Resource Usage Prediction for Optimal and Balanced Provision of Multimedia Services

Yiannos Kryftis, Constandinos X. Mavromoustakis
Department of Computer Science
University of Nicosia
Nicosia, Cyprus
kryftis.y@unic.ac.cy, mavromoustakis.c@unic.ac.cy

Jordi Mongay Batalla
National Institute of Telecommunications
Szachowa Str. 1,
04-894 Warsaw, Poland
jordim@interfree.it

George Mastorakis, Evangelos Pallis
Department of Informatics Engineering
Technological Educational Institute of Crete
Heraklion, Crete, Greece
gmastorakis@staff.teicrete.gr, pallis@pasiphae.eu

Georgios Skourletopoulos
Department of e-Banking Insights
Scientia Consulting S.A.
Athens, Greece
g.skourletopoulos@scientiaconsulting.eu

Abstract—**This paper proposes a novel network architecture for optimal and balanced provision of multimedia services, exploiting a resource prediction system. This architecture enables for the long-term prediction of multimedia services future demands, based on the history of previous network resources usage. The proposed research approach provides the opportunity for the optimal distribution of streaming data, among Content Delivery Networks, cloud-based providers and Home Media Gateways. The short-term prediction that is performed, enables for making the proper decisions by the system, according to specific network metrics, towards achieving higher Quality of Service and Quality of Experience for the end users. The validity of the proposed system is verified through several sets of extended experimental simulation tests, carried out under controlled simulation conditions.**

Keywords: Resource Prediction Engine, Content Delivery Networks, Multimedia Services Systems, Quality of Experience

I. INTRODUCTION

The tremendous evolution of multimedia-related technologies over the Internet, as well as the increasing demand for an efficient, unified and secure media distribution solution, push more pressure for further research and development on the field of multimedia distribution content. Multimedia services distribution generates today a significant part of the global Internet traffic, while the amount of this traffic is expected to double in 2015, compared to 2012 [1], reaching more than 30 PB/month out of an overall traffic of 50 PB/month. On the other hand, cloud computing has emerged as a new paradigm for hosting and delivering services over the Internet. It was initially exploited for the realization of resources sharing network infrastructures to support computing costly applications. Cloud-based solutions have now been extended to support multimedia distribution systems, acting as a virtualized complete serving infrastructure to large number of end users. Towards achieving high network performance, the today's cloud-based systems are based on data centers infrastructures that are located in multiple sites. This way, a Service Provider can leverage geo-diversity to achieve multimedia services delivery to the final users. In this framework, open and challenging problems still remain un-solved, such as issues regarding the quality of multimedia services delivery, based on end user's expectations.

Tackling such challenges, this paper goes beyond the current state-of-the-art, elaborating on a new multimedia services delivery solution that is based on the optimum allocation of the resources used for content transmission to efficiently satisfy different users' requests, through the exploitation of existing servers' infrastructures capabilities. Such capabilities are available in conventional clouds (i.e. public or private computing infrastructure configurations, usually offered by over-the-Top providers), in Content Delivery Networks (CDNs) and in Home Media Gateway Clouds (i.e. Home Gateways/Community Gateways configurations, exploited in peer-to-peer mode). The proposed approach foresees a new business model in multimedia services delivery, strongly but smoothly leveraging (in an evolutionary way) new mechanisms and systems. Following this introductory section, section II presents related work approaches, as well as the research motivation of this paper. Section III elaborates on the proposed research approach based on a novel network architecture and a resource prediction engine for optimal multimedia services provision. Finally, section IV provides the performance evaluation results and section V concludes the paper, by indicating future research.

II. RELATED WORK AND RESEARCH MOTIVATION

Towards providing to the users the desired Quality of Experience (QoE) during the multimedia services provision process, there is a need of implementing a resource usage prediction engine that provides the ability to predict future demands of the network resources. This gives the opportunity through a management plane to trigger the proper actions for keeping the desired quality for the streaming sessions. The resource prediction engine has to be developed based on novel models, able to efficiently predict and plan the needed bandwidth capacity based on bandwidth auto-scaling functions, automatically accommodating the fluctuations of the network resources for a certain multimedia event provision. In this direction, Niu et al. [2] presented some of the issues of demand forecast and performance prediction in peer-assisted Video-on-Demand (VoD) services. They use the Box-Jenkins approach [3] to predict the future population of each video channel based on a given time series in the past. They avoid periodicity, by using regression methods, as well as the seasonal ARIMA (autoregressive integrated moving average) model [3]. They infer the initial population of a new released video channel, using machine learning techniques. They also utilize pass data from newly released video channels as training data to make a prediction, based on statistical models and of the channels release time. The ARMA (autoregressive moving-average) model [3] was used to predict the server bandwidth demands by a video channel at future time. In [4], Niu et al. proposed a predictive cloud bandwidth auto-scaling system for VoD providers. Based on the history of bandwidth demand in each video channel, derived by cloud monitoring services, it estimates the expectations for near future demands. By doing so, it provides quality assurance, by deciding the minimum bandwidth reservation to satisfy the demand.

Wu and Lui [5] presented a system architecture and model for optimization of replication strategy in P2P-VoD systems. They showed that conventional proportional replication strategy is not optimal, by proposing a passive replacement strategy for the decision of the content that should be deleted when a local storage is full. They also proposed an active replication algorithm to aggressively push data to peers, in order to achieve the desired replication ratios. A similar approach to the optimal content placement for a large-scale VoD system is presented in [6]. RPS [7], [8] is a publicly available toolkit that allows the creation of online and offline resource prediction systems. It includes its own monitoring facilities but it also provides the ability to use monitoring data that comes from other sources. Part of the system is the wavelet toolkit [9] that supports analysis of the signals. In [10], the Push-to-Peer approach was presented that proactively pushes content to peers, to increase content availability, improving the use of peer uplink bandwidth. It allows performance analysis of the push policies, distributed load balancing strategies for the initial selection of serving peers and distributed strategies to cope with dynamic uplink bandwidth. A different approach is presented in [11], [12], [13], [14] with special emphasis in the energy consumption, and efficient sharing of resources. Such research approach indicates that although some work has been done and the related tools where developed, there is a lack of a system that combines the ability to predict future demands, by taking advantage of that prediction to automatically accommodate the fluctuations of the network resources for the optimal provision of the desired Quality of Service (QoS) and QoE to the end users. In this context, this paper proposes a network architecture that predicts the future content delivery demands and the future network usage, performing all the necessary adaptations to deliver the content in an optimal approach.

III. Resource Prediction Engine for Optimal Multimedia Services Provision

Towards introducing an efficient Resource Prediction Engine for optimal multimedia services provision over the future Internet architectures, a collaboration between different actors involved in the multimedia delivery process (network operators, service providers and end users) is required. Therefore, a novel network architecture is proposed in Fig. 1 with an upper layer (called DELTA M&C) that coordinates such a collaboration environment.

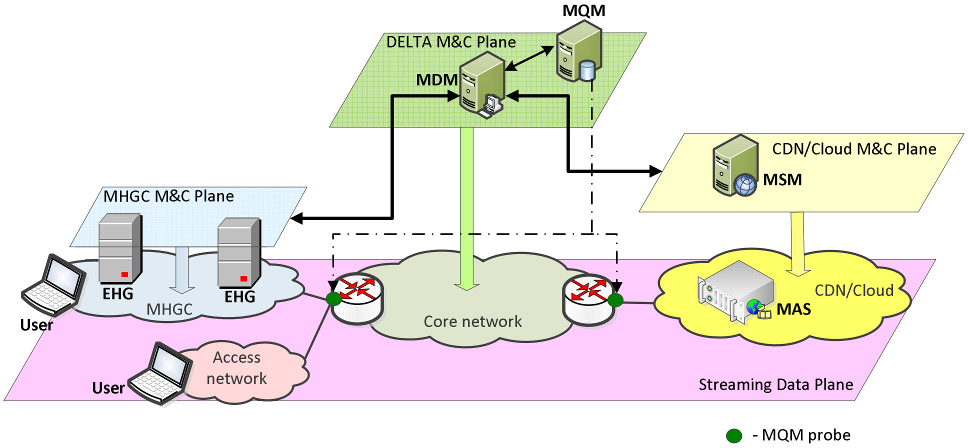

Fig. 1. Proposed Network Architecture

The entities of this plane interchange information with the current management and control (M&C) planes of the CDN and Cloud providers, as well as with the so-called Media Home Gateway Cloud (MHGC) provider, which basically is an ecosystem of users gateways (e.g. set-top-boxes with advanced functionalities) with a distributed management (sometimes the management may be centralized in the case when the home gateways are owned by the network operator). The proposed network architecture consists of the following conceptual entities: a Media Distribution Middleware (MDM), a Media QoE Meter (MQM), a Media Services Manager (MSM), an Enhanced Home Gateway (EHG) and a Media Advanced Streamer (MAS). These components cooperate together, creating different Management and Control (M&C) planes.

The MDM is the main component of the M&C plane. It executes all necessary operations and determines all data required for optimal allocation of the available resources at each Resource Provider's domain. As a result, the MDM returns guidelines, which resources should be used for handling given user's request, to achieve the best (in terms of efficiency) resource exploitation. Another component of the M&C plane is the MQM that is responsible for continuous monitoring of network metrics at the user's and the Service Provider's domain access points, as well as the user's context and preferences. Based on the data gathered by the set of the MQM probes, distributed all over the domain, this entity provides to the MDM the related data about the current network conditions and the estimated value of QoE available for a user. Moreover, the MQM sends alerts to the MDM, only if any of the monitored QoS/QoE parameters declines below the allowed level. In addition, the functionalities of the CDN/Cloud M&C plane are realized by the MSM entity, which manages all Service Provider's resources. The MSM receives content requests generated by the users, and next, according to recommendations received from the MDM, takes a decision, on which server should stream the requested media and with which bitrate. In this way, the MSM, contrary to existing solutions, performs adaptation decision, taking into account not only the available bandwidth, but also considering other important information addressed by the MDM, such as the estimated QoE value and the prediction of potential upcoming streaming sessions. Although conceptually, the MSM is one entity, which manages the whole Service Provider domain, its functionality can be distributed between several physical machines (with different streaming protocols), but interconnected to form one coherent component.

The EHG entity is placed at the user's premises. In this way the control modules of EHGs realize functionalities of the Media Home Gateway Cloud (MHGC) M&C plane – they are responsible for creating the MHGC ad-hoc system from a set of peer-to-peer connected EHGs. Similarly to the MSM, the EHG receives content requests from the users and then asks the MDM about information which MHGC peers should be involved, in order to efficiently deliver requested content. EHG collaborates with MSM entities to obtain media content requested by the user, if this content (or part of it) is not stored on any of EHGs, belonging to given MHGC. In addition, the MAS entity resides in the CDN/Cloud domain as a standalone component, whereas in case of the MHGC, its functionalities are provided by the EHG. Streaming process realized by the MAS is performed, in accordance with instructions received from the MSM/EHG entity. Finally, the MDM adopts a resource prediction engine, in order to be able to predict future demands for resources. The prediction is divided in a short-term and long-term prediction. The long-term prediction takes as input the demand for each resource in the past and it uses some statistical methods to predict future demands. This gives the chance to the system to make the optimal distribution of data in CDNs and EHGs based on the prediction before the actual need. More specifically, Fig. 2 presents an overview of the proposed resource prediction system used for the prediction of future demands for streaming of video channels.

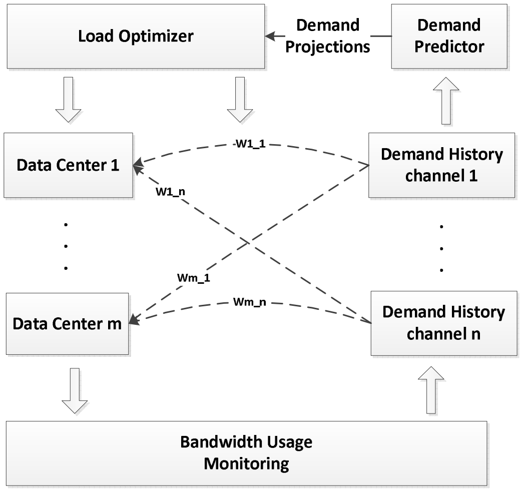

Fig. 2. Proposed predictive auto-scaling system

The demand predictor takes as input the demand history of each video channel and performs a prediction for the future demands. At the same time, the load optimizer uses as input the demand prediction to calculate the bandwidth reservation needed for each data center and for each streaming channel. It also calculates the load direction table $W=[w_{mn}]$, where w_{mn} represents the proportion of video channel n's requests, directed to data center m. The short-term prediction takes as input the data from the MQM component that performs real-time monitoring of the network status and of the content that is exchanged. The bandwidth and bandwidth utilization metrics combined with the content that is being delivered, are used as input to the prediction engine, using the RPS toolkit [7]. This toolkit is exploited to make a short-term prediction for the upcoming demands, in order to take the proper decisions for the delivery method. Fig. 3 presents an overview of the proposed online prediction system. The Monitoring Service provides a measurement stream, by sampling an attribute. This is used as an input to a Predictor, which presents the prediction as a vector of the next values of the measurement stream, each one associated with an estimated error. The Prediction Stream is the output of the predictor and applications can directly subscribe to the stream to use the data. The Buffer keeps a short history of the Prediction Stream, to allow asynchronous communication with applications when needed. The Evaluator is an optional component that constantly monitors the performance of the predictor, by comparing the predicted value with the real measured value and refits the model.

The MDM component uses the predicted future values for the metrics, to take the decisions for delivery of requested media, which may be streamed: 1) directly from the Cloud, 2) through deployed surrogate servers of the CDN, 3) by establishing a Media Home Gateway Cloud (MHGC) ad-hoc system and using a combined P2P-based technology of distribution, or 4) a combination of parts or all of them. The results are forwarded to the MSM component that is responsible for the actual streaming of the data to the user. Finally, Fig. 4 presents the internal architecture of the MDM component. The QoS/QoE Politics Traffic Data History component gathers the monitoring data that comes from the MQM, using them as input to the Media Traffic Forecast that generates the long-time prediction for the traffic in the network. The Resource allocator/scheduler uses the monitoring data for short-term prediction, feeding the MSM component with the optimal methods of content delivery. At the same time, the Bandwidth Allocation Optimizer calculates the optimal bandwidth allocation for the peer-to-peer delivery among MHGC devices. The optimizer is an online system that takes into consideration the network metrics, which come from the MQM, delivering that information to the MHGC devices.

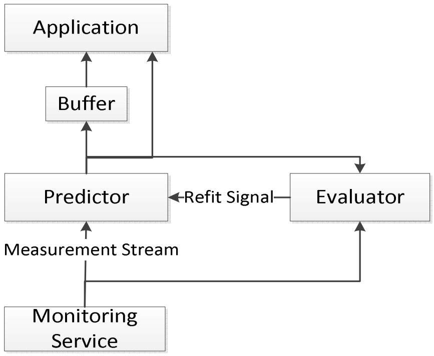

Fig. 3. Overview of the proposed online resource prediction system

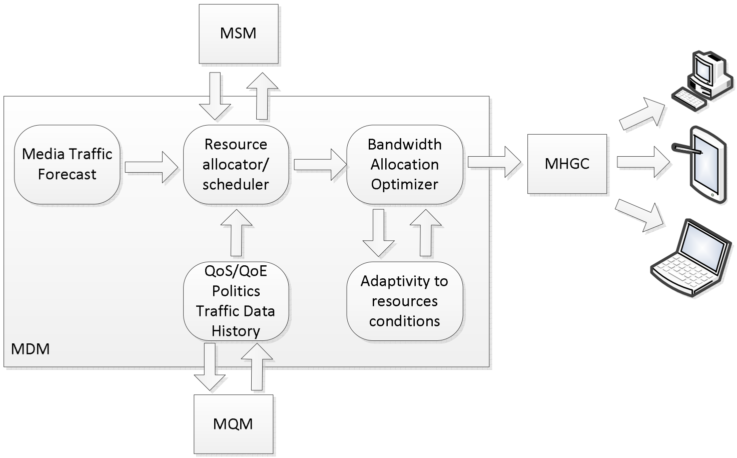
MSM
Media Traffic Forecast
Resource allocator/ scheduler
Bandwidth Allocation Optimizer
MHGC
QoS/QoE Politics Traffic Data History
Adaptivity to resources conditions
MDM
MQM

Fig. 4. Resource Prediction Engine Internal Architecture

IV. Performance Evaluation Analysis, Experimental Results And Discussion

In this section the effectiveness of the proposed system is demonstrated, by executing some simulations of usage scenarios. The experimental simulation results encompass the evaluation of the performance and the offered reliability in streaming activities, by the proposed system. Comparative performance evaluation results were extracted, to validate the reliability degree and the streaming ability of the proposed architecture, with respect to the performance efficiency. Towards implementing such scenario, a common look-up application service for video streaming is set in each node (both static and mobile), to enable nodes requesting a stream from a certain user. For simulating the proposed scenario, the varying weighted parameters (w_{nm}) described in the previous section were exploited, by using a two-dimensional network, consisting of nodes that vary between 10-100.

More specifically, Fig. 5 shows that the number of the participating nodes is increasing, when MDM-enhancing broadcasting is used, instead of a generic -or as called-staggered broadcasting [15]. This indicates the enhancement that has been done, by the MDM in the broadcasting process, whereas the Community Streaming factor W as introduced in [16], indicates the level of robustness in receiving neighboring feedback, during the process of streaming. W_n takes into account the streaming parameter of a single application S consisting of S_1, S_2,... S_n streaming delay bound, with n indicating the number of the possible intermediate nodes. W is the Community streaming factor [11], defined as the number of existing communities in the intercluster communication links over time. W can be defined, according to the download frequency of the file chunks in the intercommunity (cluster) as indicated in [11]. Fig. 6 shows the end-to-end delay time, which is shown to be reduced when the MDM mechanisms take place, in contrast to the random placement. In addition, the total delay time with the number of simultaneous transmissions is shown in Fig. 7. It is obvious that the total measured delay is significantly reduced in the presence of MHGC, whereas the utilization of the existing infrastructure increases the overall delays when multiple transmissions take place. Fig. 8 shows the respective Complementary Cumulative Distribution Function (CCDF) that represents the sharing reliability with the download time for requests up to 20 MB. It is true that by using the proposed approach in the presence of *Reyleigh fading* and mean noise of 4dB, the reliability is not importantly affected. The average throughput of the system is estimated as the average throughput of the total link utilization (1-P_{loss}). In this respect, the throughput with the number of requests per second is presented in Fig. 9, depicting that the fading channels with noise have low throughput exhibition especially, when the number of requests per second increases.

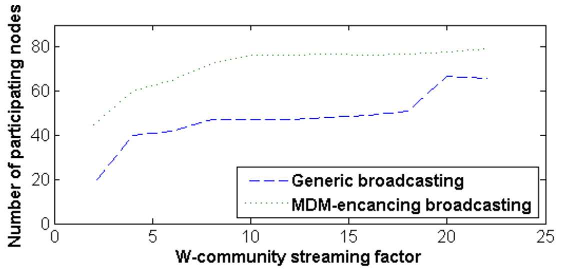

Fig. 5. Number of participating nodes with streaming factor [11]

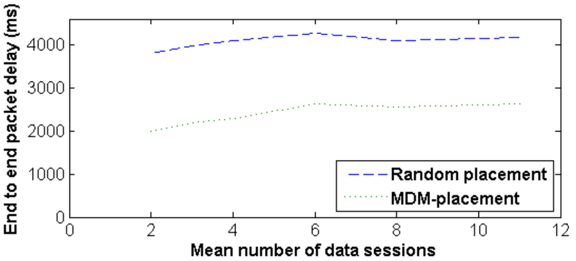

Fig. 6. End to end packet delay with number of data sessions

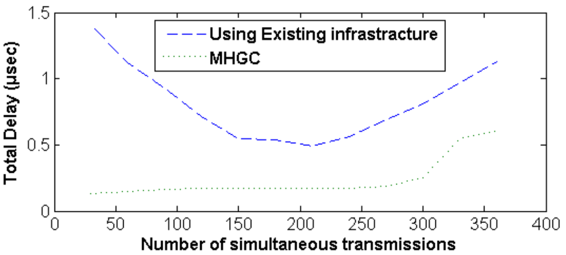

Fig. 7. Total delay (µsec) with number of simultaneous transmissions

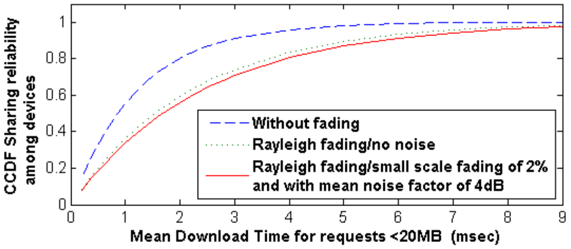

Fig. 8. CCDF sharing reliability among device with download time (msec) for requests <20 MB

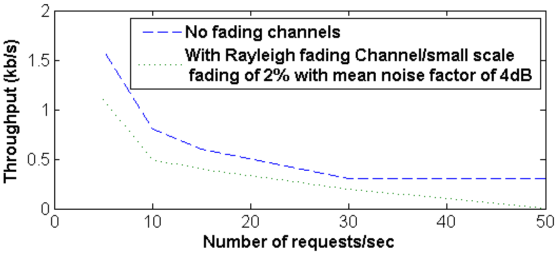

Fig.9. Throughput (kb/s) with number of requests per second

V. Conclusions

This paper proposes a novel network architecture for optimal and balanced provision of multimedia services, exploiting a resource prediction system. The proposed system utilizes resource prediction methods and algorithms in an evolutionary way for long-term and short-term predictions of the future resources demands for efficient multimedia services provision. This provides the opportunity for performing all the necessary adaptations to deliver the multimedia content in optimal and balanced way. The experimental results verified the efficiency of the proposed system and indicated fields for further research. More specifically, the proposed architecture will enable new video streaming models to be developed in the future, aiming to offer high throughput and reliability response for delay sensitive services provision.

Acknowledgment

The work presented in this paper is co-funded by the European Union, Eurostars Programme, under the project 8111, DELTA "Network-Aware Delivery Clouds for User Centric Media Events".

References

[1] CNI, CISCO. Consumer Internet Traffic 2012-2017. http://www.cisco.com/web/solutions/sp/vni/vni_forecast_highlights.

[2] D. Niu, H. Xu, B. Li, S. Zhao, "Quality-Assured Cloud Bandwidth Auto-Scaling for Video-on-Demand Applications", in proc. of IEEE INFOCOM, 2012.

[3] G. Box, G. Jenkins, G. Reinsel, "Time Series Analysis: Forecasting and Control", WILEY, 2008.

[4] D. Niu, Z. Liu, B. Li, S. Zhao, "Demand forecast and performance prediction in peer-assisted on-demand streaming systems", in proc. of IEEE INFOCOM, 2011.

[5] W. Wu and J. Lui, "Exploring the Optimal Replication Strategy in P2P-VoD Systems: Characterization and Evaluation", IEEE Transactions of Parallel and Distributed Systems, 2012, Vol. 23.

[6] D. Applegate, A. Archer, V. Gopalakrishnan, S. Lee, K. K. Ramakrishnan, "Optimal content placement for a large-scale VoD system", in proc. of CoNEXT 2010.

[7] RPS. http://www.cs.northwestern.edu/~RPS/. [Online]

[8] P. Dinda, "Design, Implementation, and Performance of an Extensible Toolkit for Resource Prediction in Distributed Systems", IEEE Transactions on Parallel and Distributed Systems, 2006, Vol. 17.

[9] J. Skicewicz, P. Dinda, "Tsunami: A Wavelet Toolkit for Distributed Systems", Technical Report NWU-CS-03-16, Department of Computer Science, Northwest University 2003.

[10] K. Suh et al, "Push-to-Peer Video-on-Demand system: design and evaluation", Thomson Technical Report, 2006.

[11] K. Papanikolaou, C. X. Mavromoustakis, G. Mastorakis, A. Bourdena, C. Dobre, "Energy Consumption Optimization using Social Interaction in the Mobile Cloud.", in MONAMI 2014.

[12] P. Mousicou, C. X. Mavromoustakis, G. Mastorakis, A. Bourdena, E. Pallis., "Performance Evaluation of Dynamic Cloud Resource Migration Based on Temporal and Capacity-Aware Policy for Efficient Resource Sharing", in proc. of HP-MOSys 2013.

[13] C. X. Mavromoustakis, E. Pallis, G. Mastorakis, "Resource management in mobile computing environments", Modeling and Optimization in Science and Technologies Series, Springer 2014.

[14] A. Bourdena, C. X. Mavromoustakis, G. Kormentzas, E. Pallis, G. Mastorakis, Muneer Bani Yassein, "A Resource Intensive Traffic-Aware Scheme using Energy-efficient Routing in Cognitive Radio Networks", Future Generation Computer Systems, Elsevier, 2014.

[15] J. B. Kwon and H. Y. Heom. "Providing vcr functionality in staggered video broadcasting", IEEE Consumer Electronics, Vol. 48, No 1, 2002.

[16] C. X. Mavromoustakis and H.D. Karatza, "Performance evaluation of opportunistic resource sharing scheme using socially-oriented outsourcing in wireless devices", The Computer Journal, Vol. 56, No. 2, 2013.