International Journal of Computing and Network Technology

Edge Caching in a Small Cell Network

Debashish Purkayastha, Jun Li, Bartosz Balazinski, John Cartmell and Alex Reznik

InterDigital Communications Corporation, LLC, Wilmington, Delaware, United States of America

E-mail address: debashish.purkayastha@interdigital.com, jun. li@interdigital.com, bartosz.balazinski@interdigital.com, john.cartmell@interdigital.com, alex.reznik@interdigital.com

Received: 26 Aug. 2013, Revised: 29 Oct. 2013, Accepted: 31 Oct. 2013, Published: 1 Jan. 2014

Abstract: An explosive growth of the use of the public internet by mobile phones along with new high speed data technologies increases the challenge of mobile data delivery. The owners of popular media typically use content delivery networks (CDN) to distribute their content. Currently there is no collaboration between the CDNs and the mobile networks.

In this paper we present a novel method for caching at the edge of small cell wireless networks. Our solution allows the small cell network operator to control and manage the access to the storage system at the edge of the small cell network while delegating to the CDN provider the task of managing the content placement.

We define a managed caching architecture that consists of new components deployed in the mobile core networks and within the small cell network. These components are interconnected amongst themselves and with an external CDN using newly defined interfaces which are designed using existing OpenAPI and TR-69 standards.

Keywords: Small cell network; edge caching; cellular networks

I. Introduction

Over the last two decades, the emergence of smart phones and tablets, the availability of high speed mobile data networks and the availability of high definition multimedia and web contents have resulted in an explosive growth of traffic over the internet and mobile core networks. Mobile networks are struggling to deliver a high-quality consumer experience. User data consumption, combined with subscriber growth and user mobility, is putting today’s and tomorrow’s networks under tremendous pressure [1]. The networks under pressure include wireless links, backhaul and mobile core cellular networks. In order to reduce the demand on the network resources, caching at the far edge of the mobile network is a technique which is getting attention.

Owners of popular multimedia content typically use a content delivery network (CDN) to distribute the content to several points of presence and create the CDN edge severs. The CDN control application applies a distribution algorithm to select which edge servers to store a copy of certain content to improve the quality of experience (QoE) for end-users [2]. In practice the CDN edge servers are located outside the mobile core network. This does not help in improving cellular end-user QoE. As a result, frustrated cellular customers abandon attempts to reach content. This abandonment is not in the best interests of the content owners or mobile network operators (MNO).

On the other hand, MNOs are interested in both improving the end user QoE and reducing the bandwidth demand on their core network components [3]. MNOs are deploying small cells along with macro cells to satisfy consumer demand for large bandwidth users. Small cells are emerging as a necessary component of the mobile network. But the small cell technology is further overloading the backhaul network. In this paper the terms mobile network operator and small cell network operator are used interchangeably and they refer to the same entity.

Nonetheless, MNOs are not willing to open up their networks to CDN providers and allow deployment of edge servers within the mobile core network and small cell network. The reasons for this unwillingness are several. The first reason is the absence of a revenue sharing model. The next reason is a concern that external parties may ruin their network. The last reason is a concern about security risks if CDN operators are permitted within the cellular network proper. Furthermore, network operators also lack the expertise to run a CDN system by themselves; they are in the cellular business, not the content providing business.

In this paper we describe a novel method of caching at the edge of a small cell wireless network, which will benefit all the stakeholders, e.g. end-users, mobile network operators and content owner/CDN providers. In this proposed method a storage subsystem is deployed at a point of presence which is closer to the end-users, such as near or at the Lte eNB (Enodeb), than in a traditional CDN configuration. The small cell network operator controls and manages the access, allocation and maintenance of the storage subsystem but delegates to the CDN provider the task of managing content placement in such storage subsystems.

We start by providing background and problems related to caching in small cell networks (SCN) in section II. In section III we describe the assumptions, design decisions, architecture framework, interfaces and an example use case. We conclude in section IV by describing the advantages and disadvantages of the proposed system as well as describe potential future work.

II. Problems of Caching in a Small Cell Network

Providing caching service in a SCN can be very different from traditional CDN services because of following:

- Content is more localized and dynamic – cache hits are less predictable based on global statistics [4], [7]
- Content is shorter and more likely to be partially consumed due to the user mobility – intra-media statistics are needed
- Content must be adaptive to terminal configuration, link quality and link cost – multi-descriptive media is needed

A. Cacheability of Content

Cacheability is defined in [3] as the ratio of revisited content to all requested content. Cacheability can be measured by the number of requests or data volume requested, which is equivalent to the request hit-ratio or byte hit-ratio when there is no cache limit, respectively. Analysis based on the internet traffic traces from a large US mobile operator [3] shows that although the cacheability is about 50% based on the number of requests, the cacheability in data volume is only 9%, which implies limited potential for bandwidth saving. Another analysis [4] shows the average cacheability drops as the number of users decreases. For example, the cacheability drops from 30% for content requests from 1 million users to 10% for content requests from only 10 users. This implies a limited benefit of caching in small cell networks. The cacheability in data volume drops further due to the viewing pattern for large size content, such as videos. Detailed analysis of YouTube video [5] and various other internet user generated content (UGC) [6] show most of users finish less than 10% of the length of the video they view.

B. Transparent versus Managed Caching

Transparent caching assumes a requested content will be revisited again in near future so that subsequent requests to the same content can benefit from caching. As the size of user group is small, the cacheability is relatively low [4]. A study based on a large DSL network [7] has shown that "One-timer" objects (content requested only once) can take up to 76% of the available cache space even though the content will not be revisited. In an SCN, one-timer objects can take an even higher percentage, due to the sparse number of users; therefore, using transparent caches is not suitable for SCNs.

Managed caching intends to pre-fetch popular content into the cache before users' request content. If the pre-fetching is made at off-peak hours, even "one-timer" objects can benefit from caching. However, it is not possible to get popularity of the one-time objects through local statistics in SCNs. Studies revealed that local statistics have only 5% correlation to global statistics based on their internet traces [8]. The small size of the cell, especially when there is extensive user mobility, offers insufficient data to estimate local popularity. On the other hand, content requests tend to be more concentrated to a smaller set of content in a region. Studies in [9] show that less than 1% of requests for video content span more than 20 geographical regions. Although a region defined in [9] is much larger than an SCN, it provides insight into the potential of SCN caching if local popularity can be correctly estimated within a subset of content.

C. Architectural Contraints

Although the cacheability in a small cell network may be low, particularly in terms of data volume, it still attracts attention of both mobile and CDN operators. An economic model analysis [10] based on a typical mobile network shows a 5.1% cache hit ratio can be economically beneficial from mobile operator's point of view.

A CDN technology survey whitepaper [11] reviews the challenges to CDN operators in today's content oriented markets. It claims CDN operators must play more specialized roles in different market sectors which may have different needs from that of conventional web servers.

Mobile operator networks require managed caching to increase the efficiency of caching. This can be achieved in two ways. One is an internal CDN that is specialized to the mobile network requirements. A research paper [12] proposes a mobile CDN design in a mobile network jointly deployed with distributed mobility management (DMM). The second option is the use of a special node at the peering point between Mobile core and Internet. Research paper [2] proposes a video streaming scheduler for a mobile CDN (mCDN) architecture. It defines a mobile CDN serving point (mCSP) working as a bridge between the external content source and the mobile network clients. This solution needs a standard interface for the mCSP to both the internal scheduler and external content sources.

In general, SCNs operated by mobile operators are not easily accessible by third party application providers. Original content owners and CDN providers, collectively called "CDN Application" in this paper, fall into this category and are restricted from accessing the SCN. Mobile operators are unwilling to open their SCNs in order to retain control over their network. On the other hand mobile operators do not have the expertise to provide managed caching services within their network. In order to overcome this stalemate situation, an architectural modification and new interfaces are required. The balance of this paper offers the architecture and new interfaces to support this architecture. The goal is to allow CDN applications to provide caching services at the far edge of the mobile network without mobile operators giving up control of their network infrastructure.

III. Managed Caching Architecture

In this proposed managed edge caching solution, mobile network operators deploy and control the storage subsystems and allow the CDN applications to manage the content such as content prepositioning, content placement, content replication, content deletion etc. The CDN applications use the storage subsystem in the SCN for caching popular multimedia contents by creating virtual edge servers via the mobile operator network. This paradigm creates a need for new services in the mobile core network to allow all CDN applications to have a single point of contact towards the wireless network and all its SCNs.

A CDN application can improve the content distribution algorithm it uses by receiving wireless network related statistics and storage related information from wireless and small cell core network components. Several wireless networks may cooperate and coordinate in the creation of virtual edge servers, which can provide a single interface to be used by CDN applications. Prepositioning and pre-fetching by CDN providers will improve when SCN related statistics are available to them along with global statistics.

Fig. 1 describes the managed edge caching scenario by showing the new architecture, the new components and the new interfaces to these entities.

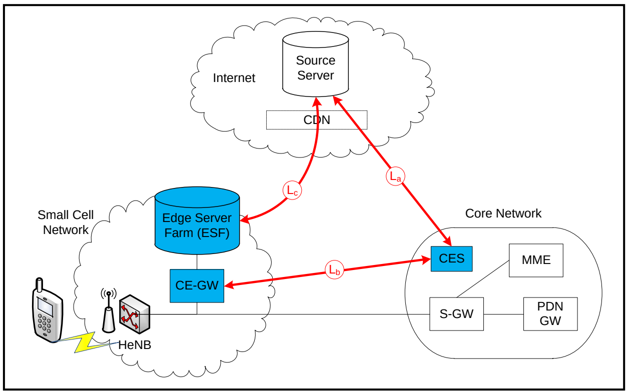

Figure 1. Managed Edge Caching Architecture

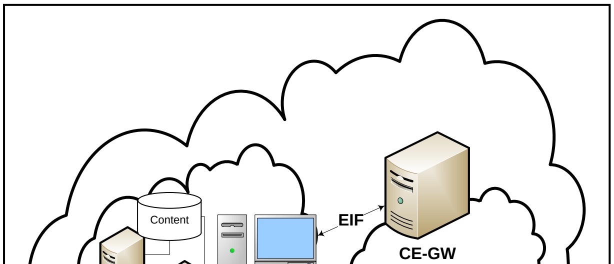

Figure 2. Edge Server Farm

The storage subsystem, called the Edge Server Farm (ESF), consists of storage and application servers. The management entity within the ESF is the ESF Control and Management System (ECMS). Fig. 2 shows the details of an Edge Server Farm (ESF).

The ESF is placed under the control of a small cell gateway or aggregator. Within this gateway, a new function is introduced, which manages the ESF and interfaces to the mobile core network. This function is the Content Enabled Gateway (CE-GW). CE-GW manages and controls the ESF over the Edge Server Farm interface (EIF). This interface is internal to SCN and not exposed to external entities. The CE-GW uses another internal interface to communicate with eNB and APs.

The node in the core network to which the CE-GW communicates is the Content Enablement Service (CES). This is the central control entity in the operator core network. The CDN control application in the public internet uses the services of CES and CE-GW (via the CES), to acquire SCN storage subsystem capabilities information, network statistics, etc. It allows a CDN application to create, modify and delete virtual storage at a particular SCN. These actions are carried over two new interfaces, La and Lb, which are described in more detail later in this paper. After creating virtual storage, a CDN application performs managed edge caching by causing the ingestion of popular multimedia content into the ESF directly using the Lc interface.

A. Assumptions

There are several assumptions required to implement this architecture and system design. The CES might be an operator owned service in a similar way to those services offered within the OpenAPI framework [13]. A CES might interact with several SCNs, where some of them contain a storage subsystem while others do not have any storage.

Any storage within a SCN can be shared by several CDN networks to deliver varied content to the users of the small cell network. The CDN requests for virtual storage and network statistics are authenticated and authorized by the CES. The interaction on the Lc interface between the CDN and Edge server is also secured.

B. Design Decisions

There were several design decisions made for the prototype developed by us. The first two decisions relate to the La and Lb interfaces. For the La interface, the GSMA One API is used. For the Lb interface, the CPE WAN Management Protocol (CWMP) based on the Broadband Forum TR-069 specification is utilized.

A platform was needed to host this functionality. We developed a small cell gateway denoted as the Converged Gateway (CGW) [14]. The CGW is a proprietary node that permits the aggregation of several eNBs and WiFi hotspots. The CGW offers features such as IP Flow Mobility (IFOM) and traffic load balancing to UEs that are capable of being simultaneously attached to the mobile network and an associated WiFi AP. Given that it sits at the edge of a small cell network, it makes a natural place to host the CE-GW. Other feasible locations for the CE-GW are the H(e)NB and the Local Gateway (LGW) defined in the 3GPP standards [15].

The CGW platform is transparently inserted on the data path between the UE and the EPC and also between the WiFi AP and the Internet. Given its location and it functionality, the CGW platform can perform Deep Packet Inspection (DPI) and can treat each user data flow individually. This ability to access user data flows themselves allows for resolving requests to the local ESF if the ESF has the content that is requested. This resolving to the ESF is handled by a local DNS Server.

C. System Architecture

The architecture is shown in Fig. 3. The CGW comprises the IP Traffic Gateway, Content Enablement Gateway (CE-GW), DHCP Server and DNS Server. The CGW is connected to HeNB(s) and the Wi-Fi AP(s) in a small cell network. It is also connected to the CES within the cellular core network, and to the CDN/content servers which are located in the public internet.

1) Content Enablement Gateway (CE-GW)

The CE-GW sits in the small cell network and facilitates the enablement of virtual storage at the ESF and content ingestion towards the ESF by the CDN. It communicates with the CES over the Lb interface as noted previously. The CE-GW interacts with the ESF using the EIF interface. In addition, the CE-GW maintains the latest information about the small cell network status and the edge server farm status and sends it to the CES using the Lb interface.

2) Edge Server Farm (ESF)

The ESF is the storage subsystem in the SCN. It provides the storage space to the CDN control applications to perform the managed edge caching. The ESF comprises storage with multimedia streaming and logging services. It may also incorporate other features or functions. The storage space and the data store services will be shared among the several CDN applications which are subscribing to the service; therefore, each CDN can get some share of the physical storage. The ESF supports the functional decomposition of the total storage space into customized virtual edge servers.

3) Content Enablement Service (CES)

The CDN control applications will have a single interface towards the CES in the operator core network. The CES will have a database comprising the SCN information, storage subsystem related information, data store service capabilities information, etc. It will facilitate the SCN service enablement by routing the virtual edge server (VES) management messages from CDN control applications towards the respective CE-GWs in the various SCNs.

4) Interfaces

This section describes the various interfaces that were introduced in Fig. 1 as part of the cache architecture shown in Fig. 3.

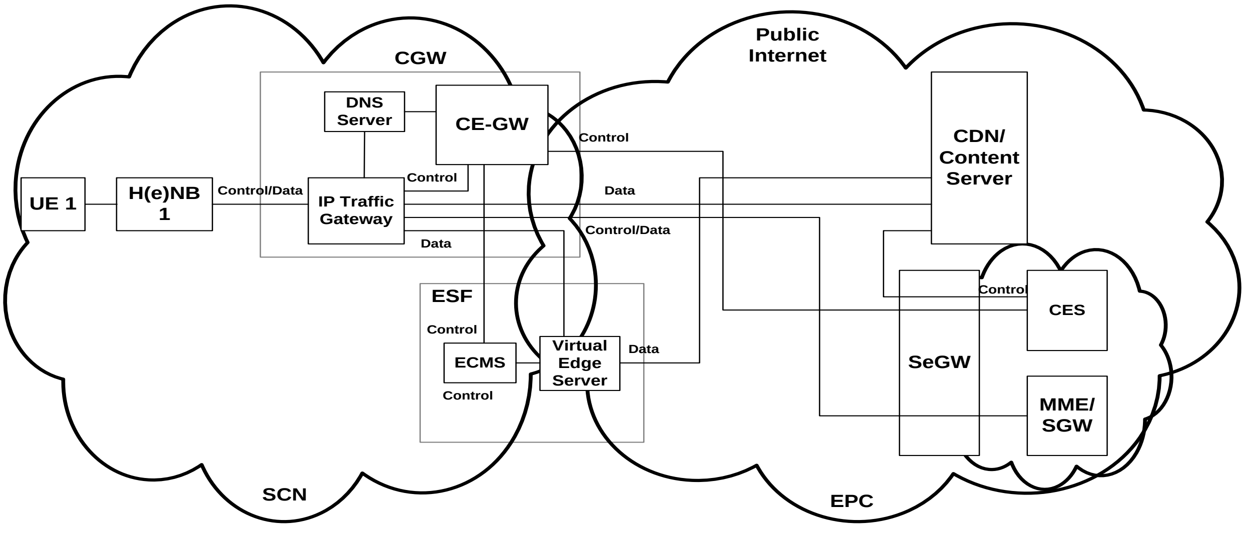

Figure 3. Managed Edge Caching Functional Architecture

a) La Interface

The La interface will be used for communication between the CDN application and the CES in the core network. The CDN application uses this interface to connect to the CES service, where the CDN account (username/password) goes through typical authentication and authorization techniques to allow access to the CES service. The communication between the CDN and CES can use secured and non-persistent connections using an HTTP protocol such as HTTPS. Subsequently, a CDN application may use this connection interface to:

- Query the CES service about the available storage subsystems within the SCNs. The response from the CES will carry a map of the SCNs with the amount of available storage in each location.
- Create, update, delete or remove some of the already allocated storage at certain small cell networks.
- Request the logging service within certain SCNs to report wireless related statistics as well as QoE related parameters.
- Configure resources for the allowance of streaming protocols such as RTSP, SIP, HTTP progressive or HTTP dynamic adaptive streaming (DASH) within the CE-GW.

The La interface functionality is realized using the Femtozone OneAPI services [13]. The OneAPI server will be placed at CES and the CDN control applications will run the OneAPI client.

b) Lb Interface

The Lb interface will be used for communication between the CES in the core network and the CE-GW in the SCN. The CES maintains the latest information about each SCN and the available amount of storage at each location using a traditional database system. Therefore, this interface is mainly used to open and close connections with several SCNs to query their latest status. Secured connections can be used as the transport protocol for this interface. The following CES prompted activities are performed over this interface:

- Query SCNs about their physical location and/or network topology in order to organize a response to a CDN in a map or graph format.
- Query an SCN about the availability of an externally exposed ingestion URI to be used by a requesting CDN for sending content to the SCN. This URI might be used by the operator's DNS Server to resolve to an IP address within the associated SCN.
- Negotiate the QoS/QoE parameters and the set of allowed data transport protocols with a particular SCN.
- Enable a detailed logging service which can be used by the charging function of a core network.

c) Lc Interface

The Lc is the communication interface between a virtual edge server and CDN control application. After receiving the ingestion URI from the CES, the CDN application can store content material in the allocated storage within the SCN. The ingestion URI is the address to which the CDN can send content to be stored at the SCN.

The CDN application has the freedom to organize the stored contents in any hierarchy, using any suitable fault tolerance technique or even using a digital encryption scheme. By using this interface, the CDN application can apply a cache placement strategy. During peak hours, the cache placement strategy might allow read-only operation on the allocated storage, while read-write operation is allowed in any other hours.

The Lc interface is used to enable the following operations:

- CDN control application pushing/updating the media content
- CDN control application deleting the media content
- CDN control application retrieving log files
- Log Server reporting the virtual edge server performance statistics
- Log Server reporting the content usage statistics

5) Use Case

A high level use case is shown in Fig. 4 and is described here to capture the managed caching functions such as querying the SCN related information, virtual edge server service procedures, content management procedures, subscription and notification procedures and fetching SCN related statistics. The following are the list of supported use cases by this architecture:

- Query of the Available SCN storage subsystems
- Query of the SCN storage subsystem capabilities
- Create virtual edge server
- Delete virtual edge server
- Update virtual edge server
- Query Status of Virtual Edge Server
- Content management on the virtual edge server by a CDN control application
- Subscribe for notification services
- Report virtual edge server performance and content usage statistics

Steps 1-4: The CDN control application logs in to the CES database using the La interface procedures described previously. It then queries the CES database to acquire the available SCN storage subsystem information.

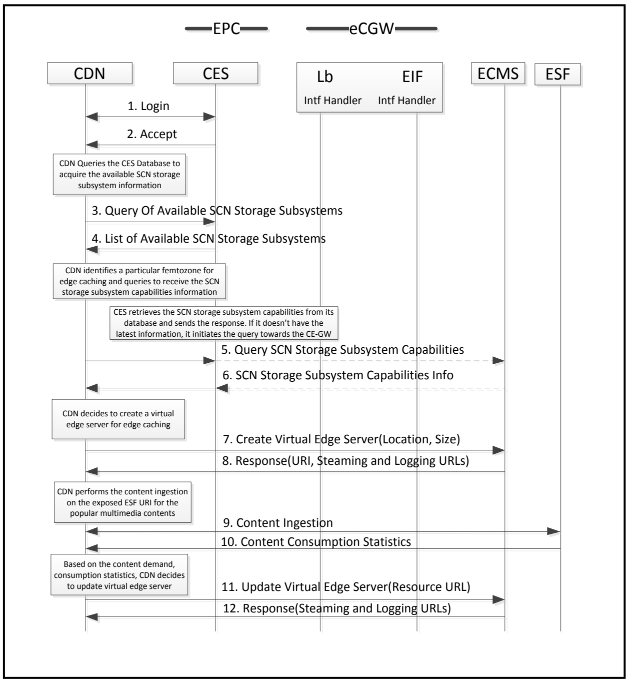
EPC
eCGW
CDN
CES
Lb
Intf Handler
EIF
Intf Handler
ECMS
ESF
1. Login
2. Accept
CDN Queries the CES Database to acquire the available SCN storage subsystem information
3. Query Of Available SCN Storage Subsystems
4. List of Available SCN Storage Subsystems
CDN identifies a particular femtozone for edge caching and queries to receive the SCN storage subsystem capabilities information
CES retrieves the SCN storage subsystem capabilities from its database and sends the response. If it doesn't have the latest information, it initiates the query towards the CE-GW
5. Query SCN Storage Subsystem Capabilities
6. SCN Storage Subsystem Capabilities Info
CDN decides to create a virtual edge server for edge caching
7. Create Virtual Edge Server(Location, Size)
8. Response(URI, Steaming and Logging URLs)
CDN performs the content ingestion on the exposed ESF URI for the popular multimedia contents
9. Content Ingestion
10. Content Consumption Statistics
Based on the content demand, consumption statistics, CDN decides to update virtual edge server
11. Update Virtual Edge Server(Resource URL)
12. Response(Steaming and Logging URLs)

Figure 4. Managed Edge Caching Procedures

Steps 5-6: The CDN identifies a particular femtozone for edge caching and queries the CES to receive the SCN storage subsystem capabilities information from the CES database. If the CES database has the latest information it provides it to CDN, otherwise it queries the CE-GW over the Lb interface. Next, the CE-GW gets the capabilities information from the ECMS over the EIF interface.

Steps 7-8: The CDN decides to create a virtual edge server for edge caching at a particular SCN. It initiates the virtual edge server creation procedure. As a result, it receives the content ingestion URI to be used for content placement on the edge server.

IV. Conclusion

In this paper we presented an overview of the problems regarding edge caching in an SCN. An architecture framework has been proposed, which allows network operators to control the network resources and CDN applications to manage the actual cache content. It also allows mechanisms to increase cacheability of content by providing local SCN statistics to CDN applications. We have presented some of the building blocks needed and new interfaces required to implement such a framework. An example use case has also been presented to illustrate the caching mechanism.

The benefits of the architecture provided in this paper are as follows:

- Clear separation of responsibilities among the stakeholders
- Easily manageable hierarchical architecture
- Capability to improve cacheability by collecting local statistics and data, which may be used as an input for cache pre-fetching strategy

We plan to enhance the architecture framework and continue work in the future to include:

- Refinement of data collected within a SCN to generate more accurate content placement
- Measurement and estimation of backhaul conditions, which can be used to determine cache placement strategy
- Delivery mechanism which adapts to cache content, wireless condition and backhaul condition

References

[1] Cisco Visual Networking Index, Global Mobile Data Traffic Forecast Update, 2012-2017, http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.pdf

[2] Yousaf, Faqir Zarrar, Marco Liebsch, Andreas Maeder, and Stefan Schmid. "Mobile CDN enhancements for QoE-improved content delivery in mobile operator networks." Network, IEEE 27, no. 2 (2013): 14-21.

[3] Content Delivery Summit 2013, http://www.contentdeliverysummit.com/2013/Agenda.aspx

[4] Ramanan, Buvaneswari A., Lawrence M. Drabeck, Mark Haner, Nachi Nithi, Thierry E. Klein, and Chitra Sawkar. "Cacheability Analysis of HTTP traffic in an Operational LTE Network.", Wireless Telecommunications Symposium (WTS), April 17-19, 2013

[5] Erman, Jeffrey, Alexandre Gerber, Mohammad Hajiaghayi, Dan Pei, Subhabrata Sen, and Oliver Spatscheck. "To Cache or not to Cache: The 3G case." *Internet Computing, IEEE* 15, no. 2 (2011): 27-34.

[6] Braun, L. ; Klein, A. ; Carle, G. ; Reiser, H. "Analyzing caching benefits for YouTube traffic in edge networks — A measurement-based evaluation", Network Operations and Management Symposium (NOMS), 2012 IEEE

[7] Lucas C.O. Miranda et al. "Characterizing Video Access patterns in Mainstream Media Portals", WWW 2013 Companion, May 13-17, 2013, Rio de Janeiro, Briazil

[8] Abid, Saloua Messaoud, and Habib Youssef. "Impact of One-Timer/N-Timer Object Classification on the Performance of Web Cache Replacement Algorithms." In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, vol. 1, pp. 208-211. IEEE, 2010.

[9] Zink, M., Suh, K., Gu, Y., & Kurose, J. (2008, January). Watch global, cache local: YouTube network traffic at a campus network: measurements and implications. In *Electronic Imaging 2008* (pp. 681805-681805). International Society for Optics and Photonics.

[10] Brodersen, Anders, Salvatore Scellato, and Mirjam Wattenhofer. "Youtube around the world: geographic popularity of videos." In *Proceedings of the 21st international conference on World Wide Web*, pp. 241-250. ACM, 2012.

[11] Catrein, Daniel, Bernd Lohrer, Christoph Meyer, René Rembarz, and Thomas Weidenfeller. "An Analysis of Web Caching in Current Mobile Broadband Scenarios." In New Technologies, Mobility and Security (NTMS), 2011 4th IFIP International Conference on, pp. 1-5. IEEE, 2011

[12] Content delivery networks: Market dynamics and growth perspectives, whitepaper, Informatandm.com

[13] GSMA OneAPI, http://www.gsma.com/oneapi/

[14] Cartmell, John, "LTE Converged Gateway IP Flow Mobility Solution," 7th IASTED CIIT 2012, May 2012, Baltimore, MD, USA

[15] 3GPP TS 32.467, "UTRAN architecture for 3G Home Node B (HNB); Stage 2 (Release 10), v10.1.0, March 2010.