisioned capacity.

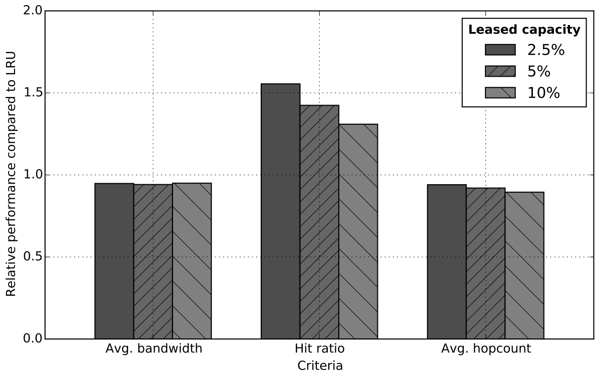


Figure 18: Relative performance of the proposed approach compared to a purely reactive approach when no blockbuster knowledge is available.

the proactive placement strategy in these scenarios. The average performance increases amount to 7.36%, 18.78% and 5.63% in terms of average bandwidth usage, hit ratio and average hop count, respectively, with 39.19% less migration overhead.

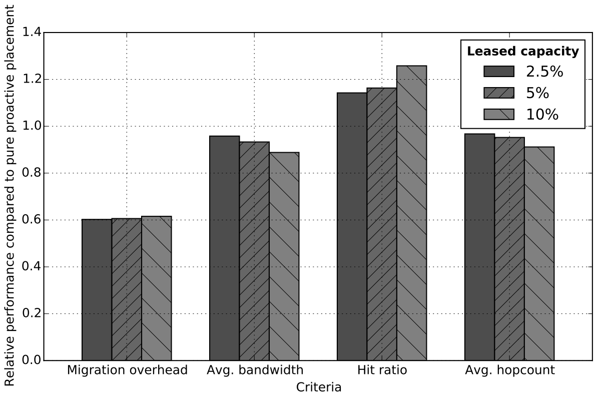


Figure 19: Relative performance of the proposed approach compared to a purely proactive approach when no blockbuster knowledge is available.

When information about 5 blockbuster movies is available on a daily basis, the same trends can be seen (graphs omitted due to space limitations). Compared to a purely reactive approach, the average performance increase in terms of average bandwidth usage, hit ratio and average hop count is equal to 9.26%, 44.34% and 8.46%, respectively. Compared to a purely proactive approach, relative performance increases of 5.89%, 10.58% and 3.94% are achieved in terms of average bandwidth usage, hit ratio and average hop count, respectively, with 39.63% less migration overhead.

7. Conclusions

In this paper, we presented a hybrid cache management approach for ISP networks in a scenario where multiple content providers lease caching capacity. In the proposed approach, a part of the leased capacity is uniformly allocated across the network and equipped with a reactive cache replacement strategy to handle inherent inaccuracies when predicting the content popularity. The allocation of the remaining caching capacity is periodically reconfigured based on predictions of the content popularity and the geographical distribution of requests.

To characterize the performance of the proposed approach in a realistic scenario, a request trace of the VoD service of a leading European telecom operator has been applied. While general characteristics of this trace, such as its popularity curve and diurnal pattern, are comparable to generally accepted popularity models, it contains real-life information about the geographical distribution of content popularity and underlying relationships between multiple content items, which is often nonexistent in synthetic models.

Thorough evaluations have shown that a hybrid cache management approach can combine the benefits of both a reactive approach and the purely proactive management approach proposed in previous work (Claeys et al., 2014a,b), outperforming both approaches in terms of multiple performance indicators for a wide range of reactive ratios. After optimizing the reactive ratio over multiple scenarios for the evaluated VoD use-case and topology, the hybrid cache management approach can increase the hit ratio by 18.78% and 42.96% on average, compared to the purely proactive and the purely reactive approach using the LRU cache replacement strategy, respectively. In terms of average bandwidth usage, the reduction amounts to 7.36% and 5.35%, respectively. Furthermore, content is shown to be stored closer to the end users, indicated by the average hop count being reduced with 5.63% and 8.15%, respectively. Compared to the purely proactive approach, the hybrid caching approach introduces 39.19% less migration overhead.

Finally, the potential benefits of adaptively changing the cache division ratio over time in order to react to changes in the prediction accuracy have been investigated. It was shown that the theoretical performance gain with respect to a static reactive ratio in terms of hit ratio is negligible compared to the required level of added complexity.

In future work, the influence of the underlying network topology on the performance of the hybrid cache management approach will be investigated. Furthermore, business models will be identified for the cooperation between the ISPs and the tenants which lease caching capacity.

Acknowledgments

Maxim Claeys is funded by grant of the Agency of Innovation by Science and Technology in Flanders (IWT). This work was partly funded by Flamingo, a Network of Excellence project (318488) supported by the European Commission under its Seventh Framework Programme.

References

Applegate, D., Archer, A., Gopalakrishnan, V., Lee, S., Ramakrishnan, K. K., nov 2010. Optimal content placement for a large-scale VoD system. In: Proceedings of the International Conference on Emerging Networking Experiments and Technologies (CoNEXT). ACM Press, New York, New York, USA, pp. 1–12.

Asur, S., Huberman, B. A., 2010. Predicting the future with social media. In: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. Vol. abs/1003.5. pp. 492–499.

Baev, I., Rajaraman, R., 2008. Approximation algorithms for data placement problems. SIAM Journal on Computing, 1–18.

Bekta, T., Cordeau, J.-F., Erkut, E., Laporte, G., dec 2008. Exact algorithms for the joint object placement and request routing problem in content distribution networks. Computers & Operations Research 35 (12), 3860–3884.

Borst, S., Gupta, V., Walid, A., 2010. Distributed caching algorithms for content distribution networks. In: Proceedings of the IEEE International Conference on Computer Communications (INFOCOM). pp. 1–9.

Cho, K., Jung, H., Lee, M., Ko, D., Kwon, T., Choi, Y., 2011. How can an ISP merge with a CDN? IEEE Communications Magazine 49 (October), 156–162.

Cisco, 2015. Cisco Visual Networking Index: Global IP traffic growth, 2014-2018. Tech. rep., Cisco.

Claeys, M., Tuncer, D., Famaey, J., Charalambides, M., Latré, S., De Turck, F., Pavlou, G., 2014a. Towards multi-tenant cache management for ISP networks. In: Proceedings of the European Conference on Networks and Communications (EuCNC). pp. 1–5.

Claeys, M., Tuncer, D., Famaey, J., Charalambides, M., Latré, S., Pavlou, G., De Turck, F., 2014b. Proactive multi-tenant cache management for virtualized ISP networks. In: Proceedings of the International Conference on Network and Service Management (CNSM). pp. 82–90.

Dai, J., Hu, Z., Li, B., Liu, J., Li, B., mar 2012. Collaborative hierarchical caching with dynamic request routing for massive content distribution. In: Proceedings of the IEEE International Conference on Computer Communications (INFOCOM). IEEE, pp. 2444–2452.

Frank, B., Poese, I., Smaragdakis, G., Uhlig, S., Feldmann, A., feb 2012. Content-aware traffic engineering. ACM SIGMETRICS Performance Evaluation Review.

Garg, N., Konemann, J., 2007. Faster and simpler algorithms for multicommodity flow and other fractional packing problems. SIAM Journal on Computing 37 (2), 630–652.

Jiang, W., Zhang-Shen, R., Rexford, J., Chiang, M., 2009. Cooperative content distribution and traffic engineering in an ISP network. In: Proceedings of the ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS). pp. 239–250.

Kamiyama, N., Mori, T., Kawahara, R., Harada, S., Hasegawa, H., apr 2009. ISP-operated CDN. In: Proceedings of the IEEE International Conference on Computer Communications (INFOCOM) Workshops. IEEE, pp. 49–54.

Koren, Y., 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In: Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD). pp. 426–434.

Korupolu, M. R., Dahlin, M., 2002. Coordinated placement and replacement for large-scale distributed caches. IEEE Transactions on Knowledge and Data Engineering (TKDE) 14 (6), 1317–1329.

Laoutaris, N., Telelis, O., Zissimopoulos, V., Stavrakakis, I., 2006. Distributed selfish replication. IEEE Transactions on Parallel and Distributed Systems (TPDS) 17 (12), 1401–1413.

Laoutaris, N., Zissimopoulos, V., Stavrakakis, I., 2004. Joint object placement and node dimensioning for Internet content distribution. Information Processing Letters 89 (6), 273–279.

Laoutaris, N., Zissimopoulos, V., Stavrakakis, I., 2005. On the optimization of storage capacity allocation for content distribution. Computer Networks 47 (3), 409–428.

Mukerjee, M. K., Naylor, D., Jiang, J., Han, D., Seshan, S., Zhang, H., aug 2015. Practical, real-time centralized control for CDN-based live video delivery. ACM SIGCOMM Computer Communication Review 45 (5), 311–324.

Sharma, A., 2013. Distributing content simplifies ISP traffic engineering. Proceedings of the ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS), 229–242.

Sourlas, V., Flegkas, P., Gkatzikis, L., Tassiulas, L., 2012. Autonomic cache management in information-centric networks. In: Proceedings of the IEEE Network Operations and Management Symposium (NOMS). pp. 121–129.

Tuncer, D., Charalambides, M., Landa, R., Pavlou, G., 2013. More control over network resources: an ISP caching perspective. In: Proceedings of the International Conference on Network and Service Management (CNSM). pp. 26–33.

Valancius, V., Ravi, B., Feamster, N., Snoeren, A. C., 2013. Quantifying the benefits of joint content and network routing. In: Proceedings of the ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS). pp. 243–254.

Wauters, T., Coppens, J., De Turck, F., Dhoedt, B., Demeester, P., 2006. Replica placement in ring based content delivery networks. Computer Communications 29 (16), 3313–3326.

Wichtlhuber, M., Reinecke, R., Hausheer, D., mar 2015. An SDN-based CDN/ISP collaboration architecture for managing high-volume flows. IEEE Transactions on Network and Service Management (TNSM) 12 (1), 48–60.

Yu, H., Zheng, D., Zhao, B. Y., Zheng, W., 2006. Understanding user behavior in large-scale video-on-demand systems. ACM SIGOPS Operating Systems Review 40, 333.

Zhou, Y., Chen, L., Yang, C., Chiu, D. M., aug 2015. Video popularity dynamics and its implication for replication. IEEE Transactions on Multimedia 17 (8), 1273–1285.