Simulation Based Evaluation of Dynamic Resource Allocation for Adaptive Multimedia Services

Krunoslav Ivesic, Maja Matijasevic, Lea Skorin-Kapov
University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, Zagreb, Croatia
{krunoslav.ivesic, maja.matijasevic, lea.skorin-kapov}@fer.hr

***Abstract*—As multimedia services may consist of several media flows with varying importance to the users, it is beneficial to consider the relative flow importance while negotiating flow parameters. By regarding all service flows at once, a session can be described by creating the optimal session configuration and several alternative ones the resources of which are decreased with respect to user preferences regarding flow importance. In this way, session configuration can be switched to a less resource-demanding but previously agreed on configuration in a situation when the network resources become scarce, thus mitigating the effect of degradation on users. We present a mathematical model for optimized resource allocation for sessions with predefined configurations, which maximizes user's utility and operator's profit subject to resource constraints, while taking the priorities of users and services into account. In order to evaluate the proposed model, we develop a simulator tool named ADAPTISE that simulates arrival, duration and resource allocation of sessions and dynamically performs optimization of available resources.**

I. Introduction

Recent achievements in the development of mobile networks and mobile user equipment devices have enabled provision of complex networked multimedia services to mobile users. These services can be composed of several media components, e.g., audio, video, and 3D graphics, the number of which, as well as their properties, can vary over time. The resource requirements of such services can be significant and it is necessary to provide a solution for smart resource allocation and reallocation in case of congestion.

In our previous work, we have proposed a structure called Media Degradation Path (MDP) to describe session resource requirements and user experience [1]. We have defined the MDP as an ordered collection of session *configurations*, where each configuration specifies operating parameters of flows (e.g., codec, frame rate), resource requirements, and an overall *utility* value, i.e., a value that quantifies the degree of user satisfaction regarding configuration quality. Each MDP contains an optimal configuration and several alternative ones, ordered in decreasing order of utility value.

In this work we focus on dynamic resource allocation mechanisms for multimedia services described by using MDP. We present a mathematical model for optimized resource allocation and a simulator which demonstrates the dynamic behaviour of multimedia sessions and resource reallocation in case of congestion. We take into account multiple session media flows and their changing number and requirements. By utilizing the MDPs of all active sessions, it is possible to optimally distribute the available resources among them, and redistribute them in case of congestion while maximizing utility. The optimization process is based on making a choice of the "right configuration" for each session in given situation. In addition to our previous work presented in [2], where the model was used statically, i.e., the optimization process was run on a dataset prepared in advance, in this work we simulate the session arrival and duration and run the optimization process automatically when necessary. Due to the complexity of the optimization problem, we use a heuristic algorithm from [3] to find solutions to the optimization problems.

II. Related Work

Literature on resource allocation mostly focuses only on services requiring a single flow (e.g., voice). Service priority-based resource reallocation has been suggested in [4]. A model for proportional degradation in [5] guarantees the ratios and frequency of degradation among different service classes. A degradation model in [6] degrades sessions with maximum assigned bit rate. In comparison, our work takes into account multiple flows and multiple media types.

Degradation based on utility functions (map utility values to assigned bit rate) can alleviate the negative effect experienced by the end user. An approach presented in [7] defines different utility functions for different service types. In [8] utility functions for multimedia services are defined per flow and the degradation is conducted by maximizing the sum of utility values of all flows of active sessions subject to resource constraints. Our work differs in that we consider operator profit, as "operator utility".

In order to address model applicability, we examine the proposed approach in the context of the Evolved Packet System (EPS) specified by the 3GPP which introduces the class-based concept for QoS management whereby each bearer is assigned a QoS Class Identifier (QCI) that specifies standardized packet forwarding behaviour [9]. Nine QCIs have been specified, each of them defining: priority (1 to 9; 1 being the highest priority), bearer type (guaranteed or non-guaranteed bit rate), packet delay budget, and packet error loss rate. In order to describe the priorities of sessions in cases when resource reallocation is necessary, the Allocation and Retention Priority (ARP) is defined [9]. The ARP specifies whether a session can lose

its resources, or acquire the resources currently assigned, to another session.

III. Model Description

We postulate the following negotiation process: at session initiation, the participants (client and server or two clients) negotiate and agree on session parameters by specifying possible operating parameters (e.g., codec, resolution, frame rate), corresponding resource requirements, and achievable utility (e.g., corresponding to user perceived quality). Negotiated parameters serve as input for calculating a session MDP.

During the negotiation process, user preferences are taken into account. The user specifies the importance of particular service components and this drives the process of creating alternative service configuration(s), e.g., if the service contains audio and video flows and the user specifies that the audio flow is more important, the alternative configuration(s) should keep the audio quality as high as possible, while degrading the video quality as much as necessary. In this way, knowledge about the service is introduced in service provisioning, thus enabling the creation of "personalized" services, where switching from one configuration to another in case of congestion is more acceptable than a change insensitive to the type of service that could degrade the service performance unacceptably.

In order to describe service dynamics, the configurations are grouped according to the service state they pertain to. By the term *service state* we refer to a set of service components that are simultaneously active at a given time interval of the service provisioning. The addition or removal of a media flow to an ongoing session triggers the change of the active service state, thus causing the switch to another configuration in a new service state. If the degradation process is due to occur, the new configuration which the session will switch to, is chosen from the set of configurations belonging to the currently active service state. An example MDP for a session with three service states is shown in Fig. 1. The service consists of a 3D graphical virtual world with the ability of adding a video stream or an audio chat. When only 3D graphics are active, the service is in *State 1* which requires a single bearer. In the case of adding an audio or a video flow, the service state is changed to *State 2* or *State 3* and a new bearer is required. Because exactly one state of a session is active at any given time, the mathematical model takes into account only the active state of each session. The objective is to maximize the total system utility subject to resource constraints by selecting a single configuration from the active state of each session. The optimization problem class is multi-choice multidimensional knapsack problem (MMKP), formulated as follows.

Let n be the number of sessions and p_u the number of configurations in the active state of the session u. Let configuration i of session u have z_{ui} media flows, with flows $1, ..., h_{ui}$ pertaining to the downlink direction and flows $h_{ui}+1, ..., z_{ui}$ pertaining to the uplink direction. The bandwidth requirements of configuration i flows can be described by the vector $\mathbf{b}_{ui} = (\mathbf{b}_{ui1}, ..., \mathbf{b}_{uiz_{ui}})$ where $\mathbf{b}_{uij} = (b_{uij1}, ..., b_{uij9})$ describes the requirements of the flow j and it is composed

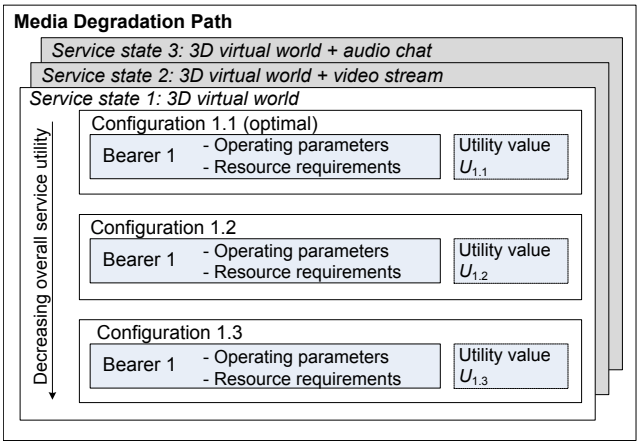

Fig. 1. MDP for an example service with three service states

of 9 components, each of them describing the requirements of a corresponding QCI. Because only one QCI can be assigned to the bearer, it is assumed that only one component of this vector is greater than zero while others are equal to zero. Let $U(\mathbf{b}_{ui})$ denote the utility value and $P(\mathbf{b}_{ui})$ the operator's profit of configuration i. Because the objective function is a weighted sum of utilities and profits of all sessions, we define the weight factor w_{ut} for the total user utility and w_{pr} for the total profit, thus converting the multi-objective problem to the single-objective one. Each session also has a weight factor w_u defined by $w_u = w_u^c w_u^s$ where w_u^c and w_u^s are the weight factors of a user's category and a service, respectively. In order to enable the fair comparison of the sessions with possibly very varying utility values and profits, we normalize these values. The configurations in each state of the MDP are ordered by their decreasing utility value so we can calculate the normalized utility values U_n by dividing all values by the first one: $U_n(\mathbf{b}_{ui}) = U(\mathbf{b}_{ui})/U(\mathbf{b}_{u1})$. The normalized profit is $P_n(\mathbf{b}_{ui}) = P(\mathbf{b}_{ui})/max_i P(\mathbf{b}_{ui})$. Let B_{kD} and B_{kU} denote the maximum bandwidth for QCI k for the downlink and uplink directions respectively as specified by the network operator (we consider there is no resource borrowing between QCIs). The optimization problem formulation is:

$$max \sum_{u=1}^{n} \sum_{i=1}^{p_u} \{w_u x_{ui} [w_{ut} U_n(\mathbf{b}_{ui}) + w_{pr} P_n(\mathbf{b}_{ui})]\} \quad (1)$$

such that:

$$\sum_{u=1}^{n} \sum_{i=1}^{p_u} \sum_{j=1}^{h_{ui}} x_{ui} b_{uijk} \leq B_{kD},\ k = 1, ..., 9 \quad (2)$$

$$\sum_{u=1}^{n} \sum_{i=1}^{p_u} \sum_{j=h_{ui}+1}^{z_{ui}} x_{ui} b_{uijk} \leq B_{kU},\ k = 1, ..., 9 \quad (3)$$

$$\sum_{i=1}^{p_u} x_{ui} = 1,\ x_{ui} \in \{0, 1\},\ u = 1, ..., n \quad (4)$$

where x_{ui} are binary variables used to denote the selected configurations. The result of the maximization is the list of the

values of the variables x_{ui} indicating selected configurations that maximize the weighted sum of total utility and profit.

If we assume exponential interarrival time distribution of each session type, our system can be modelled as an $M/G/\infty$ queue (M=exponential distribution of interarrivals, G=arbitrary distribution of duration, ∞=infinite number of servers, where "serving" is assigning the required resources to the session. We can assume the infinite number of "servers" because we are interested only in the situations when new sessions can still be admitted). Let λ be the parameter of the interarrival time distribution (mean=$1/\lambda$) and μ the mean of duration distribution. According to [10], the number of sessions in such a system is Poisson distributed with the mean $\rho = \lambda\mu$. With the expected number of sessions and known average bandwidth requirements of each service, we can estimate the total required bandwidth of such parameter setting and study the resulting resource allocation and optimization process. Table I lists interarrival time and distribution parameters and the estimated number of sessions that have been used for the experiments.

IV. Model Simulation and Evaluation

In order to demonstrate the application of our mathematical model, we simulated arrivals, durations, resource consumption and state changes of multiple sessions. We have identified different types of services (e.g., video call, voice call, video streaming, gaming) and assume that distributions for interarrival times and durations are known for each service type (examples used in our simulation are given in Table I and described in more detail later). Upon arrival of a session in the system the resources for the best configuration of the session's active state are occupied (adaptive admission control mechanisms are planned for future work).

The optimization process responsible for resource (re)allocation is triggered when resource consumption surpasses a predefined threshold (as specified by operator policy, we set 95% of QCI as a threshold). As an input to the optimization algorithm, we specify the bandwidth constraint (maximum available bandwidth for a given QCI) as 80% of the actual bandwidth allocated to that QCI. The reason for this is to prevent the optimization process from running too often, since frequent optimizations in a real network would cause significant signalling load. The events that can result in an increase in resource consumption are arrival of a new session and a state change within an active session.

To perform simulations, we have designed and implemented a tool named ADAPTISE (ADmission control and resource Allocation for adaPtive mulTImedia SErvices). The graphical user interface of the tool is depicted in Fig. 2 and displays the visualization of active configurations of sessions in the form of a matrix of squares where each session is represented by a column. The currently active configuration of a session is coloured black while the inactive configurations are coloured white. The bottom row pertains to the configurations with the highest utility value, the other rows pertain to the remaining configurations; the higher the row, the lower the utility value of

Fig. 2. Simulator ADAPTISE GUI

the respective configuration. Two sets of simulations have been performed, with QCI limits set to 100 kbps and to 90 kbps, as depicted in Figs. 3(a) and 3(b), respectively. For each set 5 simulation experiments have been run. The simulated time interval has been set to 1 hour. The horizontal axis represents time and the vertical axis represents the simulation instances. The points pertain to moments in time when the optimization process has been triggered. In Fig. 3(a) it can be seen that the number of optimization triggers goes from 0 to 5, while in 3(b) the number of optimization triggers goes from 10 to 17. While the latter might seem as too high optimization frequency, it should be noted that the mean duration of a session was 100 s, with the experiment setup as in Table I, which means that the optimization process has been run, in average, every 4 to 6 minutes. In that case it is not very likely that the optimization process will run more than once per single session. However, on both graphs it can be noted that some occurrences of the optimization process are followed by the next occurrence that happens rather quickly, e.g., in Fig. 3(a) at around 40 minutes in simulation instance no. 3 there are 2 consecutive optimization triggers. To prevent the optimization process from running to often, adequate admission control mechanisms are needed, and they will be addressed in future work.

Table II summarizes the effects of the optimizations for the two simulation sets described. The average and the standard deviation are given for the following measures: the total utility values before and after each optimization process, the number of active sessions, the number of sessions that have been processed by the optimization algorithm (those sessions that consume resources of QCIs that became scarce) and the number of degraded sessions.

By using the mean of the interarrival times of the sessions from Table I, the expected total number of sessions can be calculated by dividing the duration of the simulation (1 h) by the mean of the interarrival time, giving 3601 sessions as a result. With the average of 13 degraded sessions per optimization and at most 5 optimizations per hour for the limit

Service	Interarrival param.	Interarrival mean [s]	Duration distribution	Duration mean [s]	Expected sessions [#]
VE	$\lambda = 3 \cdot 10^{-4}$	3.33	Exponential, $\lambda = 10^{-5}$	100	30
MMO	$\lambda = 1.5 \cdot 10^{-4}$	6.67	Normal, $\mu = 10^6, \sigma = 3 \cdot 10^5$	100	15
Video call	$\lambda = 2.5 \cdot 10^{-4}$	4	Lognormal, $\mu_L = 9.5, \sigma_L = 2$	99	25
Voice call	$\lambda = 2 \cdot 10^{-4}$	5	Erlang, $r = 30, \mu = 10^{-5}$	100	20
Video streaming	$\lambda = 10^{-4}$	10	Exponential, $\lambda = 10^{-5}$	100	10

TABLE I
SESSION INTERARRIVAL TIME AND DURATION PARAMETERS

QCIs [kbps]	Utility before	Utility after	Sessions (total) [#]	Processed sessions [#]	Degraded sessions [#]	Optimizations [#]
100	65.82 ± 4.82	61.53 ± 5.30	108.42 ± 8.06	33.92 ± 2.97	13.00 ± 2.04	2.4 ± 1.95
90	63.39 ± 6.34	60.24 ± 6.23	104.86 ± 8.03	30.97 ± 2.38	10.63 ± 2.10	12.6 ± 2.88

TABLE II
OPTIMIZATION STATISTICS

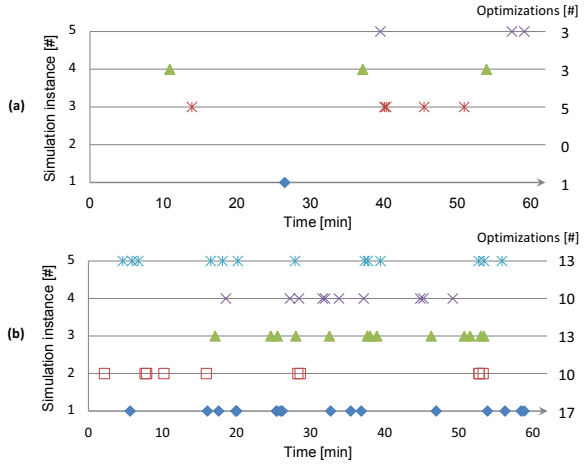

Fig. 3. Number and distribution in time of optimization occurrences with QCIs set to 100 kbps (a) and 90 kbps (b)

per QCI equal to 100 kbps, as stated in Table II, the expected number of degraded sessions is equal to 65, assuming at most one degradation per session. With the limit per QCI set to 90 kbps, the calculation yields 170 as the expected number of degraded sessions. Both 65 and 170 degraded sessions represent a minor portion of the total number of sessions, i.e., 1.81% and 4.72%, respectively. However, the problem of the consecutive optimizations that happen very quickly remains. With admission control mechanisms, it will be possible to project more realistic parameters for the real network.

V. CONCLUSIONS AND FUTURE WORK

By utilizing the structure of MDP to describe complex multimedia services, the resources can be optimally allocated and redistributed in case of congestion with total utility kept high and acceptable number of degraded sessions. The problem of consecutive optimizations remains and requires adequate admission control mechanism.

In our future work we plan to develop an admission control algorithm aware of the properties of the complex multimedia services with state changes and agreed on configurations. In situations where network resources are scarce, new sessions may be admitted with suboptimal configurations. The algorithm will take user and service priority as well as service configurations into account and it will be implemented as an extension to ADAPTISE. Furthermore, we will consider resource pre-emption and implement the possibility of varying interarrival time and duration distribution parameters, e.g., in order to model daily traffic patterns.

ACKNOWLEDGMENT

The authors acknowledge the support of the Ministry of Science, Education and Sports of the Republic of Croatia projects no. 036-0362027-1639 and 071-0362027-2329. The authors would also like to acknowledge the help of Mario Kusek and Iva Bojic in development of the Java based simulator.

REFERENCES

[1] L. Skorin-Kapov and M. Matijasevic, "A QoS negotiation and adaptation framework for multimedia services in NGN," in *10th Intl. Conf. on Telecommunications, ConTEL*, (Zagreb, Croatia), pp. 249–256, 2009.

[2] K. Ivesic, M. Matijasevic, and L. Skorin-Kapov, "Utility based model for optimized resource allocation for adaptive multimedia services," in *IEEE 21st Int. Symp. on Personal Indoor and Mobile Radio Communications (PIMRC)*, (Istanbul, Turkey), pp. 2638–2643, Sept. 2010.

[3] M. M. Akbar, M. S. Rahman, M. Kaykobad, E. Manning, and G. Shoja, "Solving the multidimensional multiple-choice knapsack problem by constructing convex hulls," *Computers & Operations Research*, vol. 33, no. 5, pp. 1259–1273, 2006.

[4] C. Lindemann, M. Lohmann, and A. Thümmler, "Adaptive call admission control for QoS/revenue optimization in CDMA cellular networks," *Wirel. Netw.*, vol. 10, pp. 457–472, July 2004.

[5] Y. Xiao, H. Li, C. L. P. Chen, B. Wang, and Y. Pan, "Proportional degradation services in wireless/mobile adaptive multimedia networks," *Wirel. Commun. Mob. Comput.*, vol. 5, pp. 219–243, March 2005.

[6] N. Nasser, "Real-time service adaptability in multimedia wireless networks," in *Proc. of the 1st ACM intl. workshop on Quality of service & security in wireless and mobile networks*, Q2SWinet '05, (New York, USA), pp. 144–149, ACM, 2005.

[7] N. Lu, J. Bigham, and N. Nasser, "Utility-based bandwidth adaptation for multimedia wireless networks," in *Adaptation and cross layer design in wireless networks*, pp. 149–181, Taylor & Francis, Inc., 2008.

[8] A. Brajdic, A. Kassler, and M. Matijasevic, "Quality of Experience based Optimization of Heterogeneous Multimedia Sessions in IMS," in *Baltic Congress on Future Internet Communications*, February 2011.

[9] "Policy and charging control architecture," 3GPP TS 23.203, Release 11, January 2011.

[10] I. Adan and J. Resing, "Queueing theory." Department of Mathematics and Computing Science, Eindhoven University of Technology, 2002.