Energy efficiency of on-demand video caching systems and user behavior

Chien Aun Chan,¹* Elaine Wong,¹,² Ampalavanapillai Nirmalathas,¹,²
André F. Gygax,³* and Christopher Leckie¹,⁴

¹Centre for Energy-EfficientTelecommunications (CEET), The University of Melbourne, Victoria, 3010, Australia
²Department of Electrical and Electronic Engineering, The University of Melbourne, Victoria, 3010, Australia
³Department of Computer Science and Software Engineering, The University of Melbourne, Victoria, 3010, Australia
⁴Department of Computer Science and Software Engineering, The University of Melbourne, Victoria, 3010, Australia
*chienn@unimelb.edu.au

Abstract: Energy-efficient video distribution systems have become an important tool to deal with the rapid growth in Internet video traffic and to maintain the environmental sustainability of the Internet. Due to the limitations in terms of energy-efficiency of the conventional server centric method for delivering video services to the end users, storing video contents closer to the end users could potentially achieve significant improvements in energy-efficiency. Because of dissimilarities in user behavior and limited cache sizes, caching systems should be designed according to the behavior of user communities. In this paper, several energy consumption models are presented to evaluate the energy savings of single-level caching and multi-level caching systems that support varying levels of similarity in user behavior. The results show that single level caching systems can achieve high energy savings for communities with high similarity in user behavior. In contrast, when user behavior is dissimilar, multi-level caching systems should be used to increase the energy efficiency.

©2011 Optical Society of America

OCIS codes: (060.4250) Networks; (060.4256) Networks, network optimization.

References and links


#155711 - $15.00 USD Received 30 Sep 2011; accepted 1 Nov 2011; published 18 Nov 2011
(C) 2011 OSA 12 December 2011 / Vol. 19, No. 26 / OPTICS EXPRESS B260
1. Introduction

Energy efficiency has become one of the main concerns in the information and communication technologies (ICT) sector due to the rapid growth in Internet traffic [1]. The ICT sector faces the challenge of offsetting ever higher network traffic, mostly driven by the rapid growth in video traffic, with ever more energy-efficient equipment [2, 3]. Based on two recent reports from comScore, the total number of people who watched video online on an average day in December 2010 has increased by approximately 30% in the U.S. and approximately 20% across European countries, when compared to the previous year [4, 5]. Furthermore, the latest Cisco Visual Networking Index (VNI) forecast reports also indicated an identical issue whereby the sum of all forms of video traffic (TV, video on demand, Internet and peer-to-peer) will continue to be approximately 90 percent of global consumer traffic by 2015 [3]. In addition, the demand for high-definition (HD) video-on-demand (VoD) is predicted to surpass standard definition (SD) by the end of the year 2011 [3]. With the increase in user requests and demand for higher video bit rates, the design of energy-efficient video distribution networks has become a key focus for improving the environmental sustainability of Internet services.

Several network energy consumption models for energy-efficient Internet Protocol television (IPTV) networks have been proposed and studied in [6, 7], concluding that popular contents should be stored on and served by storage servers close to end users. However, the use of the overall mean value of video content popularity as an indicator of which videos to cache in these storage servers may not fully reflect the actual end-user behavior, especially not for communities of diverse backgrounds, interests, and age groups. Therefore, in this paper, we present several energy consumption models to evaluate the energy savings of single-level caching and multi-level caching systems that support varying levels of end-user behavior similarity. Specifically, we show that single-level caching that is close to the users yields substantial energy savings in communities with high user behavior similarity, e.g., a large number of users request the same video content and session length. We also highlight that the less similar user-behavior is, i.e., more users request different video content and download different session lengths, the more effective a multi-level caching system is in achieving significant energy savings.

Multi-level caching, also known as hierarchical caching, was originally proposed based on solid-state drive (SSD) caches that were incorporated into a Digital Subscriber Line Access Multiplexer (DSLAM), Central Office (CO), and Intermediate Office (IO) [8]. Though a thorough analysis of hierarchical cache optimization was made in [8], the dynamic nature of end-user behavior and more importantly, its implication for the efficiency of multi-level caching and the resultant energy consumption of an IPTV network has yet to be investigated. Therefore, the goal of this paper is to explore how the similarity in user behavior affects energy savings of single and multi level caching systems.
2. Review of content distribution architectures and related work

A conventional CDN system such as the one operated by Akamai consists of a large number of content server clusters scattered across multiple Internet service providers’ (ISPs) backbone networks, and are shared among multiple service providers [9]. Alternatively, the service providers could deploy their own CDN network and bring multimedia content closer to their customers, e.g., by using the Velocix digital media delivery platform, to reduce download time [10, 11]. Instead of storing video content in the CDN system, the authors of [6] and [7] have studied the network energy consumption of storing video content at multiple locations in the network, concluding that popular content should be stored and served from storage servers close to users in order to improve the network energy efficiency.

Several caching techniques for multimedia services have been proposed in [12–14], aiming at optimizing the quality of service (QoS) to end users by using proxy servers. Instead of caching content at a single location, the authors of [9] proposed a hierarchical caching system that utilizes cache memory incorporated into network equipment close to the customers, e.g., in a digital subscriber line access multiplexer (DSLAM), central office (CO), or in an intermediate office (IO). The authors modeled the hierarchical caching system as an optimization problem in order to determine the optimal network cost and cache sizes. Although the authors conducted a thorough analysis of hierarchical cache optimization, the dynamic nature of end-user behavior and the associated energy efficiency issue has yet to be investigated.

3. Single-level and multi-level caching systems

This section discusses single-level and multi-level caching systems and the associated energy consumption models. Figure 1 illustrates a simple on-demand IPTV network model. For the conventional CDN system, server clusters are located in the data center, which is connected to the ISP’s backbone core network through data center access, aggregation, and core networks. It is assumed that all videos are stored on the video servers located in the data center. As the distance to the end users decreases, caches can be implemented in the core routers, edge routers and aggregation switches at the CO. A single level caching system consists of only caches at a single level of the network, i.e., caches incorporated into an aggregation switch at the CO. In contrast, a multi-level caching system consists of caches at multiple levels of the network, i.e., caches incorporated into an aggregation switch, edge router, and core router. In a multi-level caching system, the entire requested video session will be divided into several segments and served from different locations depending on the content availability at each caching level. If the total video session length requested is \( M \) (bits), a level 3 caching system will stream a portion of \( B_1 \) (bits) from caches located at the CO (aggregation switch) and the rest of the video portions, \( B_2, B_3 \) and \( B_4 \) will be served from caches in the edge router, the core router, and the video data center, respectively, with a total \( M \) (bits) = \( B_1 + B_2 + B_3 + B_4 \). It should be noted that for a single-level caching system, \( M = B_1 + B_4 \) (video
streaming from the CO and video data center), and for a CDN system, \( M = B_4 \) (video streaming from the data center only).

The energy consumption model of receiving a portion of \( B_1 \) bits of a video stream from caches incorporated into an aggregation switch (level 1) is shown below:

\[
E_1 = \left[ \frac{P_{ONT}}{A} + \frac{P_{OLT}}{AN_{TU}} + 6\left( \frac{P_{Eth}}{C_{Eth}} + \frac{P_{ssd}}{C_{ssd}} \right) \right] B_1,
\]

where \( P_{ONT}, P_{OLT}, P_{Eth}, \) and \( P_{ssd} \) represent the power consumption (W) of the optical network terminal (ONT), the optical line terminal (OLT), the Ethernet switch, and the SSD cache, respectively. \( C_{Eth} \) and \( C_{ssd} \) indicate the capacity (bps) of an Ethernet switch and SSD cache, respectively, and \( A \) denotes the mean access rate per user (75 Mbps) and \( N_{TU} \) represents the number of users per OLT (32 users). If a portion of the video, \( B_2 \), is streamed from caches incorporated into an edge router (level 2), the corresponding energy consumption, \( E_2 \) is:

\[
E_2 = \left[ \frac{P_{ONT}}{A} + \frac{P_{OLT}}{AN_{TU}} + 6\left( \frac{P_{Eth}}{C_{Eth}} + \frac{P_{Er}}{C_{Er}} + \frac{P_{ssd}}{C_{ssd}} \right) \right] B_2,
\]

where \( P_{Er} \) and \( C_{Er} \) represent the power consumption and capacity of the edge router. Likewise, if a portion of the video, \( B_3 \), is delivered from caches incorporated into a core router (level 3), the corresponding energy consumption, \( E_3 \) is:

\[
E_3 = \left[ \frac{P_{ONT}}{A} + \frac{P_{OLT}}{AN_{TU}} + 6\left( \frac{P_{Eth}}{C_{Eth}} + \frac{P_{Cr}}{C_{Cr}} + \frac{P_{ssd}}{C_{ssd}} \right) \right] B_3,
\]

where \( P_{Cr} \) and \( C_{Cr} \) represent the power consumption and capacity of the core router. Finally, the last portion of the video, \( B_4 \) will be delivered from the data center and the corresponding energy consumption, \( E_4 \) is:

\[
E_4 = \left[ \frac{P_{ONT}}{A} + \frac{P_{OLT}}{AN_{TU}} + 6\left( \frac{P_{Eth}}{C_{Eth}} + \frac{P_{Er}}{C_{Er}} + \frac{P_{WDM}}{C_{WDM}} + \frac{P_{DCE}}{C_{DCE}} + \frac{P_{S}}{C_{S}} \right) \right] B_4,
\]

where \( H \) denotes the number of hops in the core network and \( P_{WDM}, P_{DCE}, \) and \( P_S \) represent the power consumption of wavelength division multiplexed (WDM) components, data center network equipment, and video server, respectively. Similarly, \( C_{WDM}, C_{DCE}, \) and \( C_S \) represent the capacity or bandwidth of wavelength division multiplexed (WDM) components, data center network equipment, and video server, respectively. Since the data center network equipment consists of core switches, aggregation switches, and access switches, \( P_{DCE}/C_{DCE} = (P_{Cr}/C_{Cr} + P_{Ags}/C_{Ags} + P_{As}/C_{As}). \) The first \( P/C \) term is for the data center core switch, the second and third terms are for aggregation and access switches, respectively. A factor of 6 is included into the equations due to the overheads of over-provisioning (factor of 2), cooling (factor of 2), and redundancy (factor of 2). We assume that the delivery of a video stream from the data center to the end users requires a total of 12 hops including the access, aggregation, core, and data center networks. Therefore, the parameter \( H \) is assumed to be 6. In addition, we assume that SSDs are used for caching with \( P_{ssd} = 4 \) W and \( C_{ssd} = 1.5 \) Gbps (approximately 200 MB/s). The above equations adopt the widely used energy per bit (or power-to-capacity) ratio of \( P/C \) [1]. The power consumption and capacity for network equipment used in Eqs. (1), (2), (3), and (4), are summarized in Table 1.
Table 1. Equipment Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPON ONT, ( P_{ONT} ) [1]</td>
<td>10 W</td>
</tr>
<tr>
<td>GPON OLT, ( P_{OLT} ) [1]</td>
<td>19.2 W</td>
</tr>
<tr>
<td>Ethernet aggregation switch, ( P_{Eth/C} ) [1]</td>
<td>8 W/Gbps</td>
</tr>
<tr>
<td>Edge router, ( P_{Er/C} ) [6]</td>
<td>26.3 W/Gbps</td>
</tr>
<tr>
<td>Core router, ( P_{Cr/C} ) [1]</td>
<td>12.6 W/Gbps</td>
</tr>
<tr>
<td>WDM equipment, ( P_{WDM/C} ) [6]</td>
<td>3.5 W/Gbps</td>
</tr>
<tr>
<td>Data center core switch, ( P_{Cs/C} ) [1]</td>
<td>2.6 W/Gbps</td>
</tr>
<tr>
<td>Data center aggregation switch, ( P_{Ags/C} ) [1]</td>
<td>1.7 W/Gbps</td>
</tr>
<tr>
<td>Data center access switch, ( P_{As/C} ) [1]</td>
<td>0.38 W/Gbps</td>
</tr>
<tr>
<td>Data center server &amp; storage, ( P_{S/C} ) [15]</td>
<td>40 W/Gbps</td>
</tr>
</tbody>
</table>

4. Modeling user behavior

In this section, we discuss the modeling of user behavior in terms of the daily user requests, video popularity and session length. We model 7 different cases for video title selection and 7 different cases for the requested session length. This setup generates a combination of 49 sub-networks that represent 49 communities with different similarities in user behavior. Assuming that each sub-network consists of 20,480 users, the total simulation accounts for the behavior of 1 million users.

4.1 Total number of daily requests

According to [16] and [17], the number of video requests generally reaches a small peak in the afternoon and a daily maximum in the evening. Therefore, we model the daily access profile for on-demand IPTV services in Fig. 2. The total number of requests is low during the early morning (from 1AM to 9AM). It then reaches the first peak at around 2PM. From 3PM to 6PM, the number of requests decreases slightly before it rises to the daily peak at around 9PM.

![Fig. 2. Daily user requests.](image)

4.2 Video content popularity

According to [16–18], the distribution of video content can be modeled using a Zipf-like distribution. The request probability of video \( k \) is \( P(k) = C/k^\alpha \), and \( C \) is given as:

\[
C = \frac{1}{\sum_{i=1}^{N} 1/l^\alpha}
\]

where \( N \) is the total number of video titles, \( l \) is the index of a video title in the list of \( N \) videos sorted in order of decreasing popularity, and \( \alpha \) is the skew factor. By setting \( \alpha \) close to 0, the
resultant distribution is uniform. In this case, every video title has the same probability of being requested, and this corresponds to communities that have low similarity in user behavior when selecting video content. In contrast, setting $\alpha$ to a value higher than 1 results in a distribution that is highly skewed. This results in high popularity for a few video titles. Therefore, a higher value of $\alpha$ is used to model communities that have high similarity in user behavior. As a result, we model 7 different cases for video popularity using a Zipf-like distribution with $\alpha$ ranging from 0.01 (less similar) to 1.2 (most similar), which gives an average value of approximately 0.6.

### 4.3 Session length and video start time

Previous studies [19, 20] indicated that, depending on the video category, several distribution functions can be used to model the video session lengths, i.e., lognormal, exponential, gamma-lognormal, and double exponential. We assume that the lognormal distribution is used to model 7 different cases of session lengths. The probability density function (PDF) of a log-normal distribution is given as:

$$f_x (x | \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}.$$  \hspace{1cm} (6)

By setting $\mu$ close to zero, the log-normal distribution will be highly skewed. This is used to model communities that have high similarity in user behavior. In contrast, by setting $\mu$ to a higher value (e.g., 6), the log-normal distribution results in a uniform distribution. This is used to model communities that have low similarity in user behavior. Therefore, to model 7 different cases of session length, the log-normal parameters are set to be: $\sigma = 1.63$ and $\mu$ ranging from 0.1 (most similar) to 6.1 (less similar). In addition, we assume that approximately 80% of viewer requests correspond to watching a video from the beginning of the video and the remaining viewer requests begin watching a video at some point in between following a Pareto distribution [21].

The above setup generates a combination of 49 (7 x 7) sub-networks, which represent 49 communities with different similarities in user behavior in terms of video content popularity and session length.

### 5. Results and discussion

Discrete time event simulations were carried out to simulate the user behavior of 1 million users for two days. The first simulation day is used to calibrate the caches. Each of the 1,000 video titles is divided into 1-minute intervals. Starting from the caches that are closest to the end users, the single-level caches (caches in the CO) will store the most popular video content based on the request rate of the first simulation day, on a per-minute basis. After fully utilizing the first level caches, the second level caches (caches incorporated into the edge routers) will continue to store the popular video content that has not been stored. The same procedure is performed for the third level caches (caches incorporated into the core routers) after the second level caches are full. Then, the user requests of the second simulation day are used to evaluate the performance of the single-level and multi-level caching systems. By using Eqs. (1), (2), (3), and (4), we compute the energy consumption of conventional CDN, single-level caching, and multi-level caching systems for 49 sub-networks that have different similarities in user behavior.

#### 5.1 Single-level caching

Single-level caching utilizes the caches that are incorporated into the aggregation switches at the CO. Due to limited cache sizes, a portion of the requested video will be streamed from the caches while the rest of the video will be streamed from the video server. Figure 3 shows the daily energy consumption of 49 communities that have different similarities in terms of video title selection and requested session length with cache size of 512 GB. The seven cases of video title selection and requested session length similarity are based on the seven scenarios
modeled in Section 4. The figure shows that the network energy consumption is directly affected by the user behavior of the community. For the community that has the most similar user behavior in terms of video title selection and session length, the associated energy consumption is approximately 23 kWh/day. In contrast, for the similarity in user behavior decreases, the community that has the lowest similarity in user behavior, the daily network energy consumption increases to approximately 736 kWh/day. If larger cache size than 512 GB is used, the daily energy consumption will be reduced but follows the shape of the 3-dimension plot shown in Fig. 3.

![Figure 3: Daily energy consumption of 49 communities using single-level caching with cache size of 512 GB.](image)

Comparing the network energy consumption of a conventional CDN to a single-level caching system, Fig. 4 plots the potential energy savings (in %) of a single-level caching system with variable cache sizes of 256 gigabytes (GB), 512 GB, 1 terabyte (TB) and 2 TB. A normalized similarity value of ‘0’ indicates that the residents within a community have very low similarity in terms of video title selection and requested session length. On the other hand, a normalized similarity value of ‘1’ means that the residents are most likely to request the same video content over a similar session length. Figure 4 shows that as the caching size increases from 256 GB to 2 TB, the energy savings improve significantly from 6% to 41% for communities that have less similarity in user behavior. However, for communities with high user behavior similarity, significant energy savings of 70% can be easily achieved using small cache sizes (i.e. 256 GB) and as cache size increases, the energy savings improve accordingly (up to 82% for a 2 TB cache).

![Figure 4: Energy savings of single-level caching system with variable cache sizes of 256 GB, 512 GB, 1TB, and 2TB.](image)
5.2 Multi-level caching

In order to improve the energy efficiency of a video caching system for communities with low similarity in user behavior, multi-level caching systems can be used. Figure 5 shows the energy savings of using a two-level caching system (caches in the CO and edge routers) with variable cache sizes of 256 gigabytes (GB), 512 GB, 1 terabyte (TB) and 2 TB. As depicted in the figure, as the caching size increases from 256 GB to 2 TB, the energy savings improve from 11% to 63% for communities that have less similarity in user behavior. However, for communities with high user behavior similarity, energy savings of 74% can be achieved using small cache sizes (i.e. 256 GB) and as the cache size increases, the energy savings improve accordingly (up to 83% for a 2 TB cache).

![Fig. 5. Energy savings of two-level caching system with variable cache sizes of 256 GB, 512 GB, 1TB, and 2TB.](image1)

Figure 6 shows the energy savings of using a three-level caching system (caches in the CO, edge routers, and core routers) with variable cache sizes of 256 GB, 512 GB, 1 TB and 2 TB. For communities that have less similarity in user behavior, energy savings can be improved from 15% to 73% by increasing the cache size from 256 GB to 2 TB. Likewise, for communities with high user behavior similarity, energy savings can be improved from 76% to 83.5% by increasing the cache size from 256 GB to 2 TB.

![Fig. 6. Energy savings of three-level caching system with variable cache sizes of 256 GB, 512 GB, 1TB, and 2TB.](image2)

Comparing the performance of single-level and multi-level caching systems, Fig. 7 summarizes the energy savings of single-level, two-level and three-level caching systems for a cache size of 2 TB. The results clearly indicate that for communities with low user behavior
similarity, energy savings can be improved significantly from 41% (single-level) to 63% and 73% by using two and three-level caching, respectively. However, for communities with high user behavior similarity, the improvement of 1.5% (82% to 83.5%) in energy savings is minimal with an additional caching level. Further, no significant difference is observed between two-level caching and three-level caching. Therefore, deploying multi-level caching systems to support communities with high user behavior similarity is potentially redundant and may lead to unnecessary use of additional resources. On the other hand, the use of multi-level caching systems is critically important for communities that have low user-behavior similarity in order to achieve significant energy savings.

Fig. 7. Energy savings of single-level and multi-level caching compared to CDN system with cache size of 2 TB.

5.3 Discussion

Delivering video content from the data center requires substantial amount of network transport energy due to large number of network hops. Energy savings can be achieved by delivering video content from caches located close to the end users. Sub-sections 5.1 and 5.2 have shown that by using caches incorporated into network equipment close to the end users, energy savings compared to a conventional CDN system can be achieved. However, to further improve the energy-efficiency of the caching systems, a thorough analysis of the implications of user behavior to the design of video caching system is critical.

For communities that have less similarity in user behavior, the end users often request for different video content and session lengths, which means that the distributions of video title selection and requested session length are close to uniform. Therefore, in order to accommodate this scenario where each minute of all video titles is having similar probability to be requested, two solutions can be used to increase the overall network energy savings. The first solution is to increase the caching size to store more video content in the cache as shown in Fig. 4. However, due to limited cache size that can be incorporated into a single-level caching systems, the second option is to deploy a multi-level caching system, which virtually increases the total caching size of the end-to-end video caching system by incorporated caches into multiple levels of the network (as shown in Figs. 5 and 6).

In contrast, for communities that have high similarity in user behavior, the end users often request for the same video content and session lengths, which means that the distributions of video title selection and requested session length are highly skewed. Therefore, limited cache size is adequate to achieve substantial energy savings. As a result, deploying multi-level caching system that virtually increases the total amount of caching size is unnecessary. Also, it should be noted that the caching size requirement is directly dependent on i) the total number of video title, and ii) the length of each video content, which may vary the results shown in this paper.
6. Conclusions

The rapid growth in Internet video traffic has increased the need for energy-efficient video distribution networks. In this paper, we show that the end user behavior can greatly affect the energy consumption of a network. In this context, caching systems can help to reduce energy consumption by optimizing the flow of network traffic. In this paper, the energy consumption models for single-level caching, multi-layer caching, and conventional CDN systems were presented and compared across 49 communities of differing user behavior. Our results showed that caching systems should be designed according to the user behavior of the communities. By way of simulation, we showed that in comparison to a CDN system, a single level caching system facilitates significant energy savings of up to 82% for communities with high similarity in user behavior. Similarly, comparable energy savings of up to 73% can be achieved for communities with low similarity in user behavior through using a multi-level caching system. Therefore, the knowledge of user behavior similarity that is unique to each community and the understanding of its effect are critical in optimizing the energy-efficiency of on-demand IPTV networks.

Acknowledgments

We thank Dan Kilper, Rod Tucker, Kerry Hinton, Rob Ayre, and Arun Vishwanath for valuable discussions about Internet energy efficiency. We gratefully acknowledge financial support by Bell Labs (Alcatel Lucent), the Victorian State Government, the Centre for Energy-Efficient Telecommunications (CEET), and the Institute for a Broadband Enabled Society (IBES).