Cache as a Service: Leveraging SDN to Efficiently and Transparently Support Video-on-Demand on the Last Mile

Panagiotis Georgopoulos*†, Matthew Broadbent†, Bernhard Plattner* and Nicholas Race†
*Communication Systems Group, ETH Zurich, 8092 Zurich, Switzerland
†School of Computing and Communications, Lancaster University, Lancaster, LA1 4WA, UK
panos@tik.ee.ethz.ch, m.broadbent@lancaster.ac.uk, plattner@tik.ee.ethz.ch, n.race@lancaster.ac.uk

Abstract—**High quality online video streaming, both live and on-demand, has become an essential part of consumers' every-day lives. The popularity of video streaming has placed a heavy burden on the network infrastructure that now has to transfer an enormous amount of data very quickly to the end-user. To further exacerbate the situation, the Video-on-Demand (VoD) distribution paradigm uses a unicast independent flow for each user request. This results in multiple duplicate flows carrying the same video assets many times end-to-end. We present OpenCache: a highly configurable, efficient and transparent in-network caching service that aims to improve the VoD distribution efficiency by caching video assets as close to the end-user as possible. OpenCache leverages Software Defined Networking to benefit last mile environments by improving network utilisation and increasing the Quality of Experience for the end-user. Our evaluation on a pan-European OpenFlow testbed uses adaptive video streaming and demonstrates that with the use of OpenCache, the external link utilisation is reduced by 100%. Furthermore the streaming application receives better quality video and observes higher throughput, lower latency and shorter start up and buffering times.**

Keywords—Video-on-Demand (VoD), Caching, Quality of Experience (QoE), Software Defined Networking (SDN), OpenFlow, OpenCache

I. Introduction

Recent years have seen a huge growth in the popularity of video streaming, for both live and on-demand services. In 2012, Internet video traffic accounted globally for 57% of all consumer Internet traffic, and is predicted to increase even more, to 69% by 2017 [1]. Correspondingly, the popularity of Video-on-Demand (VoD) traffic continues to increase, with the volume of VoD traffic predicted to reach the equivalent of 6 billion DVDs per month by 2017 [2]. At the same time, High Definition (HD) video traffic has already surpassed Standard Definition (SD) traffic, making HD the de facto video quality level consumed by users [2]. There is no doubt that high quality online video streaming has become an essential part of consumers' every-day life.

With a VoD service, individuals can retrieve previously recorded content at a time after it was initially broadcast or made available. With the increasing growth of VoD and the popularity of HD content, an alarming challenge to the underlying network infrastructure is becoming apparent. The network now has to transfer an enormous amount of data to the end-user, and do so as quickly as possible. At the same time, the available content continues to improve in terms of resolution and overall video quality, and this trend will continue as we move from HD through to Ultra HD and 3D video. These changes in video quality require delivery throughput in the order of tens of Mbps for just one stream, and place additional burden on the underlying network infrastructure for supporting their distribution.

Currently, VoD requests are handled individually, leading to *an independent flow* in the distribution network serving *each user's request*. Using such a unicast content delivery paradigm naively ignores that much of the content is identical to transmissions minutes, hours or days earlier. Hence, a very large amount of identical media objects, in the order of gigabytes for each HD film, is delivered on the same network segments repeatedly. In order to efficiently support such VoD streaming, the end-to-end capacity of the network must continuously match the increasing number of Internet video users and the growing popularity of HD content. Mechanisms are thus sought to improve the VoD distribution efficiency.

In this paper we introduce OpenCache: a transparent, flexible and highly configurable in-network caching service for VoD streaming. OpenCache uses Software

978-1-4799-3572-7/14/$31.00 ©2014 IEEE

Defined Networking (SDN) to provide *cache as a service* for media content in an efficient and transparent fashion. This should directly benefit last mile environments. By leveraging SDN, and OpenFlow in particular [3], we provide a control plane that orchestrates the caching and distribution functionalities, and transparently pushes the content as close to the user as possible without requiring any changes to the delivery methods or the end-hosts.

Our approach, building an SDN based in-network caching service, has three important contributions. First, it improves network utilisation and minimises the external link usage on the last mile that is often costly. Second, OpenCache reduces the distribution load from the VoD content provider and all the transient networks along the path of the VoD server to the end-user. Third, by transparently caching the content closer to the user, OpenCache minimises the distance between the VoD streaming server and the user. This provides great improvements to the Quality of Experience (QoE) of the end-user, as the streaming application observes higher throughput, less latency and smaller start up and buffering times; key QoE differentiators [4], [5], [6].

The remainder of the paper is organised as follows. Section II provides the background of this work, whilst related work is presented in Section III. Section IV introduces the main entities and functionality of OpenCache, whereas Section V describes the key benefits achieved with SDN. Evaluation is shown in Section VI and finally, Section VII concludes the paper.

II. Background

A. *Motivation & Problem Statement*

To achieve high quality VoD streaming, a potential solution should be able to address the following two primary requirements :

1) Provide high throughput end-to-end: High quality video streaming demands quick and reliable transmission of high bitrate encoded content end-to-end. It is often the case that the intermediate networks that have to transfer the media content quickly, become the bottleneck for high quality video streaming. It is not enough to ensure adequate origin server capacity, but adequate network bandwidth must be available in all the intermediate networks between the content server and the end-user [7], [8], [9]. With the Internet being highly fragmented, namely, the largest network worldwide accounting for only 5% of user traffic and needing over 650 networks to reach 90% of access traffic [7], [10], this is a stark problem. This fragmentation means that content that is centrally hosted must travel over multiple networks to reach end-users. Therefore, the burden falls on the intermediate networks to ensure adequate capacity is available to achieve the necessary end-to-end throughput for high quality streaming.

2) Minimise distance between VoD server and user: Large geographical distance between the content server and the end-user presents the potential for higher latency and packet loss in today's best-effort Internet. High latency and packet loss are particularly important as, when present, the user notices larger start up and buffering times. In addition, frame dropping and frame freezing are observed. Ultimately, these eventually lead to lower QoE [4], [5], [6], [8], [9], [11]. To minimise packet loss and benefit from reliable transmission, major VoD content providers (e.g. Netflix, Amazon's LoveFilm, YouTube etc.) use TCP to stream VoD content [12], [10], [8], [9], [13]. However, TCP's performance is highly affected by latency and packet loss (noticeably present when the VoD server and client are far away from each other). This is because TCP's throughput is inversely related to network latency or RTT [7], [8]. Therefore, from both networking and QoE perspectives, the distance between the server and the end-user can become a significant bottleneck in maintaining high quality video streaming.

A potential solution should address the aforementioned challenges and ensure that the media content resides as close to the user as possible. Such an approach would ensure lower latency and higher throughput end-to-end, eventually leading to higher video quality and lower start up and buffering times [6], [4], [5].

B. *Software Defined Networking*

Software Defined Networking (SDN) is a new, very promising, networking approach that facilitates the decoupling of the control plane in a network (i.e. the decision making entity) from the data plane (i.e. the underlying forwarding system). OpenFlow [3], currently the prominent SDN protocol, defines the communication between the Layer 2 switches and the controller of the network in an open and vendor-agnostic manner. OpenFlow allows experimenters, protocol developers and network administrators to exploit the true capabilities of a network in an easily deployable and flexible manner. With the centralised network perspective that OpenFlow provides through its controller, an administrator has an overarching view of the current network status and has the ability to programmatically introduce new network-wide functionality without having to interact with each individual network or user device. OpenCache, our in-network caching service, uses OpenFlow to dynamically cache and distribute media content within a network in a highly efficient and transparent manner.

III. Related Work

Multicast is a technology that can deliver the same media assets to multiple users simultaneously by reducing the delivery throughput on the origin server to one stream. Setting aside the multiple real-world deployment problems that multicast entails [14], multicast is, by design, an efficient solution for live video streaming, where, all users' requests are for the same content at the same time. However, with VoD, this is not the case, as requests can occur any time after the content becomes available. Related work that has used multicast to improve the delivery efficiency of VoD often involves unnecessary complexity, such as merging streams by speeding up or slowing down the clients' viewing rate [15], [16], holding a portion of the clients' bandwidth in reserve [17], or requiring from the clients the receipt of two or more video streams simultaneously, each at the playback rate [18], [19], [17]. Also, these solutions are not transparent and require client or server modifications.

The efficiency of a **Peer-to-Peer** (P2P) based networking solution for video streaming depends heavily upon the participation of users and their willingness to share their limited storage and network resources [20]. P2P pushes the content closer to end-users and can deliver a live video stream to multiple users simultaneously. This is possible because peers can sustain the short-term retention of live video using their own resources. However, P2P is much less effective for VoD distribution, as the time between requests for identical content may be in the order of hours, days or even months. In addition, P2P peers may join and leave the service at will, making VoD streaming quality assurance very challenging. Furthermore, the distributed nature of P2P brings a lack of central control, particularly for authentication, authorisation, accounting and security. This prohibits network administrators and content providers from making informed decisions and improving the service that they provide. In addition, it prevents these parties from using intelligent caching and distribution techniques [21].

Alternatively, traditional in-network **cache and proxy** approaches aim to provide additional network and storage support by focusing on delivering the content to users locally. For example, [22] demonstrates the significant benefits of caching YouTube content, where even a very basic caching policy (i.e. a static cache with long-term popular videos) can approximately achieve a 51% cache-hit ratio. Similar benefits are demonstrated in [23], where a simple two hour expiration caching policy yields an aggregated request and byte hit rate of 24% using cache storage of a size less than 2% of the overall data transferred.

Historically though, the most common use of caching and proxying servers is to serve static web content. Thus, existing solutions (e.g. Squid [24]) are not usually optimised for the high storage, bandwidth and the very demanding application requirements of video delivery (e.g. skipping to a certain part of a video stream). Some popular caching servers are too complex to customise and configure, and require constant attention and tuning from network administrators. To the other extreme, some commercial caching solutions provide little flexibility, customisation and configurability as to the content that should be cached and their caching policies. These solutions are essentially black boxes in the network, running on dedicated hardware and requiring third-party support. Such solutions typically leave network administrators with minimal control and resource monitoring of devices located within their own network.

Another mechanism to improve the efficiency of VoD delivery is to use a dedicated **Content Delivery Network** (CDN). CDNs deploy a large number of caching servers worldwide, in order to push content to the edges of the Internet [7], [21]. CDNs are typically deployed in order to achieve goals similar to those noted in Section II-A. From a content provider's perspective, CDNs are an efficient distribution and cost effective solution. However, from a consumer ISP's perspective, CDNs do not reduce the bandwidth utilisation on last mile connections, as multiple requests for the same video content will create an equal amount of flows serving the same amount of content to end-users. Even in scenarios where a dedicated CDN cache can be deployed within an ISP's network [25], this cache is specific to a particular service and has strict hardware, software and networking requirements (e.g. video traffic higher than a threshold), that deems it unsuitable for medium-scale ISPs or last mile deployments. In addition, despite the fact that CDNs deploy their servers worldwide, it is unrealistic to expect them to deploy in all ISP networks or last mile environments.

Even CDNs themselves have recognised their inability to truly reach out to last mile environments. For example, in order to address this problem and reduce maintenance and administration cost, Akamai introduced a hybrid CDN-P2P based solution, that complements their service by pushing content closer to end-users [21]. However, such an approach has the drawbacks of P2P networking mentioned earlier, namely, requiring user involvement to download and install software, and consuming the limited storage and uploading resources of end-users. It is without doubt that a more flexible, configurable and transparent in-network caching service, located closer to the user, would complement CDNs and truly benefit last mile environments.

IV. OpenCache

OpenCache is an OpenFlow-assisted in-network caching service that provides efficient, transparent and highly configurable caching and distribution of VoD content in the last mile.

OpenCache offers a powerful interface that provides *cache as a service*. This is not intrinsically linked to a particular type of content, or to a specific hardware or software implementation. The control and decision of what content should be cached is passed on to the network administrator of the ISP, who now has the ability to optimise his network's utilisation and external link usage. This can be achieved by enabling in-network caching for specific content via a designated interface. OpenCache exposes this interface through a powerful and flexible JSON-RPC based API, which allows VoD content to be cached closer to the end-user. This placement also increases their QoE when streaming the content. Given appropriate SLAs, OpenCache's interface could also be used by content providers (e.g. CDNs) to declare their content as cacheable on last mile environments, without having to physically deploy and administer their own caching hardware. It is envisaged that OpenCache's interface will eventually be compatible with CDNI [26], a collaboration protocol currently under development by the research community. This functionality can be used by CDNs to define their interconnections, and hence ease interoperability and communication between them and potentially OpenCache.

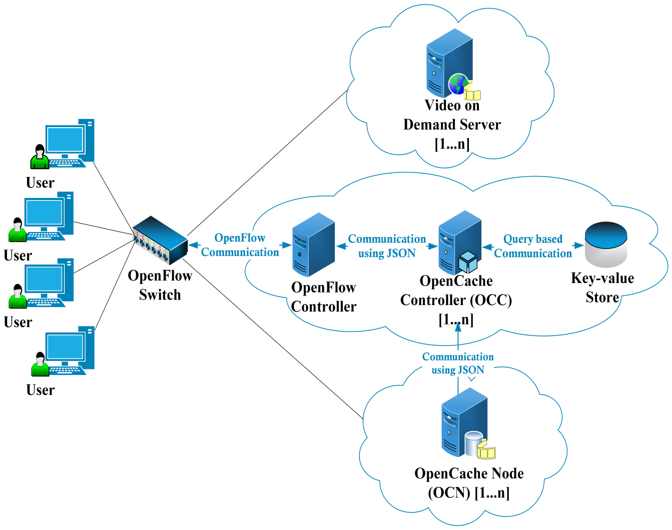


Fig. 1: OpenCache Architecture

Fig. 1 presents the entities of OpenCache when deployed on a production network. On an SDN-enabled, OpenFlow-based network, users are connected to Layer 2 OpenFlow switches. The functionality of these switches is dictated by a network controller using the OpenFlow protocol [3]. The main entity of OpenCache, namely the OpenCache Controller (OCC), orchestrates the VoD caching and distribution functionalities with the aid of a key-value store, that acts as a database. The OCC communicates with the OpenFlow controller of the network via a JSON-RPC interface. A VoD server is the primary source for the video assets and could be located anywhere on the Internet reachable by its IP address. Finally, the OpenCache Nodes (OCNs) are the caches of the service, inherently being deployed in various locations in the network.

With respect to the placement of the OpenCache's entities, the OCC, along with its partner elements (the OpenFlow controller and the key-value store) would ideally be located in the same network as the end-users. A single, widely reachable OCC would be able to coordinate caching amongst a number of OCN instances (i.e. caches). It is important to note that it is not a requirement that OpenFlow be deployed throughout the network; the connecting network hardware could also be entirely non-OpenFlow. The only specific OpenFlow switch requirement is at the last hop, closest to the user. We propose that OCNs are connected directly to OpenFlow switches on which clients are also attached. This deployment would offer the lowest latency and fastest response time, and thus ensure higher QoE for the end-users. However, the use of multiple OCNs at different points in the network (e.g. attached to aggregation switches in an enterprise or University campus network) is also entirely feasible. In fact, a hierarchical approach has the potential to provide further benefits if a greater proportion of requests can be fulfilled without leaving the LAN.

A. The OCC's Functionality and Interfaces

The OpenCache Controller (OCC) is the main orchestrator of the in-network caching functionality that OpenCache provides, and implements the following four main operations :

1) Receives requests for content of interest that should be cached in the network's OCNs (Fig. 2), by exposing a JSON-RPC based API (Table I). Network administrators or content providers invoke the methods provided by this interface, using the appropriate authentication credentials, to declare what content should be cached from this point on in the network. If there is a request for certain content to be cached (using the *start-expr* method) the OCC ensures that this interest is stored in the key-value store and that all the OCNs are initialised and aware of the content that was requested for caching. In addition, the OCC will interact with the network's controller and instruct it to add the matching OpenFlow redirecting rules for the cacheable content in the OpenFlow switches of the network. These flows ensure that all the users' requests for that content are redirected to their closest OCN. If there is a request to stop caching content previously added (using the *stop-expr* method), then the OCC updates the key-value store appropriately and ensures that all the matching content and flows are removed from the networks' OCNs and OpenFlow switches, respectively. Finally, the OCC exposes a *list-expr-all* method over its API that lists to the requester all the content that has been requested for caching.

It is important to emphasise the granularity and flexibility that this interface provides: a regular expression-like syntax can be used to define parameters in the *start-expr* and *stop-expr* methods. The level of granularity is only constrained by the matching capabilities of OpenFlow. This allows administrators to fine tune their requests. For example, a request can be made for a specific video to be cached, or the videos of a whole domain or even a certain type of video from any domain.

2) Implements the caching logic which dictates what content should be cached and to which OCN at each point in time. OpenCache's ability to control the caching logic centrally on the OCC allows the network administrator to program and deploy their desired caching behaviour. This request can be based on the specific parameters that he wants to optimise in his network. For example, an administrator could program the caching logic to minimise the streaming latency of recorded video lectures on a University's network, or to implement the pre-caching of popular content closer to end-users overnight, when the network is underutilised. It is important to note the ease and speed at which the network administrator can actually implement the caching logic with the use of OpenFlow in OpenCache.

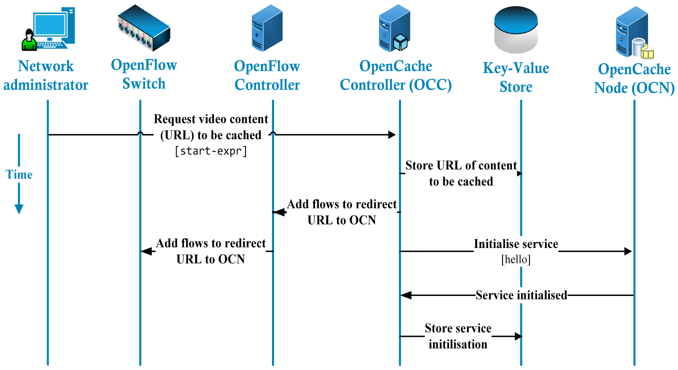

Network administrator
OpenFlow Switch
OpenFlow Controller
OpenCache Controller (OCC)
Key-Value Store
OpenCache Node (OCN)
Time
Request video content (URL) to be cached
[start-expr]
Store URL of content to be cached
Add flows to redirect URL to OCN
Add flows to redirect URL to OCN
Initialise service
[hello]
Service initialised
Store service initilisation

Fig. 2: Declaring Content of Interest as Cacheable

TABLE I: Interface for Declaring Content of Interest

Method	Parameters	Result
start-expr	{ "expr" : <expr> }	<boolean>
stop-expr	{ "expr" : <expr> }	<boolean>
list-expr-all	none	[{"expr" : <expr>, "port" : <port>}, ...]

TABLE II: Interface to Interact with OCNs

Method	Parameters	Result
hello	{ "host" : <host>, "port" : <port> }	<node-id>
keep-alive	{ "node-id" : <node-id> }	<boolean>
goodbye	{ "node-id" : <node-id> }	<boolean>

3) Manage the available OCNs' resources in the network. An important part of resource management is to be able to handle the addition and removal of caches in a network dynamically. For this reason, the OCC exposes another JSON-RPC based API that allows the communication of the OCC with a number of OCNs (Table II). When an OCN is added to the network, it invokes the *hello* method to let the OCC know that it is now available on the network. In turn, the OCC replies with a *node-id* that is assigned to this particular OCN. From that point on, the OCN will periodically send a *keep-alive* message to indicate that it is still in the network and functioning correctly. If the OCC does not receive a *keep-alive* call from an OCN every 15 seconds (a configurable option), then it assumes that the OCN is not reachable, either because of network congestion or because it has been taken offline. Consequently, the OCC will remove the "unreachable" OCN from the list of caching resources that it has at its disposal. Alternatively, an OCN may also leave the network gracefully with the transmission of a *goodbye* message (Table II).

4) Manage and maintain the OpenFlow flows in the network dynamically via a Flow Pusher API that an OpenFlow controller provides (e.g. [27]). The OCC defines dynamically the appropriate flows that should be in the OpenFlow switches, so that each user's request gets redirected to an OCN in his vicinity (Fig. 2). With the management of flows, the OCC also propagates the caching logic to the network (e.g. expire content, perform load balancing or pre-caching). This also ensures that the caching and distribution functionalities remain purely in the network and are fully transparent to end-users.

B. *The OpenCache Node's Functionality*

The OpenCache Node (OCN) is responsible for caching the appropriate video content, and delivering it to users if they request it. When the OCN comes online in a network, it communicates with the OCC (via the interface in Table II) and makes its resources available to it. Subsequently, when the OCN obtains a *node-id* from the OCC, it initialises its operation and awaits users' requests.

When a user makes a video request, and if the content has been declared as cacheable, the request will be received by the closest OCN to the user. This is in contrast to it traversing the external link, which would be the case without OpenCache in place. This is possible due to the pre-populated rules installed in the OpenFlow switches when declaring content of interest (Fig. 2). Following these rules, an OpenFlow switch redirects the user's packets appropriately. When the OCN receives such packets, it examines if it has the requested content already cached. If the particular video is not cached in the OCN (a cache-miss scenario depicted in Fig. 3), the OCN requests the video from the original VoD server. Once the first packet of this flow is received, the OCN will begin forwarding these back to the client. This process is intended to reduce any latency induced by the caching process. The delivery of this content traverses the OpenFlow switch too, and additional rules ensure that the packet received by the client appears to be from the expected source. With the completion of this process, the session has remained transparent to the user and there is no interruption to the service. Once the full flow has been handled in this way, the payload of the delivery is stored by the OCN in order to serve subsequent requests. Furthermore, the OCN informs the OCC of this transaction for resource provision and management purposes.

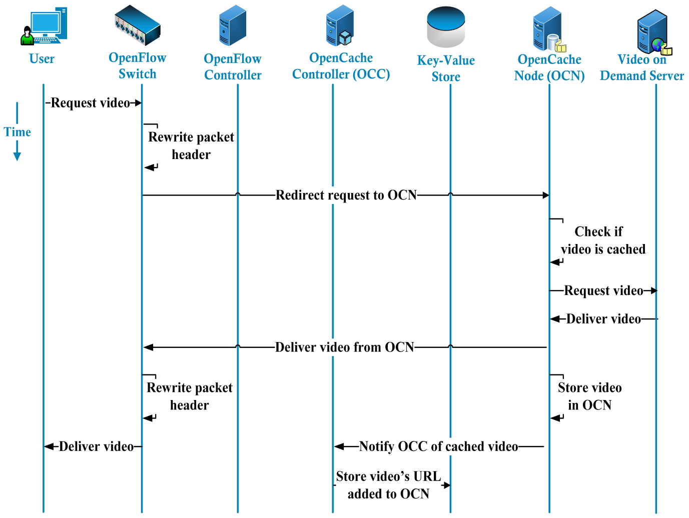


Fig. 3: Cache-miss Scenario

If the content for a particular request is already stored in the OCN (a cache-hit scenario), the OCN delivers that video directly to the user in a transparent fashion. As in a cache-miss scenario, the content always appears to originate from the VoD server the client originally requested it from. However, in a cache-hit scenario, no traffic would have left the user's network into other networks, thus saving external link utilisation and significantly reducing start up and buffering delays. The role of the OCN is such that there is an inherent need to have multiple instances of it distributed in the network to facilitate users' requests and content caching as efficiently as possible.

V. Key SDN Benefits

By using Software Defined Networking (SDN), and OpenFlow in particular, we are afforded a number of key abilities that OpenCache exploits to their fullest extent. Most critically to the operation of OpenCache is the ability to transparently redirect requests for content to a running cache instance, as mentioned previously. More specifically, this is achieved by rewriting the packet header information, and intentionally forwarding it towards a cache or a client. The use of OpenFlow to achieve this functionality allows the rewriting procedure to happen in a distributed fashion on the switches of the network, and hence removes the need to overload a specific server with the burden of doing so. It also ensures full user transparency as the content that reaches the client appears to originate from the origin server rather than the cache. It is important to note that this is all possible without the costly and time-consuming modification of existing delivery techniques (e.g. caching or proxy servers, middleboxes) or end-client devices.

To compliment this, SDN provides both hardware and software abstraction to our in-network caching service. OpenCache can be used on commodity hardware, without the need to perform complex configuration and setup. The only requirement for OpenCache to function is the presence of a single-piece of OpenFlow-capable hardware on the path from client to the server. The ability to perform the necessary operations anywhere in the network where OpenFlow is supported grants OpenCache with even greater flexibility. For example, and as presented in the paper, we can do this close to the user in a last mile environment, where traditionally there has existed no such cache process. However, although not discussed here, the exact same process can be applied in other situations and environments without modifying OpenCache or OpenFlow itself; the same logic is applicable.

Furthermore, utilising OpenFlow also allows us to monitor the networking hardware contained within our topology, and use this feedback in OpenCache itself. This gives OpenCache a perspective of the network not typically afforded to application-layer technologies. This information can be used in conjunction with the programmability that OpenFlow permits to effectively satisfy any caching requirements. This is possible without the need to consider the peculiarities of differing hardware devices in the network. In addition, monitoring information, when used in conjunction with the redirecting action described previously, allows us to load balance requests on-the-fly and in real-time. This level of reactivity has not been previously possible, particularly through the use of a single, unified API.

VI. Evaluation

In order to evaluate the efficacy of OpenCache, we carried out a number of VoD streaming experiments over a large-scale pan-European OpenFlow testbed provided by the OFELIA project [28]. OFELIA is composed of a number of OpenFlow-capable hardware switches and virtualised computing resources located in many different countries across Europe. Each site (or "island") is connected as to produce a large federated experimentation environment. For our experiments, a video client runs on a virtual machine located at the ETH Zurich OpenFlow island in Switzerland, where we also deployed an OpenFlow controller, an OCC and an OCN. Two VoD servers were deployed on virtual machines on two different OFELIA islands; at CREATE-NET in Italy and at i2CAT in Spain. Finally, the three islands are federated together using an OFELIA island in Belgium.

Three main metrics are used to evaluate the effectiveness of OpenCache; start up delay (key QoE metric [4], [5], [6]), external link utilisation and video playback bitrate. For all of metrics, we use the same evaluation environment; a VLC client in Switzerland accesses the same reference video content ("Big Buck Bunny"[1]) hosted at one of the VoD servers located in either Italy or Spain. The content is streamed using an adaptive video streaming technology, namely, DASH (Dynamic Adaptive Streaming over HTTP) [29]. MPEG-DASH facilitates the dynamic adjustment of the streaming bitrate by offering various bitrate encodings of the reference video, fragmented into fixed time chunks. MPEG-DASH aims to improve the overall QoE for the end-users by dynamically matching the bitrate requested to the available bandwidth.

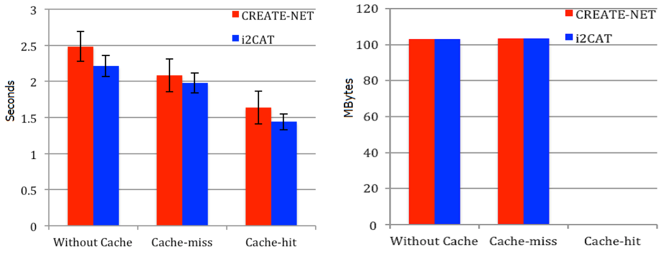
We carried out 20 video streaming requests from the video client to each of the VoD servers, in three scenarios; without a cache (as a baseline), a cache-miss (where content was not found on the OCN, and thus fetched) and a cache-hit (where the content was found and delivered from the OCN). These scenarios produced six unique sets of results (3 per island), which are shown averaged in Table III. For each experimental run, we record three metrics. First, the start up time of the VLC video player when playing back the reference video asset. Second, we recorded the number of cache-miss and cache-hit events on the OCN and the respected fetched and served bytes. This metric essentially demonstrates OpenCache's reduction in external link utilisation and origin server load, which consequently relinquishes resources for use by other users. Third, we recorded the video bitrate of each chunk requested by the client during the playback of the whole video.

The results of our experiments are shown in Table III, Fig. 4a, Fig. 4b and Fig. 5. They clearly demonstrate that OpenCache has reduced the start up delay for clients up to approximately 35%. This is reinforced by relatively low standard deviation values (taking into account that four sites across Europe are involved), which demonstrates a high level of statistical confidence. It is

[1] http://www.bigbuckbunny.org/

TABLE III: VoD Streaming Experimental Results

	CREATE-NET (Italy)			i2CAT (Spain)		
	Without Cache	Cache-miss	Cache-hit	Without Cache	Cache-miss	Cache-hit
Average Start Up Delay (s)	2.484	2.088	1.639	2.212	1.982	1.441
Improvement over baseline (%)	-	16.02	34.02	-	10.40	34.85
Standard Deviation (σ)	0.208	0.225	0.226	0.145	0.138	0.109
External Link Usage (MB)	105.7	105.8	0	105.7	105.8	0



(a) Average Start Up Delay (b) Ext. Link Utilisation

Fig. 4: Experimental Results

particularly interesting to note that we have improved the situation even when the OCN has to fetch the content (compared to the baseline), attributed to the download technique used by the OCN.

Results are similarly optimistic for the external link usage. When the content is stored on the OCN, it is delivered directly to the client and the external link usage gets reduced 100%, essentially to zero bytes. Indicatively, streaming the full 9:56 minute reference video from the OCN, saves approximately 101 MB transfer just for one client session. Without an OCN or with a cache-miss, we observe the full content traversing the external link.

A further advantage that we observed when OpenCache is present, is that the playback client estimates that more bandwidth is available between the client and the OCN. As a result, it requests higher quality video chunks when content is delivered directly from the cache. This is illustrated in Fig. 5, which graphs the bitrate of the requested chunks over the playback of the entire video, with and without OpenCache. It is clearly illustrated that in the case where OpenCache is present, the player requests a bitrate which is over 8 times higher. This is a direct improvement in the quality of the streaming video and consequently the QoE for the end-user. It is important to note that the bitrate achieved with OpenCache present (i.e. 8000kbit/s) is the greatest bitrate available in this case. Evidently, the aforementioned QoE multiplier will be dependant on network conditions and should only be seen as an example. Nonetheless, OpenCache clearly provides the potential to improve the end-user's QoE by increasing the video quality distributed to the client.

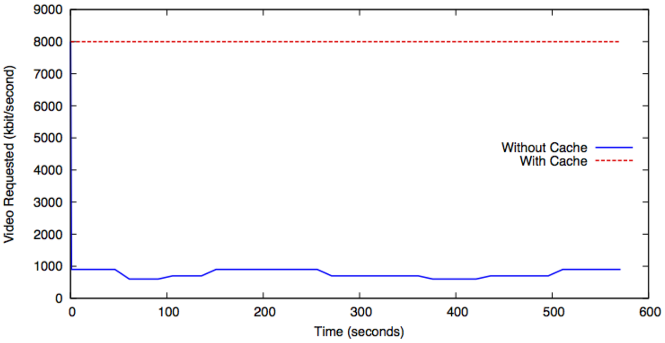


Fig. 5: Quality Requested by MPEG-DASH Client

Finally, it is important to note that the evaluation environment used within OFELIA is one which is relatively bandwidth-rich and well-connected and not necessarily representative of a typical domestic user. Nonetheless, even in such a rich environment, the OpenCache evaluation results show promising benefits to both the network and the QoE of the end-users.

VII. Conclusion

As part of our future work, we plan to improve OpenCache by exploring the cache placement problem in conjunction with different caching policies discussed in related research work. We hope to identify the beneficial position of being in close proximity to end-users whilst maximising the use of the caching resources. In addition, we plan to extend OpenCache's functionality and evaluate it further in different environments and against other commercial or research based caching services.

In this paper we presented OpenCache: an efficient, transparent and highly configurable OpenFlow-assisted in-network caching service for VoD streaming. OpenCache aims to address the underlying challenge that the network faces when the same video files are streamed to end-users repeatedly using independent unicast flows. OpenCache provides the following key benefits :

1) Provides *cache as a service* by offering an interface to declare cacheable content of interest in an open, highly configurable and flexible manner.
2) Supports centrally controlled caching that provides the forum for many additional services to be programmed on top of it with ease (e.g. load balancing, pre-caching, different expiration policies etc.).
3) Is easily deployable in a production network; there are no changes required in the underlying delivery video mechanisms and all existing hardware and software can be retained.

4) Is fully transparent to the end-user; the user does not need to install any extra software, or have to sacrifice any of his local network or storage resources to stream video content with high efficiency, which other technologies require [21], [20].
5) As demonstrated from inter-island experiments on a pan-European testbed using adaptive video streaming, OpenCache provides caching very close to the user with three important benefits. Firstly, the external link usage gets reduced 100% and the network utilisation gets improved as end-user requests are now served locally. Secondly, OpenCache reduces the distribution load from the VoD content provider and all the transient networks along the path of the VoD server to the end-user. Thirdly, since the content is served locally, the video client observes higher throughput, lower latency, higher video quality and smaller start up and buffering times, eventually leading to higher QoE for the end-user.

References

[1] "Visual Networking Index: Forecast and Methodology, 2012-2017," CISCO, Tech. Rep., May 2013.

[2] "Visual Networking Index: Forecast and Methodology, 2011-2016," CISCO, Tech. Rep., May 2012.

[3] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner, "OpenFlow: Enabling Innovation in Campus Networks," *SIGCOMM CCR*, vol. 38, no. 2, pp. 69–74, Mar. 2008.

[4] S. S. Krishnan and R. K. Sitaraman, "Video Stream Quality Impacts Viewer Behavior: Inferring Causality Using Quasi-experimental Designs," in *ACM SIGCOMM IMC 2012*, 2012, pp. 211–224.

[5] F. Dobrian, V. Sekar, A. Awan, I. Stoica, D. Joseph, A. Ganjam, J. Zhan, and H. Zhang, "Understanding the Impact of Video Quality on User Engagement," in *ACM SIGCOMM 2011*, 2011, pp. 362–373.

[6] X. Liu, F. Dobrian, H. Milner, J. Jiang, V. Sekar, I. Stoica, and H. Zhang, "A Case for a Coordinated Internet Video Control Plane," in *ACM SIGCOMM 2012*, 2012, pp. 359–370.

[7] E. Nygren, R. K. Sitaraman, and J. Sun, "The Akamai Network: A Platform for High-performance Internet Applications," *SIGOPS OS Rev.*, vol. 44, no. 3, pp. 2–19, Aug. 2010.

[8] S. Sen, J. Rexford, and D. Towsley, "Proxy Prefix Caching for Multimedia Streams," in *19th IEEE INFOCOM 1999*, vol. 3, 1999, pp. 1310–1319 vol.3.

[9] J. V. D. Merwe, S. Sen, and C. Kalmanek, "Streaming Video Traffic: Characterization and Network Impact," in *Int. Web Content Caching and Distribution Workshop*, 2002.

[10] K. Sripanidkulchai, B. Maggs, and H. Zhang, "An Analysis of Live Streaming Workloads on the Internet," in *4th ACM SIGCOMM IMC 2004*, 2004, pp. 41–54.

[11] P. Georgopoulos, Y. Elkhatib, M. Broadbent, M. Mu, and N. Race, "Towards Network-wide QoE Fairness Using Openflow-assisted Adaptive Video Streaming," in *ACM SIGCOMM 2013 Workshop on Future Human-centric Multimedia Networking (FhMN)*, 2013, pp. 15–20.

[12] B. Wang, J. Kurose, P. Shenoy, and D. Towsley, "Multimedia Streaming via TCP: An Analytic Performance Study," *ACM Trans. MCCA*, vol. 4, no. 2, pp. 16:1–16:22, 2008.

[13] A. Rao, A. Legout, Y.-s. Lim, D. Towsley, C. Barakat, and W. Dabbous, "Network Characteristics of Video Streaming Traffic," in *7th ACM CoNEXT 2011*, 2011, pp. 25:1–25:12.

[14] C. Diot, B. Neil, L. Bryan, and K. D. Balensiefen, "Deployment Issues for the IP Multicast Service and Architecture," *IEEE Network*, vol. 14, pp. 78–88, 2000.

[15] L. Golubchik, J. Lui, and R. Muntz, "Reducing I/O Demand in Video-on-demand Storage Servers," *ACM SIGMETRICS Performance Evaluation Review*, vol. 23, no. 1, pp. 25–36, May 1995.

[16] C. Aggarwal, J. Wolf, and P. Yu, "On Optimal Piggyback Merging Policies for Video-on-demand Systems," in *ACM SIGMETRICS 1996*, 1996, pp. 200–209.

[17] D. Eager, M. Vernon, and J. Zahorjan, "Bandwidth Skimming: A Technique for Cost-Effective Video-on-Demand," in *SPIE Multimedia Computing and Networking 2000*, 2000, pp. 206–215.

[18] K. A. Hua and S. Sheu, "Skyscraper Broadcasting: A New Broadcasting Scheme for Metropolitan Video-on-demand Systems," in *ACM SIGCOMM 1997*, 1997, pp. 89–100.

[19] K. A. Hua, Y. Cai, and S. Sheu, "Patching: A Multicast Technique for True Video-on-demand Services," in *6th ACM MULTIMEDIA 1998*, 1998, pp. 191–200.

[20] J. Pouwelse, J. Taal, R. Lagendijk, D. H. J. Epema, and H. Sips, "Real-time Video Delivery using Peer-to-Peer Bartering Networks and Multiple Description Coding," in *IEEE Int. Conference on Systems, Man and Cybernetics 2004*, vol. 5, 2004, pp. 4599–4605 vol.5.

[21] M. Zhao, P. Aditya, A. Chen, Y. Lin, A. Haeberlen, P. Druschel, B. Maggs, B. Wishon, and M. Ponec, "Peer-assisted Content Distribution in Akamai Netsession," in *13th ACM SIGCOMM IMC 2013*, 2013, pp. 31–42.

[22] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System," in *7th ACM SIGCOMM IMC 2007*, 2007, pp. 1–14.

[23] M. Chesire, A. Wolman, G. M. Voelker, and H. M. Levy, "Measurement and Analysis of a Streaming-media Workload," in *3rd USENIX USITS 2001*, 2001, pp. 1–1.

[24] Squid Proxy Server, http://www.squid-cache.org/.

[25] Netflix Open Connect Platform, http://www.netflix.com/openconnect.

[26] B. Niven-Jenkins, F. L. Faucheur, and N. Bitar, "Content Distribution Network Interconnection (CDNI) Problem Statement," IETF RFC 6707, Sep 2012.

[27] The Floodlight Controller, http://floodlight.openflowhub.org/.

[28] A. Köpsel and H. Woesner, "OFELIA: Pan-European Test Facility for OpenFlow Experimentation," in *ServiceWave*, 2011, pp. 311–312.

[29] ISO-IEC 23009-1:2012 Information Technology, "Dynamic Adaptive Streaming over HTTP (DASH)."