Efficient Entertainment Services Provision over a Novel Network Architecture

Yiannos Kryftis, George Mastorakis, Constantinos X. Mavromoustakis, Jordi Mongay Batalla, Evangelos Pallis, and Georgios Kormentzas

Abstract

The provision of entertainment services through current network infrastructures generates a significant part of global Internet traffic. Such traffic is created by users who are interconnected through social media networks, distributing multimedia content, as well as through their online games interaction, leading to a significant increase in data traffic demands. At the same time, users demand the high-quality content delivery necessary for such multimedia-based services and applications. In this article, a novel network architecture is proposed that exploits a resource prediction engine to predict the upcoming demands on resources for the optimal distribution of streaming data among content delivery networks, emerging mobile networks, cloud-based providers, and home media gateways. Performance evaluation experiments carried out under controlled conditions verify the validity of the proposed network architecture and establish its effectiveness for efficient resources prediction during the provision of entertainment services.

Introduction

Given the tremendous demand for multimedia services provision over emerging mobile networking platforms, more pressure is being applied on further research and development for the efficient support of entertainment applications. A significant part of global Internet traffic is generated by entertainment services like video/audio on demand and online gaming services, while the amount of this traffic is expected to double in the future, leading toward the future media Internet. Recent advances in social media networks and the rapid distribution of information due to them are the main impetus for the upcoming entertainment services provision framework. The broadband infrastructure growth, advances in wireless technologies, and cloud computing, which has emerged as a new paradigm for hosting and delivering services over the Internet, are the keystones of its success. Citizens are surrounded by an environment that rapidly evolves, leading them to demand community-centric experiences with better quality of experience (QoE). Quality of entertainment services presupposes efficient content delivery as an important part of any multimedia or online gaming activity. In this direction, the pressure is carried into the research on novel network architectures, as well as the relative components that allow efficient and balanced multimedia content delivery, while there is a need for the adoption of novel algorithms and models in emerging networking platforms to predict network resources, toward efficient multimedia content provision.

Several existing research attempts elaborate on the combination of different delivery methods in order to achieve better QoE for the end users. A content distribution network/peer-to-peer (CDN-P2P) hybrid architecture for cost-effective streaming media distribution combines the advantages of using CDN for providing high QoE [1] with the low cost of using P2P-based streaming, which can be used for live video streaming or other entertainment services. Current research approaches focus on how to benefit from the combination of the different delivery methods, but they do not take into consideration handling each resource separately. In comparison to such approaches, the proposed solution does handle each resource separately (i.e., streaming channel) based on measured values of the resource metrics and predictions of future values. A resource prediction engine constitutes an important part of an entertainment content delivery system in order to offer the desired QoE to end users. Its role is to provide a mechanism to efficiently predict future values based on previous values of each metric, like bandwidth capacity or round-trip delay time, which can forecast upcoming usage rising or network fluctuations. The prediction engine has to be based on novel methods and models that can accurately forecast future demands in order to trigger, through the management plane, the proper actions to keep the desired quality of multimedia streaming sessions. Time-series analysis techniques can be utilized for the prediction of a server bandwidth demand, while P2P content delivery can support the process in video-on-demand (VoD) services [2]. For the prediction of the future population of each video channel, Box-Jenkins models [3] are possible to exploit with input on the population of the video channel in the past. The auto-regressive moving average (ARMA) model and sea-
sonal auto-regressive integrated moving average (ARIMA) model can also be exploited to avoid periodicity. Therefore, the near future demand expectations are estimated based on the history of demands as monitored by cloud monitoring services. For a resource prediction engine able to forecast future demands, the recent advances in connected media technologies and social networks should be taken into account. In this framework, social media networks play a significant role in multimedia content delivery by providing ways of interactions among users that can lead to lightning fast spread of content [4]. A probabilistic resource provisioning approach is suggested in [5] that uses standard models developed for epidemiology spreading to represent sudden and intense workload overflow in the VoD delivery process. Epidemic model spreading in scale-free networks has also been intensively studied in [6] with the main effort on the spread of computer viruses, which really resembles the epidemic spread of human diseases.

In this context, this article exposes and proposes an entertainment services delivery solution based on the combination of heterogeneous delivery methods/systems for the efficient content transmission to satisfy different types of users' requests. The underlying delivery methods include direct content delivery from conventional clouds that can be public or private computing infrastructures, delivery through content delivery networks (CDNs), and peer-to-peer delivery among users through home media gateway clouds. Entertainment services can be decomposed in content delivery for multimedia services like video and audio on demand and online gaming services support, which involves more interactive concepts, requiring higher upload bandwidth for streaming as well as lower latency. This solution considers the context of each resource, utilizing different models for each type of entertainment service based on its nature. Among others, an epidemic spread model is proposed that accurately defines a VoD spread in the case of viral spreading through social media networks. Based on the resource usage predictions in each case, decision algorithms can make the selection among the most efficient multimedia delivery methods. In the following sections, the proposed network architecture is presented, emphasizing the resource prediction engine to efficiently optimize multimedia content delivery. Then the performance evaluation results are provided to validate the proposed approach, while the final part concludes this article by highlighting the main concept presented.

A Resource Prediction Engine for Multimedia Content Delivery Optimization

The synergy among traditional CDNs, P2P, and geographically distributed cloud-based streaming approaches is a very promising solution that utilizes the stable edge transport capability of CDNs, the scalable, last-mile transport capability of P2P approaches, and the resource elasticity offered by cloud computing, while avoiding excessive use (geographic inter-domain span) of P2P delivery, which is Internet service provider (ISP)-unfriendly. Based on the above, this section elaborates on an innovative and scalable media distribution system for mediating resource reservation requests to the underlying physical resource infrastructures (i.e., cloud data centers, telecommunication infrastructures, including CDNs, as well as user-created media home gateway clouds), able to correlate the architectural solution with the target number of users and media events to be accommodated. The anticipated user-driven QoE enhancement calls for technologies that can offer optimized delivery of media events in an end-to-end approach, even over highly heterogeneous network environments, making imperative the use of real-time interaction mechanisms, via which users can enhance their QoE under various contexts or diversified operating conditions (home, work, mobile, etc.). Additionally, the design takes into consideration the recent advances in mobile services, providing support for future 5G wireless communication systems. The proposed network architecture is depicted in Fig. 1 with an upper layer (called the delta management and control [M&C] plane) that coordinates the collaboration environment of the four separated planes as detailed below.

Network Architecture for Entertainment Services Provision

The delta M&C plane is a central point of the proposed system from the location perspective, as it is placed in the telco’s/provider’s domain, and from the functional perspective since it takes the role of coordinator for optimal exploitation of the resources that are available at each resources provider (telecommunication network operators, CDN providers, conventional cloud operators, users’ home gateways, 5G access networks). Coordination is an intensely dynamic process that considers not only all ongoing streaming sessions, but also all potential upcoming ones based on predictions of future needs. The CDN/cloud M&C plane is placed at the premises of the entertainment service provider. It is responsible for deciding on the bit rate and the server that handles each user’s requests. The decision process utilizes the data acquired from the delta M&C plane, related to the quality level of the CDN/cloud domain access points and the knowledge about resources available in this domain. The MHGC M&C plane is a distributed management plane, consisting of the users’ home gateways, which together establish and management of an MHGC ad hoc system. The streaming data plane is in charge of the delivery of the requested multimedia content to the user’s application in accordance with the rules worked out by the CDN/cloud M&C (or MHGC M&C) plane. For interactive game delivery, the streaming data plane includes dedicated game servers responsible for game streaming. The access network includes support for concepts covered by the upcoming 5G mobile networks [7, 8]. Device-to-device (D2D) communication enables the exchange of data traffic directly between devices without the requirement for base stations, supporting new usage models based on the prox
The proposed architecture benefits entertainment service providers by offloading the traffic from one origin hosting server/servers structure to multiple dispersed localizations, which usually results in significant cost reduction in hosting and bandwidth consumption. The consumers benefit from better quality, availability of the content, and the speed of its delivery.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.

Figure 1. The proposed overall network architecture.
tion for the traffic in the network. It utilizes:

- Epidemic models for resource prediction if the data-volume demands correspond to epidemic spread of the content
- A combination of suitable time series models for the forecast of future levels of the values of the resources' parameters based on previous data

The strength of the RPE originates from the ability to use historical data to select, from a pool of possible models, the model that fits best in order to be utilized for the prediction. The outcome of the forecasts is used as input to the resource allocator/scheduler, which uses algorithms to combine the prediction methods, allowing it to decide on the optimal delivery methods. The decision about the optimal delivery method to be used and the recommendations on which the CDN server should stream the requested media, or which game server should be utilized for interactive game support, are forwarded to the MSM component. In the case of using P2P delivery, the bandwidth allocation optimizer calculates the optimal bandwidth allocation for the P2P delivery between the MHGC devices. MHGC devices exchange management information between them and together constitute the M&C plane that manages the P2P network between the EHG devices. Additionally, the bandwidth allocation optimizer is responsible for handling the special features of 5G access networks.

**Epidemic Models for Download Rates Prediction of Video on Demand Services**

The high contact rate and the ability to instantly share information over the Internet can cause a spread of content that resembles the epidemic spread of diseases. In this context, epidemic models can help in the prediction of VoD usage as part of the general issue of optimal content delivery for entertainment services provision. The majority of disease models are based on a splitting in compartments of the individuals in a population based on their disease status. The basic susceptible-infectious-recovered (SIR) model provides the foundations based on mathematical epidemiology [10]. The proposed model extends the SIR model, which includes only three states, and divides the population into six compartments based on the percentage of the entertainment service subscribers in each state. The subscribers (S) group includes the subscribers of the service who can download the video; the active (A) group includes users who are currently downloading the video; while the infected (I) group contains users who have downloaded the video and can spread their opinion about it through social networks. In addition, the recovered (R) group contains users who have passed from the infected phase but do not spread the video anymore (after some period of time), while the deleted (D) group includes users who have deleted the video after watching it, or the case in which the video was auto-removed from the cache after some period of time. The turned down (T) group includes users who belonged to the S category but consciously rejected the video, so they will never download it.

In the model, $N_{S(0)}$, $N_{A(t)}$, $N_{I(t)}$, $N_{R(t)}$, $N_{D(t)}$, $N_{T(t)}$, $t > 0$ are stochastic processes representing the time evolution of each population. Suppose there are $n$ clients of the VoD provider; they all belong to group S at time $t = 0$, when a VoD is initially uploaded by the provider. The transition rate from state S to A consists of the probability of having a new spontaneous viewer plus the probability of having learned about the video from their social contacts, multiplied by $p$, that is, the possibility for a user to download in case it comes to that dilemma. The rate can be expressed as $N_{S(t)} = (\gamma + \beta N_{S(t)} N_{R(t)}) \cdot p$, where $\beta$ is the social network contact rate for users based on the specific video and $\gamma$ is the number of spontaneous viewers. For the transition from state A to I, there is a need for consideration of the download rate of users and the length of the video, but since the video download is happening directly during viewing, the transition rate $\lambda$ can be expressed as a constant delay with value the duration of the video. The transition rate $\kappa$ from I to R, and $\delta$ from R to D, are random variables and can be expressed as Poisson processes with mean time the average period of time of spreading after viewing of a VoD (1–2 days) for $\kappa$ and

![Figure 2](image-url). Internal architecture and workflow of media distribution middleware: a) internal architecture of MDM; b) workflow of MDM.
For the prediction of the bandwidth usage for the VoD, the prediction engine managed to forecast the continuity of the increment of the bandwidth usage, while in the case of the RTT for the online gaming, it managed to forecast that future values would be around the mean value.

The mean time of keeping the VoD on the device before deleting (7 days based on auto-remove) for $\lambda$. Figure 3a presents the transitions between available states, and Fig. 3b and Fig. 3c simulation results executed in Matlab for representative scenarios of VoD delivery with epidemic spread and without epidemic spread over discrete time($t$). At the beginning the whole population belongs to the $S$ state. In Fig. 3c, which presents the scenario without epidemic spread, the population of users in the Active ($A$) state remains below 0.02 during all the delivery process. In Fig. 3b, which presents the scenario with epidemic spread at $t = 50$ time steps, the population of $A$ increases, showing that there are active users downloading the video, while at $t = 70$ time steps, the population of $I$ increases significantly.

Line $A(t)$ depicts the bandwidth need to cover the needs of the active downloads. In the case of epidemic spread of a specific video, the population of simultaneous downloads is significantly increased, something that increases the difficulty in delivering high quality of service. An important observation is that by the time the active users introduce a significant increase in their population, there is always an important number of users in the $I$ and $R$ states that can seed the video for them through P2P delivery. The model provides the ability to make predictions of the future spread of the VoD based on its initial spread. The collected monitoring data of usage metrics are fit into the prediction engine, which is able to foresee whether it follows the epidemic spread model that could cause an enormous spread of the VoD to become viral.

**Algorithms for Resource Prediction**

The MDM component uses the predicted future values for the metrics in order to make decisions on delivery of the requested media, which may be streamed:

1. Directly from the cloud
2. Through deployed surrogate servers of the CDN
3. By establishing an MHGC ad hoc system and using a combined P2P-based technology of dis-

![Figure 3. Epidemic model for efficient multimedia content provision: a) the proposed epidemic model for efficient content provision; b) simulation results for percentage of population in each state (epidemic spread); c) simulation results for percentage of population in each state (no epidemic spread).]
distribution with multi-source, multi-destination congestion control algorithms.

4. A combination of parts or all of the above

The delivery over a combination of delivery methods is performed thanks to a stream-switching adaptation technique, which assumes the encoding of the content at increasing bit rates and at multiple locations. The flowchart of Fig. 2b presents a simple algorithm for deciding on the delivery methods concerning serving a specific VoD. If the VoD spreads epidemiologically, it enables all delivery methods. Otherwise, it makes use of the highest value for bandwidth need, among current and predicted values, to select the delivery methods to be used. This variable is called \( \text{neededBW} \), since it is corresponds to the needed bandwidth capacity over the network to satisfy users’ demands. In the case that \( \text{neededBW} \) is below a predefined low threshold, only cloud delivery will be used. However, if it is above an intermediate threshold, CDNs will be used, while the number of CDNs used can be increased according to the demand. After the top threshold is reached, a P2P delivery method is exploited in addition.

The quality alert is performed when the MDM predicts a QoE lower than the expectations. As shown in the workflow of Fig. 2b, MDM utilizes the network metrics to identify whether the performance issue is on server or client side. If the issue is on the client side, it triggers the adaptation procedure to deliver content at a lower representation rate, while if it is on server side, it triggers reselection of the optimal server to deliver the content.

**Performance Evaluation Analysis and Experimental Results**

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, we collected, from an entertainment services provider, monitoring data of the bandwidth usage for serving the need of a specific VoD and the round-trip delay time (RTT) for the users of an online game. The collected measurements were for a total of 60 min with a period of 5 s, but to avoid periodicity of data, the mean value per minute was used. Eighty percent (48 min) of the data was exploited to feed the prediction engine, while the rest of the data was used for evaluation through a comparison between the predicted and actual value, as shown in Fig. 4a. The important part of the graph is after the first 48 min, where it is clearly depicted that the measured values remain very close to the predicted ones for the specific sample. For the prediction of bandwidth usage for the VoD, the prediction engine managed to forecast the continuity of the increment of bandwidth usage, while in the case of the RTT for online gaming, it managed to forecast that future values would be around the mean value, as shown in Fig. 4b.

Figure 4c presents the scenario where the RTT is increasing. At time \( t = 50 \) min the prediction engine is able to forecast that in the next 10 min, at \( t = 60 \) min, the RTT will go above 50 ms, something that will affect the QoE for the users. As described in the use case discussed earlier, this will trigger the adaptation process through the components of the architecture that will perform management actions in order to alter the game server that serves the specific online game. The adaptation process was successful, and it led to a new average RTT around 35 ms. The obvious benefit through the whole procedure was the proactive response based on the prediction that kept the QoE at high standards through the whole process.

The remaining extracted results show the performance response of the proposed framework in contrast to the diversity characteristics of the utilized communication technologies through simulation experiments. Toward implementing such a scenario, a common lookup application service for video streaming is set in each node (both

---

**Figure 4.** Performance evaluation results: a) bandwidth usage for service of a specific VoD; b) average RTT for users of an online game; c) adaptation process after the forecast of high RTT for users of an online game.
static and mobile) to enable nodes requesting a stream from a certain user. Figure 5 presents experimental simulation results for the evaluation of the performance and the offered reliability in streaming activities offered by the proposed system with special attention to wireless transmissions through 5G mobile networks. Figure 5a presents the respective complementary cumulative distribution function (CCDF) that represents the sharing reliability with the download time for requests up to 25 MB. The simulation results depict that by using the proposed framework in the presence of Rayleigh fading and mean noise of 4 dB, the reliability is not importantly affected. Finally, the throughput with the number of requests per second is presented in Fig. 5b, depicting that the fading channels with noise have low throughput exhibition, especially when the number of requests per second increases.

**CONCLUSION**

This article presents a novel network architecture that exploits a resource prediction engine, able to utilize innovative models, including epidemic spread and time series schemes, for forecasting future usage demands for network resources. The proposed architecture is able to increase the QoE during the provision of entertainment services by performing optimal distribution of the streaming data among content delivery networks, cloud-based providers, and home media gateways. At the same time, it provides support for future 5G wireless communication systems. The effectiveness of the proposed system is validated by obtaining experimental and simulation results. Future directions in our ongoing research encompass the inclusion of information diffusion and spread models as well as models that describe entertainment services provision and optimize the multimedia content delivery over 5G mobile networks.

**ACKNOWLEDGMENT**

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

**REFERENCES**


**BIOGRAPHIES**

**YIANNOS KRYFTIS** (kryftis.y@unic.ac.cy) holds a graduate degree in electrical and computer engineering from the National Technical University of Athens (NTUA). He also holds an M.Sc. in advanced computing from Imperial College London. He is a Ph.D. candidate at NTUA in management architectures and systems for future Internet networks. Currently he is working as a researcher on the Department of Computer Science at the University of Nicosia. His research interests include policy-based network management, network architectures for the future Internet, virtualization information models and algorithms in federated environments, and multimedia services provisioning.

**GEORGE MASTORAKIS** received his B.Eng. (Hons) in electronic engineering from the University of Manchester Institute of Science and Technology, United Kingdom, in 2000, his M.Sc. in telecommunications from University College London, United Kingdom, in 2001, and his Ph.D. in telecommunications from the University of the Aegean, Greece, in 2008. He is serving as an assistant professor at the Department of Computer Science at the University of Nicosia. His research interests include cognitive radio networks, networking traffic analysis, radio resource management, and energy-efficient networks. He has more than 100 publications in various international conference proceedings, workshops, scientific journals, and book chapters.

**CONSTANTINOS X. MAVROMOUTAKIS** is currently an associate professor in the Department of Computer Science at the University of Nicosia, Cyprus. He received a five-year dipl. Eng (B.Sc./B.Eng./M.Eng., KISATS approved) in electronic and computer engineering from the Technical University of Crete, an M.Sc. in telecommunications from University College London, and his Ph.D. from the Department of Informatics at Aristotle University of Thessaloniki, Greece. He is...
leading the Mobile Systems Lab. (MOSys Lab., http://www.mosys.unic.ac.cy/) in the Department of Computer Science at the University of Nicosia. He has extensive experience in simulation of wireless systems, MP2P and opportunistic systems, mobile computing, and web technologies. He serves as an Associate Editor for the International Journal of Communication Systems (Wiley) and an Editorial Board member of the Simulation Practice and Theory Journal (Elsevier), and as a Technical Program Committee member for various international conferences.

Jordi Mongay Batalla received his M.Sc. degree from Universitat Politècnica de València (2000) and Ph.D. degree from Warsaw University of Technology (2009), where he still works as an assistant professor. In the past he worked at Telcordia Poland (Ericsson R&D), and he is still with the National Institute of Telecommunications, where, since 2010, he is head of the Internet Architectures and Applications Department. He has taken part (coordination and/or participation) in 12 international ICT projects, five of them inside EU Framework Programmes. His research interest focuses mainly on quality of service (Diffserv, NGN) in both IPv4 and IPv6 infrastructures, future Internet architectures (content-aware networks, information-centric networks) as well as applications for future Internet (Internet of Things, smart cities, IPTV). He is author or co-author of more than 100 papers published in books, and international and national journals and conference proceedings.

Evangelos Pallis received his B.Sc. in electronic engineering from the Technological Educational Institute of Crete in 1994, his M.Sc. in telecommunications from the University of East London in 1997, and his Ph.D. in telecommunications from the University of East London in 2002. He currently serves as an associate professor at the Technological Educational Institute of Crete in the Department of Informatics Engineering and acts as the director of research and development of the Telecommunication Systems Laboratory. His research interests are in the fields of wireless broadband and mobile networks and network management. He has more than 100 publications in international scientific journals, and conference and workshop proceedings.

Georgios Kormentzas received his Diploma in electrical and computer engineering and Ph.D. in computer science from NTUA in 1995 and 2000, respectively. He is currently serving as an assistant professor in the Department of Information and Communication Systems Engineering at the University of the Aegean. His research interests are in the fields of traffic analysis, network control, resource management, and quality of service in broadband networks. He has published extensively in international scientific journals, and edited books and conference proceedings. He is a member of renowned professional societies, and an active reviewer and Guest Editor for several journals and conferences.