# Resource Usage Prediction Algorithms for Optimal Selection of Multimedia Content Delivery Methods

Yiannos Kryftis, Constandinos X. Mavromoustakis
Department of Computer Science
University of Nicosia
46 Makedonitissas Avenue, 1700
Nicosia, Cyprus
kryftis.y@unic.ac.cy,
mavromoustakis.c@unic.ac.cy

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

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

Joel J. P. C. Rodrigues
Department of Informatics
Instituto de Telecomunicações
University of Beira Interior, Portugal
and University ITMO, St. Petersburg, Russia
joeljr@ieee.org

Ciprian Dobre
Faculty of Autonomic Control and Computers
University Politehnica of Bucharest
313, Splaiul Independentei, 060042
Bucharest, Romania
ciprian.dobre@cs.pub.ro

Georgios Kormentzas
Department of Information and Communication Systems Engineering
University of the Aegean
Samos, Greece
gkorm@aegean.gr

***Abstract*—This paper proposes two algorithms adopted in a prototype network architecture, for optimal selection of multimedia content delivery methods, as well as balanced delivery load, by exploiting a novel resource prediction engine. The proposed architecture exploits both algorithms for the prediction of future multimedia services demands, by providing the ability to keep optimal the distribution of the streaming data, among Content Delivery Networks, cloud-based providers and Home Media Gateways. In addition, the prediction of the upcoming fluctuations of the network, provides the ability to the proposed network architecture, achieving optimized Quality of Service (QoS) and Quality of Experience (QoE) for the end users. Both algorithms were evaluated to establish their efficiency, towards effectively predicting future network traffic demands. The experimental results validated their performance and indicated fields for further research and experimentation.**

***Keywords: Resource Prediction Engine, Resource Usage Prediction Algorithms, Content Delivery Networks, Multimedia Services Systems***

## I. INTRODUCTION

The continuously increasing users' network activities and the escalating amount of information that they generate within the Internet have set the basis for a new area of convergence between networks and media, paving the way towards the Future Media Internet. Driving forces are the recent advances in connected media technologies and social networks, supported by the widespread deployment of broadband infrastructures and cloud computing facilities, along with a new breed of user-equipment that integrate communication and computation capabilities. All of them are rapidly transforming the environment(s) that citizens are surrounded by, as they introduce new kinds of interactions between humans and objects. Hence, in this evolving Future Media environment, it is normal for citizens to demand the kind of experiences that they are accustomed to in their daily/real lives. Towards conveying Media Events, following a user/community-centric approach with the maximum possible Quality of Experience (QoE), media delivery plays a key role. Media delivery can be defined as the assembly of communication protocols and mechanisms, through which they are delivered over a given combination of transport networks. Towards achieving the maximum efficiency in Future Media Internet contexts, media delivery calls for new architectures, along with associated technologies (i.e. functionalities and mechanisms), supporting their synergies to meet emerging requirements and increase the quality that citizens experience.

In this context, this paper proposes a novel network architecture and two resource prediction algorithms for multimedia services delivery, based on the optimum allocation of the resources used for audiovisual content transmission. The media delivery methods include existing servers' infrastructures capabilities, available in conventional, public and private clouds, in Content Delivery Networks (CDNs) and in Home Media Gateway Clouds (i.e. peer-to-peer networks between Home Gateways). The proposed system introduces several advances in media delivery, in order to achieve the maximum QoE, by maintaining the use of network resources controlled. For this, the system introduces new mechanisms and algorithms, which are responsible for the end-to-end delivery of media content with the required quality. 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 two algorithms, adopted in a resource prediction engine for optimal multimedia services provision. Finally, section IV provides the performance evaluation results and section V concludes the paper, by highlighting fields for future research.

## II. Related Work and Research Motivation

Several existing research attempts elaborate on the combination of different delivery methods, in order to achieve better QoE for the users. Xu et al. in [1] proposes a CDN-P2P hybrid architecture for cost-effective streaming media distribution that combines the advantages of using CDN for providing high QoE with the low cost of using P2P-based stream. Yin et al. in [2] presents the design and deployment of a Hybrid CDN-P2P System for Live Video Streaming, demonstrating the improvement in startup delay time and in stability. Current research approaches focus on how to benefit from the combination of the different delivery methods but they do not take consideration of handling each resource separately. In comparison to such approaches, this paper goes beyond the current state-of-the-art, by handling each resource (i.e. streaming channel), according to the prediction of the future demand for the resources, based on the current, as well as the predicted network metrics.

The resource prediction engine constitutes an important part of the proposed system, in order to offer the desired QoE to the end users during the multimedia services provision process. Its role is to provide the ability to efficiently predict the needed bandwidth capacity and the upcoming network fluctuations. The prediction engine has to be based on novel methods and models that can accurately forecast the future demands, in order to trigger through a management plane the proper actions for keeping the desired quality for the streaming sessions. Niu et al. in [3] presents a number of time-series analysis techniques to predict the server bandwidth demand and the peer upload for content delivery in peer-assisted Video-on-Demand (VoD) services. For the prediction of future population of each video channel, they use the Box-Jenkins approach [4] with input the population of the video channel in the past. The seasonal ARIMA (autoregressive integrated moving average) model [4] is exploited, for avoiding the periodicity. They use machine learning techniques for inferring the initial population of a new released channel, by utilizing pass data from newly released videos as training data. For the prediction of the server bandwidth demands by a video channel at future time, the ARMA (autoregressive moving-average) model was used [4]. They prove that the entire procedure has reasonable computation cost. Niu et al. in [5] presents a predictive bandwidth auto-scaling system for VoD providers in the Cloud. The near future demands expectations are estimated based on the history of the demands as monitored by the cloud monitoring services. This provides the opportunity to reserve the minimum bandwidth needed for satisfying the demand in the desired quality.

Wu and Lui in [6] present their model and system architecture for the optimization of replication strategy in P2P-VoD systems. For the decision of the content to be deleted when the local storage is full, they propose a passive replacement strategy and for achieving the desired replication ratios, they propose an active replication strategy that pushes data to the peers. The validation of the results is done with proper simulations. A similar approach is presented in [7], where the so-called Push-to-Peer proactively pushes the content to peers, to increase the availability of the multimedia services and improve the use of the peer uplink bandwidth. The proposed system allows performance analysis of the push policies, load balancing strategies for the initial selection of serving peers and strategies for dynamic uplink bandwidth. Applegate et al. in [8] presents an intelligent algorithm for optimal content placement for a large-scale VoD system. In an energy consumption based approach, Mavromoustakis et al. [9], [10], [11] faces the issue of optimization of the energy efficiency on mobile cloud applications, presenting a framework that utilizes resource offloading techniques, which are proven to optimize the energy consumption in the mobile devices. Although there is impressive research in multiple content delivery methods and combination of such methods, there is a lack of a system that combines the ability to predict future demands and automatically select the optimal delivery method for the optimal provision of the desired Quality of Service (QoS) and QoE to the end users. In this context, this paper elaborates on 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 and balanced way.

## III. Resource Prediction Engine for Optimal Multimedia Services Provision

The introduction of a Resource Prediction Engine, for the optimal multimedia services provision over the future Internet architecture, demands a novel network architecture with management components that cooperate during the multimedia delivery process. These components are located in the equipment of network operators, service providers and end users. The proposed network architecture is shown in Fig. 1. The upper layer (called DELTA M&C) coordinates the collaboration environment, by interchanging information with existing management and control (M&C) planes of the CDN and Cloud providers, as well as the distributed M&C plane of the Media Home Gateway Cloud (MHGC) provider, consisted of user gateways, forming a Peer-to-Peer (P2P) network.

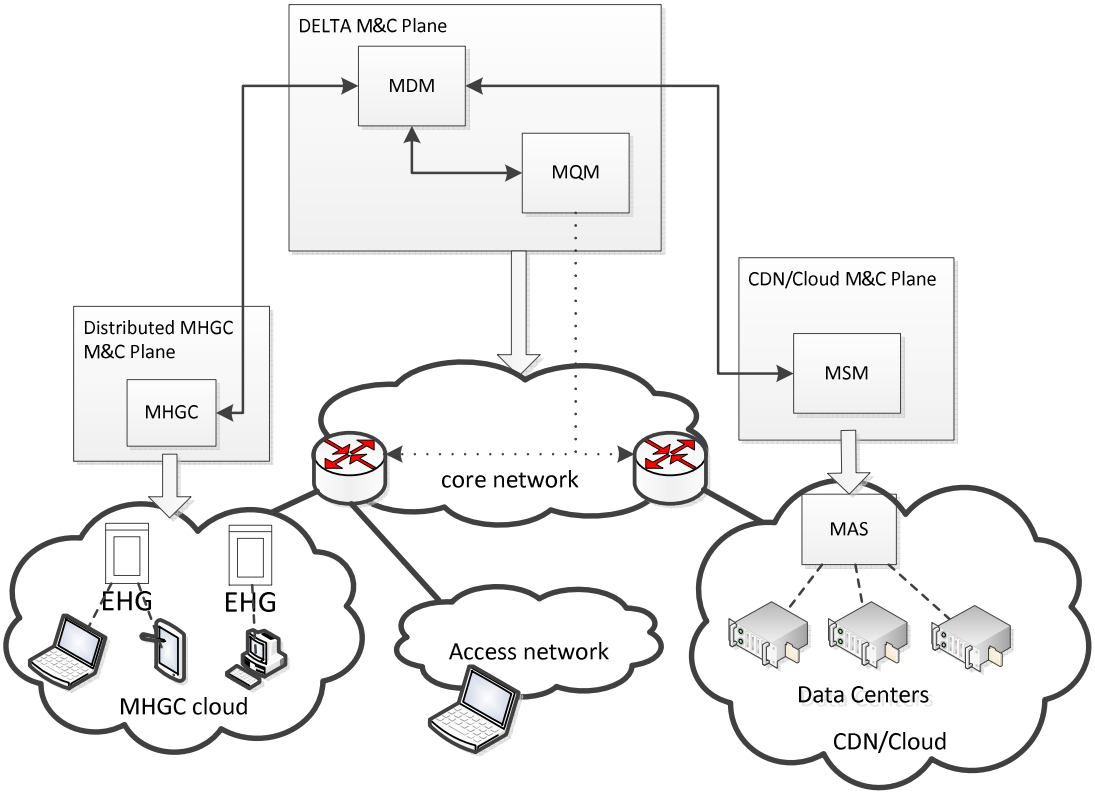


Fig. 1. Network Architecture for Optimal Multimedia Services Provision

The proposed network architecture consists of the following 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 with each other, as depicted in Fig. 1, creating different Management and Control (M&C) planes. The EHG entity is placed in the home equipment of each user. The control modules of EHGs constitute the Media Home Gateway Cloud (MHGC) M&C plane and they are responsible for creating the MHGC ad-hoc system from a set of peer-to-peer connected EHGs. Each EHG receives content requests from the users, requesting data from the MDM, about which MHGC peers should get involved to efficiently deliver requested content. EHG collaborates with MSM entities that reside in CDN/Cloud M&C planes and manage all Service Provider's resources, to obtain media content requested by the user, if the content is not stored on any of EHGs belonging to given MHGC. The MSM, according to the 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 the 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 the potential upcoming streaming sessions.

The MAS entity resides in the CDN/Cloud domain as a standalone component. Its role is to perform the streaming process, according to the instructions received from the MSM/EHG entity. MQM component is responsible for continuous monitoring of network metrics at the users and the Service Provider's domain access points, as well as the users' 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. 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. The MDM adopts a resource prediction engine, in order to be able to predict future demands for resources. The prediction is divided in long-term prediction for future demands for resources and short-term prediction for some important network metrics like throughput. The long-term prediction takes as input the demand for each resource in the past and it uses statistical methods to predict future demands. This provides the opportunity to the system to make the optimal distribution of data in CDNs and EHGs based on the prediction before the actual need. On the other hand, the short-term prediction is used for predicting and preventing upcoming network congestion issues, by triggering the proper actions. Fig. 2 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 and uses it as input to the Media Traffic Forecast, generating the prediction for the traffic in the network. The outcome of the forecast is used as an input to the Resource allocator/scheduler that takes the decision for the optimal delivery methods, feeding the MSM component with the recommendations, on which server should stream the requested media. At the same time, the Bandwidth Allocation Optimizer calculates the optimal bandwidth allocation for the P2P delivery between the MHGC devices. It is an online system, which takes into consideration the network metrics that come from the MQM, delivering that information directly to the MHGC devices. MHGC devices exchange management information between them and together they constitute a M&C plane that manages the P2P network between the EHG devices.

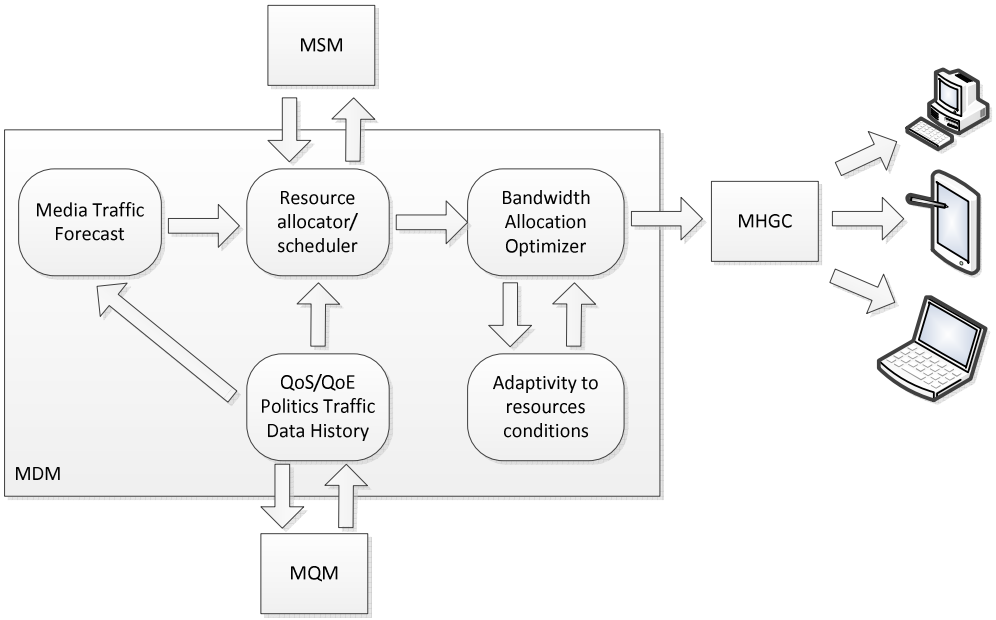


Fig. 2. Media Distribution Middleware Internal Architecture

Fig.3 presents the prediction engine that is implemented as a part of Media Traffic Forecast of the MDM. Input comes from the Monitoring Service and more specifically the MQM through the QoS/QoE Politics Traffic Data History component. The prediction engine is implemented in Java, using the JRI, Java Interface [12] for the interactions with the R-system [12]. R is a very popular free software environment for statistical computing and graphics. The standard stats package of R-system includes multiple time series models and prediction methods. The forecast package [14], [15] of R implements automatic forecasting with multiple methods, including ARIMA models, exponential smoothing methods, Theta method [16], cubic splines [17] and many others. Hyndman and Khandakar in [14] present the implementation of exponential smoothing methods and the ARIMA modelling approach in the forecast package. The proposed prediction engine uses the aforementioned packages, extending them in order to achieve optimal prediction. If a long-term (i.e. for the next hour) prediction has to be achieved for the needed bandwidth of a specific Video On Demand (VoD), the prediction engine needs to exploit the history of the bandwidth reservation for the specific VoD. The history data will be used to fit in the proper statistical model, suitable for the specific VoD and then the forecast function will be used to make the prediction. The result will be the estimated need of the bandwidth for the specific VoD after one hour. This value will be used by the Resource allocator/scheduler to decide how to serve the estimated future need.

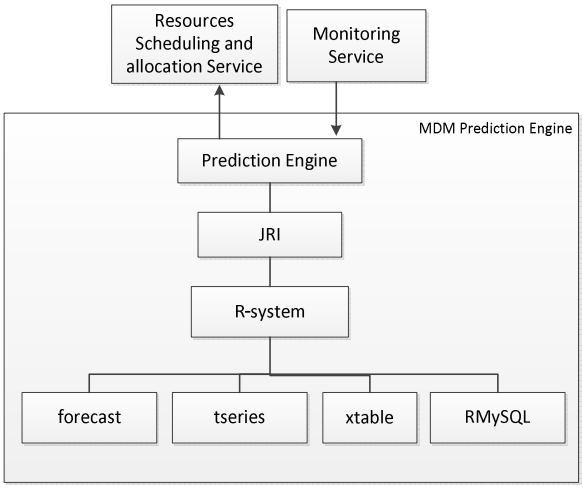


Fig. 3. Proposed Online Resource Prediction System

The MDM component uses the predicted future values for the metrics, in order 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 with multi-source, multi-destination congestion control algorithms, or 4) a combination of parts or all of them (thanks to stream-switching adaptation technique). The results are forwarded to the MSM component that is responsible for the actual streaming of the data to the user.

The mathematical formula used for the delivery methods is the following:

$$D = A * D_{cloud} + W * D_{CDN} + B * D_{MHGC} ,$$

where D=[$D_x$], x=1,..,n represents the combined delivery method for each of the n VoD channels, $D_{cloud}$=[$D_x$], x=1,..,n corresponds to the delivery over the cloud, $D_{CDN}$=[$D_x$], x=1,..,n to the delivery with use of CDN and $D_{MHGC}$=[$D_x$], x=1,..,n to the use of P2P network between EHG devices. In the formula W=[$w_{xy}$], x=1,..,n, y=1,..,m, where $w_{xy}$ represents the proportion of video channel x's requests directed to Data Center(DC) y. A=[$a_x$] and B=[$b_x$], x=1,…,n, where the variables can only take the values 0 and 1 to enable or disable the use of Cloud and P2P delivery respectively. In case of enabling P2P delivery, for a specific VoD, EHGs devices implement proper P2P algorithms to take the decisions about the load distribution. The corresponding algorithm as implemented in the MDM is presented below.

**Algorithm 1** Delivery Method Selection Algorithm

```
1: procedure SELECTCONTENTDELIVERYMETHODS
2:     neededBW ← maximum(currentBW,predictedBW)
3:     switch neededBW do
4:         case neededBW < lowLimit
5:             Use only Cloud.
6:         case lowLimit < neededBW < highLimit
7:             NumberOfDCstoUse ← ((neededBW−lowLimit)/(highLimit−lowLimit)) * AvailableDCs
8:             Use Cloud and NumberOfDCstoUse DCs.
9:             Run load balancing algorithm for DCs.
10:        case neededBW > highLimit
11:            Use Cloud, all available CDNs and P2P.
12:            Run load balancing algorithm for DCs.
```

It can be observed that the higher value can be taken among the current and the predicted bandwidth need. If it is below a preset limit (based on administrative high level decisions and network status), only the Cloud delivery will be used. If it is above the limit, CDNs will be used and the number of CDNs used is increasing according the demand. After the top limit is reached, a P2P delivery method is exploited. This algorithm provides the advantage that the data is not distributed before the actual need. In the case of a VoD with low customers demand, the CDNs will not been used for its distribution. On the other hand, if a popular VoD is requested, an early prediction will occur, while the number of CDNs, distributing the VoD will rapidly increase based on the demand. Finally, the P2P delivery method will be used, only when needed, while at the time that this happens, the number of the users already possessing the specific video will be satisfactory with those users, acting as seeders to distribute the VoD to the others.

The special algorithm for the creation of the weight matrix is presented above. This algorithm divides number 1 (the whole percentage) to the number of Data Centers (DCs), which will be used, to calculate the portion of delivery requests that each DC should handle. Then, it finds out which column represents the specific VoD or assigns a new column for a new VoD. Finally, it selects which DCs will be used, starting with those that have more zero values in their row, meaning that they do not serve many VoD channels.

**Algorithm 2** Load Balancing Algorithm

```
1: procedure LOADBALANCING
2:     portion ← 1 / NoOfDCsToUse
3:     if VoD=new then
4:         Add a column to the W[x,y] table
5:     column = Column that represents the current VoD
6:     Create a sorted table with the rows,
   based on the number of 0 they include
7:     W[x,column]=portion,
   where x takes the first NoOfDCsToUse values of the sorted table
```

## IV. Performance Evaluation Analysis, Experimental Results And Discussion

This section demonstrates the effectiveness of the resource prediction engine as a subsystem, by presenting experimental results, in terms of the outcome of the engine compared to the actual values measured, as well as the effectiveness of the whole proposed system, by executing simulations of the usage scenarios. For the evaluation of the forecast algorithms, it the monitoring data of the bandwidth usage for serving the need of a specific VoD, was utilized. The collected measurements were for a total of 30 minutes with a period of 5 seconds, but to avoid periodicity of data, the mean value per minute was used. The 80% (24 minutes) of the data was exploited to feed the prediction engine, while the rest of the data was used for the evaluation, through a comparison between the predicted and the actual value as shown in Fig. 4. The important part of the graph is after the 24 first minutes, where it is clearly depicted that the measured values remain very close to the predicted ones.

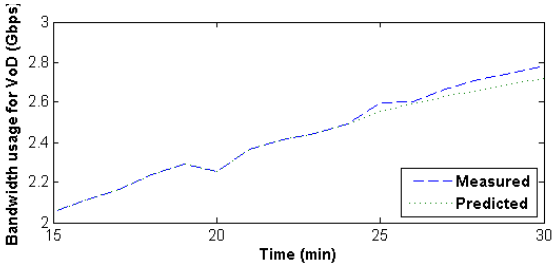


Fig. 4. Measured vs Predicted value for the bandwidth usage for VoD

The prediction performance of the engine for longer term is presented in Fig. 5. It includes the predicted value showing also the limit of 95% confidence and the corresponding (after the time passes) measured values of bandwidth needs for a VoD channel. The test scenarios presented are for 5, 30 and 60 min prediction. It is clear that the predicted values are near the actual values measured and in all cases the upper and lower limits of the 95% confidence interval include the measured value.

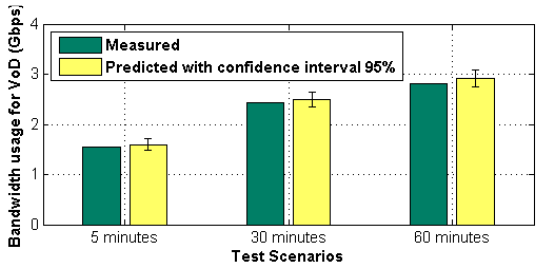


Fig. 5. Test scenarios for bandwidth usage of VoD prediction

The rest of the section presents experimental simulation results for evaluation of the performance and the offered reliability in streaming activities, offered by the proposed system. More specifically, Fig. 6 shows that the number of the participating nodes is increasing, when MDM-enhancing broadcasting is used, instead of a generic broadcasting. This indicates the enhancement that has been done by the MDM in the broadcasting process, whereas the Community Streaming factor *W,* as introduced in [18], indicates the level of robustness in receiving neighboring feedback during the process of streaming. The total delay time with the number of simultaneous transmissions is shown in Fig. 7. The total measured delay is significantly reduced in the presence of MHGC (P2P delivery), whereas the utilization of the existing infrastructure increases the overall delays when multiple transmissions take place.

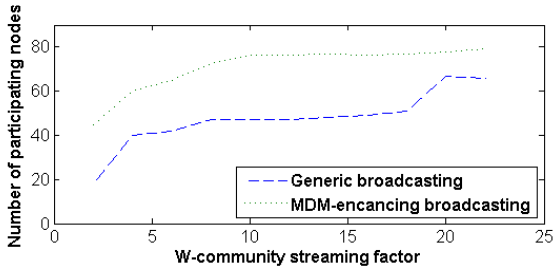


Fig. 6. Number of participating nodes with community streaming factor

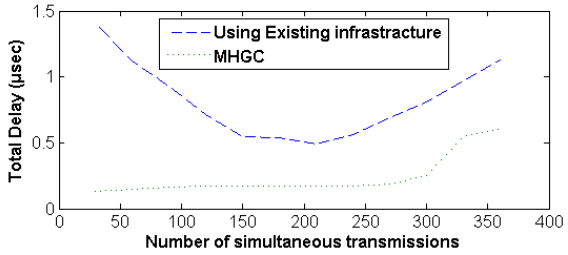


Fig. 7. Total delay (μsec) with number of simultaneous transmissions

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 framework in the presence of *Reyleigh fading* and mean noise of 4dB, the reliability is not importantly affected. Finally, 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.

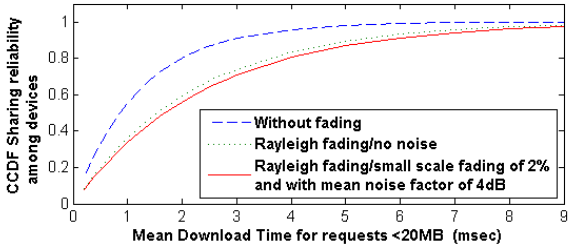


Fig. 8. CCDF sharing reliability among device with download time (msec) for requests <20 MB

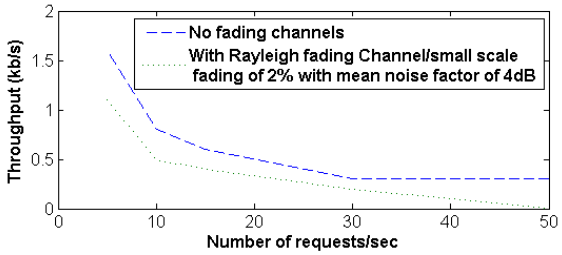


Fig.9. Throughput (kb/s) with number of requests per second

## V. CONCLUSIONS

This paper presents a novel network architecture and two algorithms for optimal selection of multimedia content delivery methods. It also presents a novel resource prediction engine utilized for balanced delivery load. The proposed system based on a resource prediction engine performs all necessary adaptations to deliver the content in an optimal and balanced way, in order to keep a high QoS and QoE for the end users. The experimental results prove that the prediction engine is accurate and that the whole system improves the whole delivery process characteristics. Future directions in our on-going research encompass the development of a load balancing algorithm that combines the delivery method and the load balancing under one scheme or multiple schemes that not only use the predicted values but the error estimations of the predictions as well.

## 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". Additionally, this work is partially supported by Instituto de Telecomunicações, Next Generation Networks and Applications Group (NetGNA), Portugal, by National Funding from the FCT - /Fundação para a Ciência e a Tecnologia through the PEst-OE/EEI/LA0008/2013 Project and by Government of Russian Federation, Grant 074-U01.

## REFERENCES

[1] D. Xu, S. S. Kulkami, C. Rosenberg, H. Chai, "Analysis of a CDN–P2P hybrid architecture for cost-effective streaming media distribution", Multimedia Systems 11(4), 2006, pp.383-399.

[2] H. Yin, X. Liu, T. Zhan, V. Sekar, F. Qiu, C. Lin, H. Zhang, B. Li, "Design and Deployment of a Hybrid CDN-P2P System for Live Video Streaming: Experiences with LiveSky" in proc. of ACM international conference on Multimedia, 2009, pp.25-34.

[3] 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, pp. 460-468.

[4] G. Box, G. Jenkins, G. Reinsel, "Time Series Analysis: Forecasting and Control", John Wiley & Sons, 2013.

[5] 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, pp.421-425.

[6] W. Wu and J. Lui, "Exploring the Optimal Replication Strategy in P2P-VoD Systems: Characterization and Evaluation", IEEE Transactions on Parallel and Distributed Systems, 23(8), 2012.

[7] K. Suh, C. Diot, J. Kurose, L. Massoulie, C. Neumann, D. Towsley, M. Varvello, "Push-to-Peer Video-on-Demand system: design and evaluation", IEEE Journal on Selected Areas in Communications, 25(9), 2007, pp. 1706-1716.

[8] R. J. Hyndman, M. L. King, I. Pitrun, B. Billah, "Local linear forecasts using cubic smoothing splines." Australian & New Zealand Journal of Statistics 47(1), 2005, pp. 87-99.

[9] C. X. Mavromoustakis, G. Mastorakis, A. Bourdena, E. Pallis, G. Kormentzas, J. J. P. C. Rodrigues, "Context-oriented Opportunistic Cloud Offload Processing for Energy Conservation in Wireless Devices", to appear in IEEE Globecom 2014, CCSNA Workshop, Texas, USA, 2014.

[10] C. X. Mavromoustakis, A. Andreou, G. Mastorakis, A. Bourdena, J. M. Batalla and C. Dobre, "On the Performance Evaluation of a Novel Offloading-based Energy Conservation Mechanism for Wireless Devices" in proc. of 6th International Conference on Mobile Networks and Management 2014, MON-AMI 2014, Wuerzburg, Germany, 22-24 September, 2014.

[11] K. Papanikolaou, C. X. Mavromoustakis, G. Mastorakis, A. Bourdena, C. Dobre, "Energy Consumption Optimization using Social Interaction in the Mobile Cloud", in proc. of 6th International Conference on Mobile Networks and Management 2014, MON-AMI 2014, ELEMENT Workshop, Wuerzburg, Germany, 22-24 September, 2014.

[12] http://rforge.net/rJava/index.html. [Online], last accessed June 2014.

[13] R. C. Team, "R: A language and environment for statistical computing", 2013.

[14] R. Hyndman, Y. Khandakar, "Automatic Time Series Forecasting: The forecast Package for R", Monash University, Department of Econometrics and Business Statistics, (Report No. 6/07), 2007.

[15] http://CRAN.R-project.org/package=forecasting. [Online], last accessed June 2014.

[16] V. Assimakopoulos, K. Nikolopoulos. "The theta model: a decomposition approach to forecasting." International journal of forecasting 16(4), 2000, pp. 521-530.

[17] C. X. Mavromoustakis, H. D. Karatza, "Performance evaluation of opportunistic resource sharing scheme using socially-oriented outsourcing in wireless devices", The Computer Journal, 56(2), 2013.

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