Optimal Video Delivery in Mobile Networks Using a Cache-Accelerated Multi Area eMBMS Architecture

Ioannis M. Stephanakis, Ioannis P. Chochliouros, George L. Lymperopoulos, Kostas Berberidis

Optimal Video Delivery in Mobile Networks Using A Cache-Accelerated Multi Area eMBMS Architecture

Ioannis M. Stephanakis[1], Ioannis P. Chochliouros[2], George L. Lymperopoulos[3], Kostas Berberidis[4]

[1] Hellenic Telecommunication Organization S.A. (OTE),
99 Kifissias Avenue, GR-151 24, Athens, Greece
stephan@ote.gr

[2] Research Programs Section, Hellenic Telecommunication Organization S.A. (OTE)
99 Kifissias Avenue, GR-151 24, Athens, Greece
ichochliouros@oteresearch.gr

[3]Head of Network Evolution Dept., Fixed & Mobile,
COSMOTE Mobile Telecommunications S.A.
Pelika & Spartis 1 Str., GR-151 22, Maroussi, Athens, Greece
glimperop@cosmote.gr

[4] Head of Signal Processing and Communications Lab
Dept. of Com. Engineering & Informatics, University of Patras, 26500 Patras, GR
berberid@ceid.upatras.gr

Abstract. Long-Term Evolution (LTE) evolved into enhanced *Multimedia Broadcast/Multicast Service* (eMBMS) that features improved perfomance, higher and more flexible LTE bit rates, Single Frequency Network (SFN) operations and carrier configuration flexibility. Multiple eMBMS service areas allow for efficient spectrum utilization in the context of mobile *Content-Delivery-Networks* (CDNs) as well as for delivering broadcast and push media over modern broadband networks. This paper intends to highlight novel service architectures for efficient content delivery. Content caching is a widely used technique in the networking industry that brings content closer to end-users improving, thus, servive performance and latency. Next generation wireless networks use edge located cache as well as core located cache in order to reduce traffic between the gateway and the internet and make the response of mobile network faster. This is referred to as hierarchical/distributed caching. Dimensioning of individual scenario-based traffic models for health care, museum virtual tours and interactive educational use cases in the context of the *LiveCity* European Research Project is attempted in this work. ON/OFF source models featuring state dependent active and inactive periods are used in order to describe multi-threaded resource transmissions. A group of chain ON/OFF models holds the parameters of the basic services and defines cache requirements. An architecture that associates distributed cache deployed at eMBMS Gateways per service area is proposed in order to reduce backhaul traffic. A simple algorithm that determines the optimal cache size in a mixture of chain ON/OFF modeled services is presented.

Keywords: LTE, Evolved Multimedia Broadcast/Multicast Service (eMBMS), Single Frequency Network transmission, DVB-NGS, cache optimization, ON/OFF Chain Models, traffic modeling

1 Introduction

Recent advances in the delivery of multimedia content are investigated in this paper. They allow for enhanced user experience and for novel offering from telecom providers. The Third Generation Parnership Project (3GPP) defined multimedia broadcast/multicast service in 2005 in order to optimize the distribution of video traffic in 3GPP specifications release 6 (Rel-6) for Universal Mobile Telecommunications System (UMTS). The standard refers to terminal, radio, core network, and user service aspects. MBMS standard eventually evolved into enhanced MBMS (eMBMS) [1] that features improved perfomance, higher and more flexible LTE bit rates, single frequency network (SFN) operations and carrier configuration flexibility. The introduction of MBMS over Single Frequency Network (SFN) transmission (MBSFN), which is described in Rel-7 of 3GPP specification version, overcomes cell-edge problems of MBMS and allows for the increase of the capacity of broadcast channels by a factor up to 3 or 4 in certain deployment conditions [2]. An identical waveform is transmitted from multiple cells with a tight synchronization in time in SFN operation. Nevertheless, it is not possible to use the same frequency for MBMS and non-MBMS transmissions in UMTS deployments making. Current 3GPP Rel-11 (2012) specifies improvements in the areas of service layer and Coordinated Multipoint Operation (CoMP) and allows for offloading the LTE network and mobile backhaul through eMBMS. Pushed content via user equipment caching as well as machine-to-machine services are distinctive enhancements. The evolution of the MBMS standard and its performance are presented in many scientific and technical papers [3,4,5,6]. Its efficiency in content delivery is addressed in [7].

eMBMS enables the possibility to deliver premium content to many users with secured quality of service in defined areas. It efficiently delivers rich media to mass users over unicast and broadband/multicast transmissions. A user may send and receive data individually over one-to-one transmissions in the context of such applications as Video-On-Demand, e-mail services, web-browsing and media downloads. One-to-many transmissions on the other hand utilize mobile spectrum more efficiently and result in lower cost for common content. They are used for live video and audio streaming, push media, e-publications, application downloading and other services. Several telecommunication companies adopt eMBMS as an efficient and low-cost means to deliver multimedia content. Qualcomm and Ericsson[1] demonstrated their eMBMS solution at Mobile World Congress (MWC) in 2012. Alcatel-Lucent and Huawei have also introduced their end-to-end solution to support broadcast video delivery in LTE networks. Verizon and Telstra announced plans to launch a live video broadcast service for sport events based on eMBMS technology after conducting live tests. Vodafone Germany has conducted live tests with LTE Broadcast in collaboration with Ericsson and Qualcomm as well[2].

[1] Ericsson at *http://www.ericsson.com/res/docs/whitepapers/wp-lte-broadcast.pdf* last accessed 30th of April 2014.

[2] GlobeNewswire at *http://globenewswire.com/news-release/2014/02/25/612970/10069866/en/Europe-s-first-live-trial-of-LTE-Broadcast-revolutionizing-video-delivery-across-mobile-networks.html* last accessed 30th of April 2014.

Cache-accelerated architectures are suggested for current deployments of broadband and content delivery networks (CDNs). Dynamic site acceleration is a suite of technologies that make websites reliant on dynamically served content. The adoption of such technologies makes internet applications perform better and load faster. Traditional CDNs improved performance of telecommunication networks by caching critical content closer to end users. *Software-as-a-Service* (SaaS) and novel enterprise applications (B2B and B2C) base their functionality upon such notions as personalized recommendations, transactional and secure check-out and shopping carts. This poses stringent requirements on the delivery of dynamic, transactional content, as well as on the demand for e-commerce and web retailers. Traditional cache-accelerated architectures implement most or all of the following technologies (most of which are dealing with optimizing bit delivery across the network):

- **TCP optimization,** which includes algorithms designed to handle network congestion and packet loss.
- **Route optimization**, which is a technology that optimizes the route of the request from the end-user to the origin and ensures the reliability of the connection.
- **Connection management**, that includes persistent connection provision and HTTP multiplexing. Reusable and optimized HTTP connections from the edge servers to the origin servers as well as between the edge servers themselves are maintained rather than initiating a new connection for every request.
- **On-the-fly compression**, that compresses text objects shortly after they leave the origin servers alleviating, thus, computational burden from the origin servers without requiring additional bandwidth or hardware.
- **SSL offload** in order to speed up the critical secure transaction processes such as a check-out at an online store.
- **Pre-fetching**, that will parse through a served HTML file and will prefetch from the origin server the cacheable objects embedded in the file. A CDN can prefetch only cacheable content. Dynamic content by definition is contextual and can only be requested by the user.
- **Whole-site delivery,** that identifies cacheable and dynamic content and whether dynamic site acceleration techniques are applicable to the dynamic content of a site instead of simple fetches from the origin server.

Novel approaches have introduced such techniques as preresolving (performing DNS lookup in advance), preconnecting (opening a TCP connection in advance) and prewarming (sending a dummy HTTP request to the origin server) [8]. Response time for a mobile application is defined as the time between clicking to request information and loading a web page. Specifying cache size requires accurate modeling of application traffic. ON/OFF internet traffic models are frequently used for such a cause [9]. Measuring analysis and modeling of traffic has still been one of the main research challenges. Several studies have been carried-out over the last fifteen years on analysis of network traffic in the internet [10,11], traffic measurements in high speed networks [12] as well as measurements in next generation networks [13]. A brief description of the novel features and the architecture of eMBMS is presented in *Section* 2 whereas a comparison with the DVB-NGH standard - which is currently under development - is attempted is

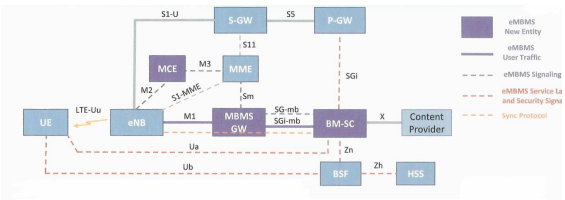

Fig. **1**. eMBMS new entities

Section 3. Dimensioning issues for cache acceleration are considered in *Section* 5 and a novel approach for dynamic parameterization of buffers at edge servers based upon optimal storage for chain ON/OFF modeled services in the context of eMBMS is proposed and evaluated.

2 Evolved Multimedia Broadcast/Multicast Service (eMBMS) in LTE-Advanced

2.1 Features And Services

An evolved architecture is required to support eMBMS transmission in LTE network. The novel logical network entities proposed for eMBMS operation are the following [14]:

- **BM-SC** (**Broadcast Multicast Service Center**), which is the entity that connects the Content Provider and the Evolved Packet Core. This entity plays the role of traffic shaper and authorizing content provider/terminal request. It is in charge of SYNC protocol to synchronize transmitted data among eNBs. SYNC protocol associates a specific header to IP packets, providing Time Stamps and session information.
- **eMBMS Gateway**, which is the entity located between BM-SC and all eNB. Its principal function is to deliver MBMS user data packets to eNBs by means of IP Multicast. When an MBMS session arrives, it allocates IP mulicast address to which the eNB should join to receive MBMS data and maintains the IP Multicast group. Furthermore, eMBMS Gateway is responsible for MBMS session announcement and it also performs MBMS Session control Signaling (Session Start/Stop) toward E-UTRAN.
- **MCE** (**Multi-cell Coordination Entity**), which is the logical entity whose function is admission control and allocation of the radio resource use for MBSFN operation. The MCE is expected to be part of e-UTRAN and can be integrated as a part of a network element. MCE may be either part of eNB (a "distributed MCE architecture") or be a stand-alone MCE ("centralized MCE architecture").

Interfaces M1, M2 and M3 (see Fig. 1) connect the aforementioned new entities with each other. M1 interface is a user plane interface that connects e-MBMS-GW and eNB. M2 interface is a control plane interface between MCE and eNBs whereas M3 interface connects MME and MCE. M3-Application Part allows for MBMS Session Control Signaling on E-RAB level.

MBMS defines three MBMS-specific radio bearers, i.e. **MICH** (MBMS indicator channel) which is used to notify terminals about the imminent start of an MBMS transmission session, **MCCH** (MBMS control channel), which carries control information about all ongoing MBMS transmission sessions and **MTCH** (MBMS traffic channel), which carries the actual data of an

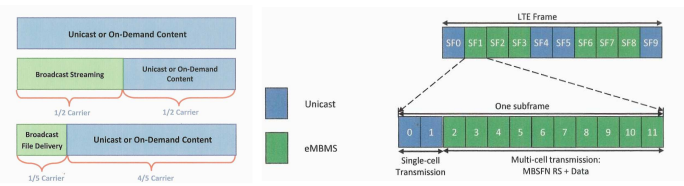

Fig. **2**. Flexible spectrum allocation between unicast-broadcast services

MBMS transmission session. Since HTTP cannot be used on unidirectional links, a unidirectional protocol was introduced for delivery of files or file segments, the so-called **FLUTE** protocol (File Delivery over Unidirectional Transport - RFC 3926). FLUTE protocol is carried over the User Datagram Protocol (UDP) and IP multicast toward end-user devices.

3 GPP Rel. 11 enhancements of eMBMS specify a mechanism that allows end-user devices to retrieve missing parts of a file after a broadcast session is over. This is due to the fact that it is impossible to use hybrid automatic repeat request (HARQ) on lower layers of unidirectional bearers. This mechanism is called file repair. Adaptive HTTP streaming, that is adopted by the Moving Pictures Experts Group as the baseline of MPEFG-DASH is specified in 3GP-DASH [15] as a compatible to MPEG-DASH profile. LTE resources reserved for eMBMS only when needed and there is no impact on LTE unicast capacity at other times (see Fig. 2).

2.2 Network Architecture And eMBMS Service Areas

A synchronous LTE network allows broadcast over a *Single Frequency Network* (MBSFN). An MBSFN area is defined as the set of cells participating in the same SFN transmission. Should different services be transmitted in the same cell but have different coverage areas, the cell transmits them in different time intervals where a different set of cells participate to the same SFN transmission (the cell/eNB may belong to up to eight MBSFN areas). The maximum extension of MBSFN areas is determined by the size of areas where cells are synchronized, which is called synchronization area. MBSFN transmissions not only include the data of the MBMS services but also periodic MBMS control and scheduling information messages indicating which services are currently transmitted and in which time intervals they may be transmitted. Efficient signal combining at user terminals achieves high operating SNR. Key Use Cases that take advantage of such an architecture include streaming video for real-time situational awareness, Push-To-Talk, file delivery for non-real-time services such as all points bulletin (e.g. amber alert), software updates/upgrades etc. Scalable push file delivery and delivery of content services are described in [16,17].

3 Comparison With Digital Video Broadcasting (DVB-NGH)

At the time MBMS over UMTS terrestrial radio access network (UTRAN) was standardized, the industry was focused on the mobile TV use case. Digital video broadcasting for handheld (DVB-H) networks was being deployed in several countries, and MBMS seemed like a way for mobile broadband operators. Alternative technologies for multimedia broadcasting that emerged were DMB (Digital Multimedia Broadcasting) deployed mainly in South Korea, ISDB 1-seg in Japan and in South America and MediaFLO (Forward Link Only) [18] in the USA. Qualcomm had conducted MediaFLO technical trials internationally at the time but discontinued development in 2010 whereas Japan has moved forward with ISDB-Tmm (Terrestrial mobile multi-media), which is a variant of the existing standard.

Digital video broadcasting for handheld (DVB-H) networks is a standard which is compatible with digital terrestrial TV (DVB-T) and transmits through digital

terrestrial channels taking advantage of the digital dividend. Nevertheless DVB-H has been a commercial failure and the service is no longer on-air. Finland was the last country to switch-off signals in March 2012. The DVB group made a "call for technologies" for a successor system (DVB-NGH – Next Generation Handheld) in November 2009 in order to update and replace the DVB-H standard for digital broadcasting to mobile devices. The schedule was for submissions to be closed in February 2010. The new ETSI standard [19] published in 2013 and rollout of the first DVB-NGH devices is planned for 2015. DVB-NGH defines the next generation transmission system for digital and hybrid broadcasting to handheld terminals, i.e. a combination of digital terrestrial and digital satellite transmissions. The standard is based on the DVB-T2 system and reuses or extends lot of novel concepts introduced in the DVB-T2 specification. It adopts SFN transmission.

5 Cache Acceleration in CDNs And Dimensioning Considerations For Several Use Cases

Web is considered as a set of servers and clients, i.e. web browsers and any other software that is used to make a request of a Web server. Requests and responces of the hypertext transfer protocol (HTTP) before the establishment of a new connection determine the delays in retrieving some resource. It is estimated that the amount of time that is required can be approximated by two round-trip times plus the time to transmit the response and any additional DNS resolution delays. Caching is performed in various locations throughout the Web including at the two end locations. Proxy caches are intermediary caches between the client machine and the origin server. They generate new requests on behalf of the users if they cannot satisfy the requests themselves. Next generation wireless networks use edge located cache as well as core located cache in order to reduce traffic between the gateway and the internet and make mobile network response faster. This is refered to as hierarchical/distributed caching. *Internet Cache Protocol* (ICP) is an example of a popular mechanism used for coordinating caches in many different locations. The ICP protocol is described in RFC 2186 and its application to hierarchical web caching in RFC 2187. *Hypertext Caching Protocol* (TCP) (RFC 2756) is designed as a successor to ICP. *Object hit rate* and *byte hit rate* are used as measures in dimensioning cache size by network operators. Response time for mobile applications is determined by server processing time, delays in the network (bandwidth availability, round-trip time-RTT and tower connect time) as well as delays in the client device (parse time, resource fetches, layout and render processing and JavaScript delays). Less bandwidth, which is subject to frequent changes, is dedicated to downloading in mobile networks compared to the bandwidth which is dedicated to downloading in fixed networks. Furthermore processing speeds of mobile devices are up to ten times slower as compared to desktops. The path to faster mobile delivery therefore lies in reducing the number of round-trips as much as possible, reducing content payload size as much as possible, defer as much as possible and parse as little as possible.

One needs to classify and model traffic patterns at the edge of the mobile network in order to dimension cache. Chain ON/OFF models are adopted to this end. The design and development of a chain ON/OFF model relies on an accurate description of traffic entities from link level to application level. The model is generally used,

when it is necessary to capture the scaling behaviors of network traffic. A chain ON/OFF model uses two states, namely ON & OFF for each state of a Markov model [9]. ON and OFF periods are usually heavy tailed processes. ON times correspond to resource transmissions while the OFF times correspond to intervals of browser inactivity. Furthermore, OFF times are classified either as *Active* (which account for client processing delays like document parsing and resource rendering) or as *Inactive* (i.e. user think time). The Weibull [20] and the Pareto [21] distributions are used to model ON and OFF times. Data generated by a specific type of application/service are discribed by a distinctive chain ON/OFF model. A queue that stores content at the edge of the mobile network is shared by different chain ON/OFF sources that are assumed to be statistically independent. The overall traffic through the edge aggregation link is modeled as a mixture of N chain ON/OFF sources belonging to M different groups, i.e. $\left\{\frac{n_1}{N}Proc_1, \frac{n_2}{N}Proc_2 ... \frac{n_M}{N}Proc_M\right\}$. The chain ON/OFF sources which belong to a specific group are characterized by the transition probabilities between states as well as the mean number of packets/cells generated during the ON states and the relationship between ON and OFF periods. The probability of accessing one of the resources stored in a proxy cache resembles accessing a word in a written text of a natural language and can be calculated through the Zipf distribution [22,23]. Specifically, the number of references to cached item *i, NR_i,* is given as follows,

$$NR_i \sim \frac{\alpha}{rank(i)^Z}\ , \qquad (\mathbf{1})$$

where *rank(i)* denotes the position of cached item *i* after the sorting of the cache population on the basis of references (i.e., the most popular item is ranked first, the second most popular is ranked second and so forth), α is a constant value while Z is the Zipf parameter, which assumes a value close to 1. Should one choose α so as to normalize the sum of all references to a unit probability, he gets,

$$P_{hit} = \sum_{i=1}^{cache\ size} NR_i = \sum_{i=1}^{cache\ size} \frac{a}{rank(i)^z}, \qquad (\mathbf{2})$$

LiveCity use cases include medical tele-monitoring for patients' treatment, city experiences (like interactivity in museums and cultural institutes, educational activities etc), municipal services and a school channel. One may use the chain ON/OFF model, which is illustrated in Fig. 3 as process of *Type* A, in order to describe such use cases as interactive guided tours in museums and other cultural institutes and the chain ON/OFF mo-del illustrated as process of *Type* B in order to describe tele-monitoring sessions and several municipal services. Streaming video for real-time situational awareness, *Push-To-Talk* municipal services and file deli-very for non-real-time ser-vices such as all-points bulletin (e.g. amber alert), soft-ware updates /upgrades etc as well as a

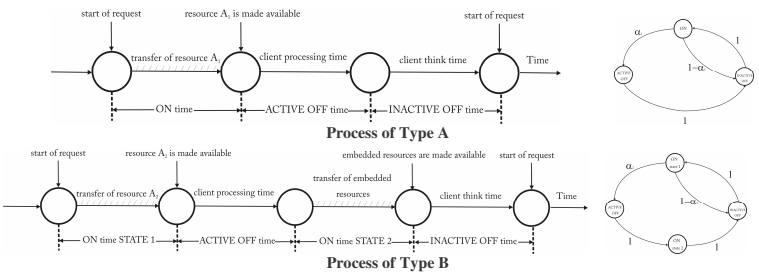

Fig. **3**. Chain ON/OFF models used in the numerical simulations

regional school channel may be implemented as broadcast services in the context of eMBMS deployment. Accessing a set of WWW resources (like HTML pages with embedded images) is modelled as a process of *Type* B in [24,25].

6 Numerical Simulations Using ON/OFF Chain Models

Not all content is cacheable. Personal data such as digital medical images and examinations as well as patients' records in the context of a telemedicine session are in general not cacheable. This holds true for personal data that may be transmitted in the context of a distance learning broadcasting. Overall 19.80-32.20% of unique URLs are uncacheable as estimated in [26]. This accounts for 21.50-28.32% of the total requests and 10.48-14.81% of the total bytes. HTML and JavaScript are dynamically generated and, thus, less cacheable. Analysis of real data in [26] indicates that, finally (after a certain buffer size) the slope of *byte_hit_rate*, which is denoted as $\lambda_{service}(buffer\_size)$, decreases as the size of the buffer increases. This leads to the following straightforward proposition regarding the optimal buffer size,

Proposition

The maximum value of the total (*byte_hit_rate* x *users* x *reguests*) for a constraint size of a buffer used for caching requests for a mixture of *M* services is given as the sum of the buffer sizes allocated to each service for which the following condition holds,

$$\frac{\partial\,(byte\_hit\_rate \text{ x } users \text{ x } requests)}{\partial\, buffer\_size}\Big|_{per\ service} = \lambda_{service}(buffer\_size) = \lambda_{opt} \qquad (3)$$

for all of the *M* services.

A mixture of two services modeled as $\{0.67\ Process\ Type\ A, 0.33\ Process\ Type\ B\}$ is simulated. Three distinctive reference distributions are associated with *Process Type* B, one with state ON1 and two with state ON2. It is assumed that the sizes of the embedded objects are equal for each distribution. The selection of parameters of chain ON/OFF models is based upon actual measurements for web traffic caching like those presented in [26] and results for modelling data emanating from medical instruments in clinical environment ([27]). A 10 Gbps link is assumed to connect cached content at the gateway to the backbone network. The gateway is assumed to provide services for 500 concurrent users. Active ON times for non-cached objects are modeled according to Pareto distribution [21] with mean equal to *Object_size/User_bandwidth.* The delay for all cached objects is assumed to equal 200 ms. *Active* OFF or client processing delay is modeled according to the Weibull distribution [20]. It ranges from 1 msec to 1 sec. *Inactive* OFF times follow the Pareto distribution. All parameters are summarized in **Table** 1. Simulation results are illustrated in Fig. 4. Both hit rate and byte hit rates are given. Optimal buffer size is estimated as $s_1(\lambda') + s_2(\lambda')$. Total *byte hit requests* (*byte hit rate* x *users* x *requests per user*) are added for the same slope. The optimal slope decreases for higher budgets of buffer storage. Hit rates are up to 30%. Changing the mixture of the two services results in a different optimal value for the buffer size.

	Process of type A	*Process of type B*		
Transitional probability ON state	0.25	0.6		
ON (Pareto)	shape=1.05 and scale=0.2	ON1:shape=4.5 -scale=0.2 (object 1e5) ON2:shape=1.5 -scale=0.2 (object 1e6) ON2:shape=1.05-scale=0.2 (object 1e7)		
ACTIVE_OFF (Weibull)	shape=3.0 and scale=0.045	shape=3.0 scale=0.045		
INACTIVE_OFF (Pareto)	shape=1.5 and scale=2.0	shape=1.5 scale=2.0		
Reference distribution(s) (**Zipf parameters**)	z=0.85 bytes per object 1e7	z=0.8 object bytes 1e5	z=0.9 object bytes 1e6	z=0.95 object bytes 1e7

Table 1. Parameters of the chain ON/OFF models according to Fig. 3

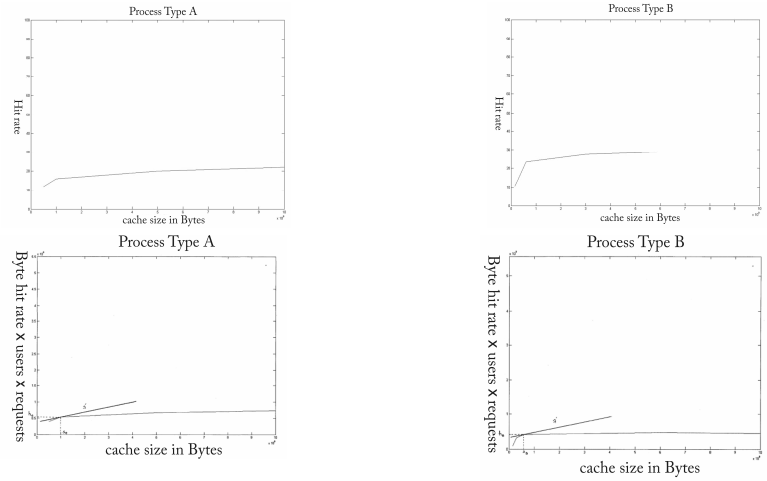

Fig. **4**. *Hit rate* (upper row) & *byte hit rate* (lower row) per service (left for service modeled as *Type* A, right for service modeled as *Type* B). Buffer sizes indicate optimal allocation.

7 Conclusion

LTE eMBMS takes advantage of single frequency network features and allows for flexible content delivery via unicast, multicast and broadband services. Live tests have been conducted by several companies. There is an increasing interest in launching commercial deployments and offering localized services through enhanced infrastructure. Service delivery and user experience may be boosted by web caching and acceleration by storing frequently accessed content at the edge of the network. A dynamic dimensioning scheme for optimal cache allocation is herein proposed. It requires traffic modeling of the services offered to the users. It assumes that the frequently accessed content is stored at the eMBMS gateway, which receives unicast streams and delivers MBMS user data packets to eNBs by means of IP multicast. Future work will emphasize upon dynamic buffer dimensioning according to the proposed approach taking into consideration such novel cache-accelerated techniques as request multiplexing and multi-resolution chunking. The proposed optimality condition may be applied in conjunction with such novel approaches as well. Simulation results are consistent with actual traffic measurements that are reported in the literature [27].

Acknowledgments. This research work has been funded by the *LiveCity* European Research Project supported by the Commission of the European Communities – *Information Society and Media Directorate General* (FP7-ICT-PSP, Grant Agreement No. 297291).

References

1. 3GPP TS 36.300: Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2.
2. 3GPP TR 25.905: Improvement of the Multimedia Broadcast Multicast Service (MBMS) in UTRAN.
3. Alexiou, A., Asimakis, K., Bouras, C., Kokkinos, V., Papazois, A., Tseliou, G.: Cost optimization of MBSFN and PTM transmissions for reliable multicasting in LTE networks. *Wireless Networks* 11, vol. 18 (issue 3), 277-293 (2011).
4. Hartung, F., Horn, U., Huschke, J., Kampmann, M., Lohmar, T., Lundevall, M.: Delivery of Broadcast Services in 3G Networks. IEEE *Trans. on Broadcasting*, vol. 53, issue 1, 188-199 (2007)
5. Hartung, F., Horn, U., Huschke, J., Kampmann, M., Lohmar, T.: MBMS – IP Multicast/Broadcast in 3G Networks. *Int. J. Digital Multimedia Broadcasting*, ISSN 1687-7578 (2009).
6. Nguyen, N-D, Knopp, R., Nikaein, N., Bonnet, C.: Implementation and Validation of Multimedia Broadcast Multicast Service for LTE/LTE-Advanced in OpenAirInterface Platform. EURECOM, October 2013.
7. Wang, X, Wang, Y., Zhang, Y.: A Novel Transmission Policy for Reliable eMBMS Download Delivery. *Wireless Communications and Networking Conference* (WCNC), 1-6, 18-21 April 2010.
8. Cohen, E., Kaplan, H.: Prefetching the Means for Document Transfer: A New Approach for Reducing Web Latency. 19th Annual Joint Conference of the IEEE Computer and Communications Societies, IEEE Proceedings, vol. 2, pp. 854-863, INFOCOM 2000.
9. Adas, A.: Traffic Models in Broadband Networks. IEEE *Communications Magazine*, 82-89, July 1997.
10. Abrahamsson, H.: Traffic measurement and analysis. *Swedish Institute of Computer Science* (1999).
11. Williamsson, C.: Internet traffic measurement. IEEE *internet computing*, vol. 5, no. 6, 70-74 (2001).
12. Celeda, P.: High-speed network traffic acquisition for agent systems. Proc. IEEE/WIC/ACM *International Conference on High Speed Network Traffic Acquisition for Agent Systems*, Intelligent Agent Technology, November 2-5, 477-480 (2007).
13. Pezaros, D.: Network Traffic Measurement for the next Generation Internet. Computing Department Lancaster University (2005).
14. Lecompte, D. and Gabin, F.: Evolved Multimedia Broadcast/Multicast Service (eMBMS) in LTE-Advanced: Overview and Rel.-11 Enhancements. IEEE *Communications Magazine*, 68-74, November 2012.
15. 3GPP TS 26.247 Release 10: Transparent end-to-end packet-switched streaming Service (PSS); Progressive download and dynamic adaptive Streaming over HTTP (3GP-DASH),"http://www.3gpp.org/ftp/Specs/html-info/26247.htm.
16. Lohmar, T., Slssingar, M., Puustinen, S., Kenehan, V.: Delivering content with LTE Broadcast. *Ericsson Review*, 2-7, February 2013.
17. Lohmar, T., Ibanez, J.-A., Blockstrand, M., Zanin, A.: Scalable Push File Delivery with MBMS. *Ericsson Review*, vol. 1, 12-16 (2009).
18. ETSI TS 102 589: Forward Link Onlv Air Interface; Specification for Terrestrial Mobile; Multimedia Multicast, V1.1.1 (2009-02).
19. ETSI EN 303 105.
20. http://en.wikipedia.org/wiki/Weibull_distribution
21. http://en.wikipedia.org/wiki/Pareto_distribution
22. Glassman, S.: A Caching Relay for the World Wide Web. *Computer Networks and* ISDN *Systems*, vol. 27, no. 2 (1994).
23. Cunha, C. et al.: Characteristics of WWW Client-based Traces. *Technical report* BU-CS-95-010, Computer Science Department, Boston University, July 1995.
24. Hadjiefthymiades, S., Merakos, L.: Using Proxy Cache Relocation to Accelerate Web Browsing in Wireless/Mobile Communications. Proc. 10th international conference on World Wide Web, 2536 (2001).
25. Barford P., Crovella, M.: Generating Representative Web Workloads for Network and Server Performance Evaluation. *Proceedings of* ACM *Sigmetrics*, July 1998.
26. Sunghwan, Ihm: Understanding and Improving Modern Web Traffic Caching. Ph. D. Thesis, Department of Computer Science, Princeton University, September 2011.
27. Ahmad, A., Riedl, A., Naramore, W. J., Chou, N-Y.: Scenario-Based Traffic Modeling for Data Emanating from Medical Instruments in Clinical Environment. World Congress on Computer Science and Information Engineering, ISBN 978-0-7695-3507-4, 529-533 (2009).