Joint Network and Rate Allocation for Video Streaming over Multiple Wireless Networks

D. Jurca[1], W. Kellerer[2], E. Steinbach[3], S. Khan[3], S. Thakolsri[2] , P. Frossard[1]
[1]Signal Processing Institute, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
[2]Future Networking Laboratory, DoCoMo Communications Laboratories GmbH, Munich, Germany
[3]Media Technology Group, Technische Universitaet Muenchen, Germany
Email: dan.jurca@epfl.ch

***Abstract*— We address the problem of video streaming over multiple parallel networks. In the context of multiple users, accessing different types of applications, we are looking for efficient ways of allocating network resources and selecting network paths for each application, in order to maximize the overall systems performance. Our optimization joint problem consists of finding the appropriate application rate allocation and network parameters for each individual user, such that a universal system quality metric is maximized. A specific mapping between the requirements of each considered application and the overall quality metric is introduced, and our results are compared to other solutions based on throughput optimization strategies. The superiority and robustness of our approach is shown through extensive simulations in constant and dynamic systems, when clients can join/leave the access networks. Furthermore, we introduce heuristic algorithms which can obtain good results and are inexpensive in terms of computation and execution time.**

I. Introduction

The fast expansion of internet coverage and the increasing availability of wired/wireless network services encourage the development of QoS demanding applications. End users can seamlessly choose from a variety of parallel wireless services (e.g., UMTS/GPRS/WiFi) [1], in order to access these applications. Moreover, efforts towards inter-networking different wireless technologies are underway [2], to better meet QoS and cost requirements [3]. In such a context, managing the overall network resources, in the presence of multiple clients accessing simultaneously different applications, becomes of crucial importance for network operators.

With the latest wireless technologies, clients have parallel access to different applications, like web browsing/downloading, voice conversations and media streaming, each having their distinct QoS requirements and hence, their particular need of network resources. Standardized protocols for network resource allocation in application dedicated networks exist, e.g., GSM systems for voice applications, or the internet congestion control mechanisms for data traffic; however, they might prove suboptimal in a more complex environment, where different types of applications share common network resources.

In this paper we consider a multiple user scenario, where clients can access various applications with different Quality-of-Service (QoS) requirements over possibly multiple access networks (Figure 1). We discuss and solve a global optimization problem that periodically computes the optimal rate allocation and network selection for each user/application, given a universal quality metric. To this end, we take into account the parameters of the networks available to each user, and the specific characteristics of wireless applications. One by one, the behavior of each considered application is designed as a function of the user's network access parameters. Specifically, we derive a distortion model for streaming applications, which depends on the available data rate, transmission loss process at each client, and specific video sequence characteristics. Similarly, voice and data transfer applications are analyzed. Then, we define a universal quality metric that maps the QoS behavior of all applications as a function of the network parameters. Our final goal is to maximize the overall QoS of the system, under the given network resource constraints.

Real systems will often offer a limited choice in the mode of operation of the accessed applications; e.g. different voice transcoders operating at different rates in the case of voice conversations, a limited number of scalable encoded video layers for streaming applications, or a set of standard download rates for data transfer applications. Our final solution consists of an optimal decision on the mode of operation (total required rate) and network resource allocation for each client accessing a specific application. Such a global solution requires the computation over the whole set of application modes, for every user. Given the time varying nature of the wireless connections and the dynamics of users leaving/joining the system, the optimality of our solution is insured by iterative computations that take into account the actualized system status. To this end, we provide fast heuristic algorithms that can be used in real time system optimizations, based on the utility trade-off between system performance improvement and required resources [4]. We show that our QoS metric behaves well in a large set of system setups, and outperforms other traditional QoS metrics based on throughput, in terms of overall achieved quality, user fairness and adaptability to dynamic system setups. Finally, we show that our proposed heuristic algorithms obtain a close to optimum system performance with a low computational effort.

Our contributions in this paper are three-fold:

- First we introduce a video distortion model for scalable video coding. The model takes into account the overall encoding rate of the layered video, and the transmission loss process that affects the video packets of the different layers. The model is validated through extensive video experiments;

- In the context of multiple parallel applications over wireless networks, we discuss the opportunity of a single unifying quality metric that maps the specific requirements of each considered application to a single value. Later, this quality metric is used in our optimization framework for improving the overall system performance;
- Finally, we propose a fast heuristic algorithm which computes a close to optimum resource allocation solution in an iterative process, by taking into account the network access characteristics at each active client, along with the specific requirements of its desired application.

The rest of this paper is organized as follows: we review the relevant state of the art in Section II. Section III presents the considered applications and available access networks. We present our joint optimization problem in Section IV and explain our heuristic approach to solving it in Section V. We offer a concrete modelling example in Section VI. Extensive simulation results are presented in Section VII, while Section VIII concludes this paper.

II. Related Work

Media streaming applications over wireless environments have drawn the attention of the research community. The overview work of [5] gives a complete presentation of potential streaming systems in wireless networks and discusses the standardization efforts. The authors of [6] evaluate different mechanisms for robust streaming over WiFi networks. They propose an adaptive cross-layer protection strategy for robust and efficient scalable video streaming, by performing trade-offs between throughput, reliability and delay, depending on the channel conditions and application requirements. On the other hand efficient techniques for streaming over wireless networks which offer some QoS guarantees (e.g., UMTS networks) are presented in [7]. Here, channel efficiency is improved by using the common UMTS channel for streaming, along with proactive hybrid ARQ protocols. Furthermore, the authors of [8] present a resource allocation framework based on service differentiation and analyze the capacity benefit achieved through service prioritization and dynamic rate adaptation. Most of these works address the problem of media streaming alone, and do not consider the larger setup, when different applications, with possibly different quality requirements share the same wireless medium. In the same time, they do not address systems where multiple wireless services can be interworked in order to improve the end user experience.

Service interworking is slowly emerging as a viable commercial solution in order to achieve a better end-user application quality, over unreliable wireless transmission mediums. While initial commercial products already exist [1], standardization efforts are paving the way towards more advanced products and services [2], [9]. The authors of [10] present handover possibilities between WLAN and cellular wireless systems and discuss the possible issues and problems. We rely on these sustained efforts, and introduce a mechanism for the global optimization of system performance, when multiple clients, in the service area of more wireless networks access parallel applications. We rely on specific network access parameters at each client in order to take an optimal decision regarding the network resources allocation.

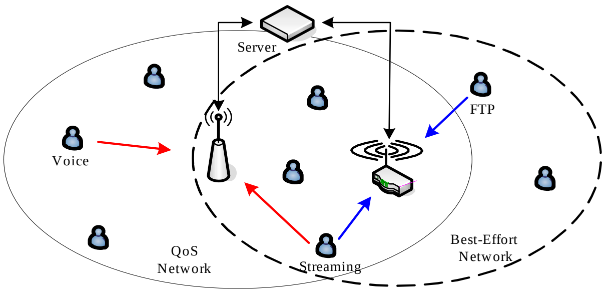

Fig. 1. Multiple wireless networks framework.

Finally, the recent works in [11]–[14] present a new framework for resource allocation and optimization in wireless systems. They exploit the information available at different layers of the network architecture in order to optimize the overall system performance. The authors of [15] describe a framework for the joint performance optimization of multiple parallel applications sharing the same wireless channels, under a universal quality metric. However, none of these early works address the problem of resource allocation and network selection when multiple users have access to several heterogeneous networks, administered by the same operator.

III. System Model

A. Multiple Applications

Assume N active users that simultaneously access via a server S any of three different types of applications, namely voice conversation (V), real-time media streaming (M) and FTP download (F). Let user i, $1 \leq i \leq N$ access one of the available applications k, $k \in \{V, M, F\}$, and let $\mathcal{M}_i = r_i$ be the mode of operation of user i, decided by S. It describes the average rate allocated to user i that has chosen application k. We assume that S can scalably adapt the transmission process to the channel conditions of user i. To this end, for each application k, the server can choose the right transmission parameter, from a predefined set of available parameters $\mathcal{P}_k$.

We first consider a multimedia streaming application that transmits a scalable encoded stream to the end user. Let L be the number of available encoded media layers available at the server S, where the layer $l \leq L$ is characterized by its average encoding rate ρ_l. The distortion of the multimedia, as perceived by the end client can generally be computed as the sum of the source distortion (D_S), and the channel distortion (D_L). In other words, the quality depends on both the distortion due to a lossy encoding of the media information, and the distortion due to losses experienced in the network. A commonly accepted model for the source rate distortion is a decaying exponential function on the encoding rate, while the channel distortion is roughly proportional to the number of lost packets and is differentiated by the importance of the video layer containing the lost packets. Hence we can explicitly formulate the video distortion metric as:

$$D = \alpha \cdot (\sum_{j=1}^{l} \rho_j)^{\xi} + \beta \cdot p_1 + \sum_{j=2}^{l} (p_j \cdot (D_{j-1} - D_l) \cdot \prod_{s=1}^{j-1} (1 - p_s))$$

where l is the total number of streamed video layers and α, ξ and β are sequence dependent parameters. D_j represents the source distortion of the first j layers of the video stream, and p_j is the average loss rate experienced during the transmission process by the video packets of layer j. Notice that our model for the loss distortion D_L separates the packet losses in the base layer (seen as more severe, because of frame loss and the activation of error concealment strategies at the decoder) and the losses in the enhancement layers (seen as affecting only the total quality of the given frame, in the absence of temporal prediction encoding in the higher video layers). We validate the distortion model with streaming experiments in Section VI.

Additionally, we assume that the server S can protect each media layer against transmission errors, with one of E systematic forward error correction schemes $FEC(n_e, k_e)$, $e = 1 \dots E$. The loss probability for each video layer j, protected by $FEC(n, k)$ can be computed starting form the total error probability affecting the transmission process p (considered as an iid random variable). Let p_j be the error probability affecting video layer j, after FEC decoding. It can be computed as the average probability of loosing i video packets from the FEC block ($1 \leq i \leq k$), and at least $\lfloor n - k - i + 1 \rfloor$ redundant packets [16].

$$p_j = \frac{1}{k} \cdot \sum_{i=1}^{k} i \cdot p_i(n, k),$$

where $p_i(n, k)$ is the probability of loosing at least $n - k + 1$ packets from the FEC block, out of which, i packets are video packets. For an iid loss process, $p_i(n, k)$ can be easily computed:

$$p_i(n, k) = \binom{k}{i} p^i (1 - p)^{k-i} \sum_{l=\lfloor f+1-i \rfloor}^{f} \binom{f}{l} p^l (1 - p)^{f-l},$$

where $f = n - k$.

We define $\mathcal{P}_M = \{\rho_m : 1 \leq m \leq O\}$ as the set of available streaming modes, where $O = L \cdot E$ represents the total number of feasible combinations between the media encoded layers and FEC schemes, and ρ_m is the total rate imposed by mode m. The final perceived quality at the end user depends on the number of media layers transmitted, and the loss process that affects the media packets after FEC decoding, according to the distortion model proposed above.

Finally, we model the voice and data download applications. We consider N_V available voice transcoders at the server S. Each transcoder v is characterized by its encoding rate ρ_v. We define $\mathcal{P}_V = \{\rho_v : 1 \leq v \leq N_V\}$ as the available parameter set for the voice application. The perceived quality of the voice application at the end client depends on the complexity of the transcoder v, and hence the allocated rate ρ_v, and the error process p that affects the data transmission. We also assume $\mathcal{P}_F = \{\rho_f : 1 \leq f \leq N_F\}$ as the available parameter set for the FTP application. ρ_f represents the download rate of the FTP session. The perceived quality of the application will depend on the total download time, hence on the allocated download rate and error process that affects the data transmission.

We define the QoS metric $\Gamma(\mathcal{M}_i) = f(r_i, p_i)$ as a function of the allocated rate r_i and the average loss probability p_i affecting the data transmission of application k, towards user i. A concrete example of such a QoS metric, along with the appropriate mappings between this metric and the perceived quality of the applications presented above is given in Section VI. Finally, we define $\mathcal{M} = \{\mathcal{M}_i : 1 \leq i \leq N\}$ as the global operation mode of the system, when the server S allocates the rate $r_i = \rho_k \in \mathcal{P}_k$ to each active user i, accessing application k.

B. Multiple Networks

Even if the problem formulation proposed here is generic, we constrain ourselves to a scenario with two active networks that relay application data between the server S and user i. Q_Net is a QoS modelled network, characterized by a guaranteed service to all active users when network loads are inferior to the congestion point (e.g., through spreading codes and transmission time intervals assignment in the case of an HSDPA system), and high blocking probability in saturated regime. Its total resources are characterized by the instantaneous total throughput R^Q, which takes into account the channel conditions of all active users in the network. R^Q is preferentially distributed among active users according to the importance of their accessed application (e.g., HSDPA systems prioritize voice conversations over streaming applications and FTP downloads). R^Q is periodically estimated on time intervals T, possibly with a certain prediction error, which translates into a generally small packet error probability p_i^Q that equally affects all active users.

The second network, BE_Net, is modelled as a Best Effort network that provides services to clients on a first-come-first-serve basis (e.g., a WiFi hotspot). Each active client i in this network can access resources at a maximum data rate R_i^B and is affected by an average loss process p_i^B, over time intervals T. While channel conditions in wireless environments change on very short time scales (e.g., up to a few tens of ms), we assume that R_i^B and p_i^B represent average values computed on larger time scales T (e.g., one to a few seconds), and represent the average channel conditions for user i on the given period T.

Let $[r_i^Q, r_i^B]$ be the rate allocation of user i over the two networks, with $r_i = r_i^Q + r_i^B$. Please observe that application rates $r_i^Q = 0$ or $r_i^B = 0$ imply that user i is inactive in the given network. Finally, let the tuple $\tau_i = [r_i^Q, p_i^Q, r_i^B, p_i^B]$ characterize the application rates and channel conditions for each user i in the two networks. The following resource constraints apply:

$$\sum_{i=1}^{N} r_i^Q \leq R^Q, \qquad \sum_{i=1}^{N} \frac{r_i^B}{R_i^B} \leq 1. \tag{1}$$

for Q_Net and BE_Net respectively. While the first constraint refers to the total available throughput on the Q_Net, the second one refers to the maximum available time for transmission on the downlink at the access point of the BE_Net. Finally, under these conditions, the total error probability that affects the transmission to user i, reads : $p_i = \dfrac{r_i^Q \cdot p_i^Q + r_i^B \cdot p_i^B}{r_i^Q + r_i^B}$.

IV. Network Selection and Rate Allocation Problem

We assume that the server S periodically solves the optimization problem, in full knowledge of the connection parameter tuple τ_i, $\forall i : 1 \leq i \leq N$, and of the application parameter sets $\mathcal{P}_k$, $\forall k \in \{V, M, F\}$. Within each time interval T, we optimize the allocation of network resources among the N clients, with the final goal of maximizing the overall quality of the system. In other words, we are looking for the optimal global operation mode $\mathcal{M}^* = \{\mathcal{M}_i^* : 1 \leq i \leq N\}$ containing the optimal application mode for each client i, where $\mathcal{M}_i^* = r_i^* \in \mathcal{P}_k$, k being the application accessed by client i:

$$\mathcal{M}^* = \arg\max_{\mathcal{M}} \sum_{i=1}^{N} \Gamma(\mathcal{M}_i) \tag{2}$$

under the constraints provided by Eq. (1). A discrete search through all operation modes leads to the solution $\mathcal{M}^*$ with optimal overall QoS. Alternatively, in the next section, we offer a heuristic algorithm that achieves close-to-optimal results with a faster convergence time.

V. Utility Based Rate Allocation Algorithm

In this section we introduce our heuristic approach for solving the rate allocation optimization problem. We build on the utility framework introduced in [4], and present an algorithm that iteratively takes a locally optimal decision on each user's application mode.

Let $\mathcal{P}_k$, $k \in \{V, M, F\}$ be the sets of application modes ordered in increasing order of their required rates, and let $\mathcal{M}_i$ be the allocated mode of user i at a given iteration of our algorithm. We define $i \rightarrow \mathcal{M}_i^{'}$ as the transition of user i to the next application mode $\mathcal{M}_i^{'}$ requiring the next higher application rate $r_i^{'}$. The utility of this transition can be computed as:

$$U_i = \frac{\Gamma(\mathcal{M}_i^{'}) - \Gamma(\mathcal{M}_i)}{r_i^{'} - r_i},$$

and represents the trade-off between the system quality improvement and the extra resources required by user i's transition. During each iteration, the proposed algorithm finds the user i^* that brings the highest utility to the overall system by its transition:

$$i^* = \arg\max_i U_i,$$

The extra resources will be allocated to user i^* starting with the resources of Q_Net. Once the resources of Q_Net are depleted, the algorithm finds a different user j that can free the required resources for user i^*, by reallocating part of its rate $r \leq r_j$ on the other network BE_Net. Let $G(j, r)$ be the operation by which rate $r \leq r_j$ of user j is redirected through BE_Net, and let H_j be the loss in system utility caused by the switch. This operation is performed as long as the overall utility of the system is still improved ($U_i - H_j > 0$), and as long as free network resources still exist in the overall system. The algorithm stops when there are no more free resources in the network system, or when no other possible user transition can bring any improvement in the overall system utility.

Algorithm 1 Utility based rate allocation algorithm

```
Input:
$R_Q$, $p_i^Q$, $R_i^B$, $p_i^B$, $\forall$ user $i$;
$\mathcal{P}_k$, $\forall k \in \{V, M, F\}$, ordered in ascending order of $\rho_k$;
$\mathcal{M}_i = 0$, $\forall$ user $i$;
Output:
Global Rate Allocation Mode $\mathcal{M}$;
Procedure RateAllocation
While (1)
for $i = 1$ to $N$ do
    Compute the utility of $i \rightarrow \mathcal{M}_i^{'}$:
    $U_i = \frac{\Gamma(\mathcal{M}_i^{'}) - \Gamma(\mathcal{M}_i)}{r_i^{'} - r_i}$;
end for
find $i^* = arg\max_i U_i$;
Push($i^*, \mathcal{M}_{i^*}^{'}, Q\_Net$);
Procedure Push($i, \mathcal{M}_i^{'}, Q\_Net$)
if Q_Net has enough free resources then
    $i \rightarrow \mathcal{M}_i^{'}$;
    update free resources on Q_Net;
else
    Switch($i, \mathcal{M}_i^{'}, Q\_Net$);
end if
Procedure Switch($i, \mathcal{M}_i^{'}, Q\_Net$)
find user $j$ that can transfer part of his allocated rate $r_j$
to BE_Net with minimum $H_j$;
if $U_i - H_j > 0$ then
    perform the switch of user $j$ rate: $G(j, r)$;
    $i \rightarrow \mathcal{M}_i^{'}$;
    update free resources on Q_Net and BE_Net;
else
    Break;
end if
```

`Algorithm 1` represents a sketch of the proposed algorithm. The **Push** procedure always attempts to increase the

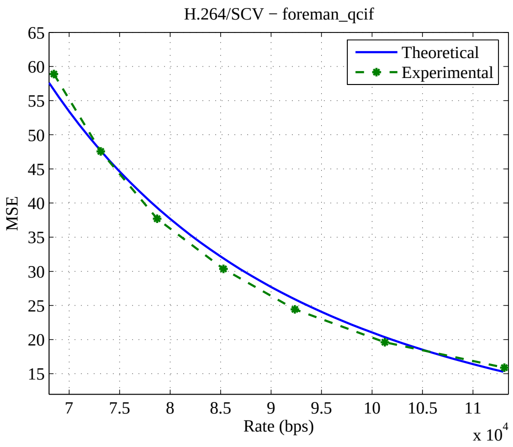

Fig. 2. Video Model Validation - Source Distortion: H264SVC encoder, $foreman\_qcif$, 30 fps, one BL and one EL, $\alpha = 4.41 \cdot 10^4$, $\xi = -1.34515$.

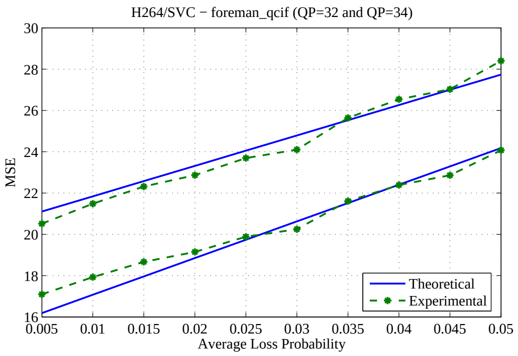

Fig. 3. Video Model Validation - Loss Distortion: H264SVC encoder, $foreman\_qcif$, 30 fps, one BL and one EL, $\beta = 147$.

system's utility by allocating the free Q_Net resources to the best user. If the free resources are not enough, the **Switch** procedure tries to find a new user that can free up enough resources by reallocating parts of its allocated rate through the BE_Net. As long as the network resources allow it, the procedures repeat until no higher modes are available at any client, or no extra utility improvement can be brought to the overall system.

The complexity involved in the search for i^* is $O(N)$, the same being valid for the **Switch** procedure. In the worst case, the algorithm requires $O(N \cdot |\mathcal{P}_k|)$ iterations to pass through every application mode of every user. Hence the total complexity of the algorithm is $O(N^2 \cdot |\mathcal{P}_k|)$. For a reasonable number of wireless users, and a finite set of available application modes, the algorithm will converge rapidly to a global rate allocation vector $\mathcal{M}$. Its performance is further studied in Section VII.

VI. Video Model Validation and MOS Quality Metric

In this section we validate the distortion model introduced in Section III-A, and we exemplify on a concrete quality metric Γ based on the MOS (Mean Opinion Score) value.

First we encode the $foreman\_qcif$ sequence (300 frames, 30 frames per second) in one base layer (BL) and one enhancement layer (EL), with the help of the H.264/SVC encoder. The total rate of the encoded sequence is varied, by encoding at different quantization parameters (QP) for the BL. The encoder always uses a QP for the EL, 6 points below the QP of the BL. We are considering one network packet per frame and per video layer. On the sequence of packets we are inflicting transmission packet losses according to an independent loss probability $p \in [0, 0.05]$, and we compare the decoded video quality with the original one, by averaging over 100 simulation runs. Results for the validation of the source distortion are presented in Figure 2, while Figure 3 presents the validation of the loss distortion model. We observe that the model closely follow the experimental results[1].

Next, we introduce the quality metric based on MOS. MOS reflects the average user satisfaction on a scale of 1 to 4.5. The minimum value reflects an unacceptable application quality, and the maximum value refers to an excellent QoS. The perceived quality of each of the three applications is converted into an equivalent MOS value, which is later used in the optimization problem.

The performance of different voice transcoders as a function of network losses is mapped to MOS values using the $PESQ$ algorithm on a representative set of voice samples [15] in Figure 4. We observe that, while good network conditions lead to increased user experience, high packet error rates degrade the perceived quality of the voice communication.

The perceived media streaming quality is initially mapped into an MSE (mean square error) distortion measure, as presented in Section III-A. Later on, a nonlinear mapping between MSE and MOS values is used, as illustrated in Figure 5.

Finally, the perceived quality of the FTP application is mapped to MOS values according to a logarithmic function of the achieved throughput: $MOS = a \cdot \log(b \cdot r(1-p))$. The variables a and b are system dependent parameters, and can be set by the network operator (Figure 6).

VII. Simulation Results

A. Simulation Setup

We test the performance of our proposed rate allocation and path selection method, and we compare its performance against a classic optimization solution that uses application throughput as a quality metric.

We use 4 voice transcoders, namely G.723.1B, iLBC, SPEEX and G.711 with average encoding rates of 6.4, 15.2, 24.6 and $64kbps$ respectively. To simulate the media streaming application, we encode the $foreman\_qcif$ sequence (300

[1] For a complete validation of the video distortion model please see [17].

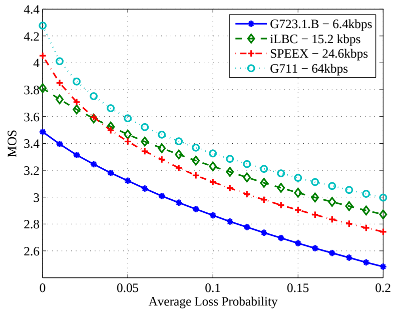

Fig. 4. Voice Application MOS

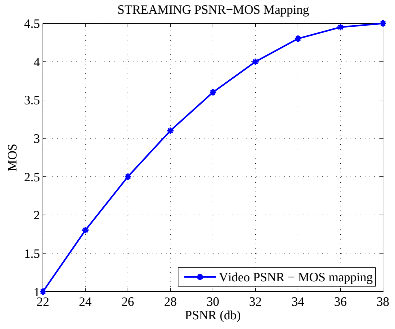

Fig. 5. Streaming Application MOS

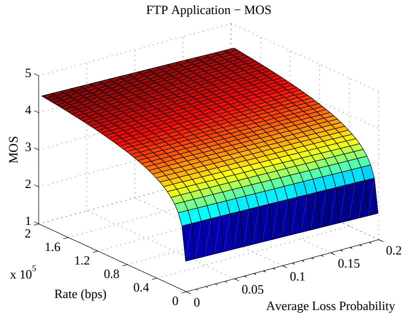

Fig. 6. FTP Application MOS

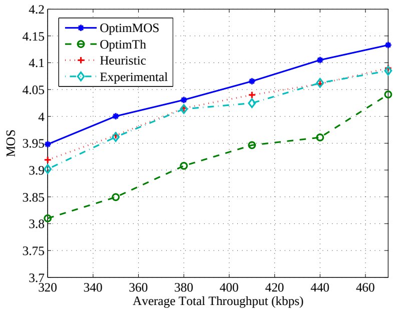

Fig. 7. Average system MOS values: MOS vs. Throughput Optimization

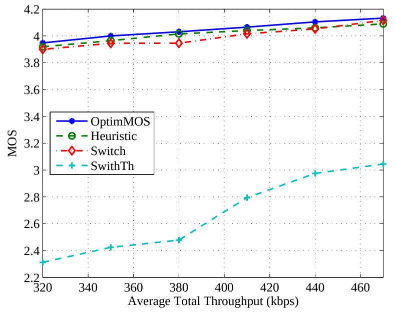

Fig. 8. Average system MOS values: Heuristic algorithms.

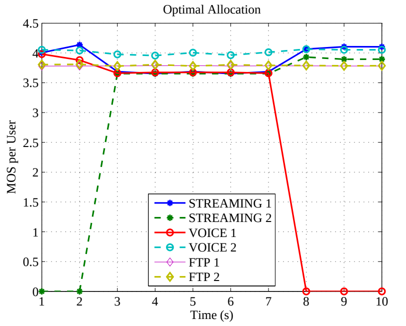

Fig. 9. Client performance when users are added/removed to/from the system: $OptimMOS$ algorithm.

frames) with the H.264/SVC codec. We encode one base layer and one enhancement layer, each of $70kbps$. Additionally, we use one forward error correction mode $FEC(20, 17)$ which can correct up to 3 packet errors in a block of 20 packets. For FTP downloads, we set 4 available download rates of 50, 100, 150 and $200kbps$ respectively.

Due to the high complexity of the full search algorithm for finding the overall optimal rate allocation solution, we use small network scenarios (5 or 6 users) in order to validate the MOS quality metric, and the proposed heuristic algorithm. Later we compare our proposed heuristic algorithm with other heuristics in larger network setups. For comparison purposes we define as $OptimMOS$ and $OptimTh$ the full search algorithms which optimize the network resource allocation based on the MOS, and respectively $Throughput$ QoS metrics. In the same time we define `Algorithm 1` as $Heuristic$, while $Switch$ represents the same heuristic algorithm, with the constraint that no user can be allocated resources from both networks in the same time (e.g., when the algorithm decides to switch one client from one network to another, its whole allocated rate is rerouted through the new network). $SwitchTh$ is similar to $Switch$, but acts according to the $Throughput$ QoS metric.

B. Small Network Scenarios

A total of 6 clients are placed in the coverage area of both networks (3 voice, 2 FTP, and one streaming user). Server S performs the optimization of the rate allocation periodically, every $T = 1s$. The average throughput R^Q of Q_Net varies in the interval $[100, 150]kbps$ and the prediction error p_i^Q is kept around 1%. The connection data rate R_i^B of the users in the BE_Net is set in the interval $[220, 310]kbps$, and the individual average loss probabilities p_i^B are randomly chosen in the interval $[1, 15]\%$. We average our results over 100 simulation runs of 10 seconds each.

We first compare the average performance of the overall system, when the optimization is performed according to the MOS and throughput quality metrics. We start by identifying the traffic distribution obtained by each optimization metric over the two networks. Table I presents the fraction of traffic that passes through both networks, for each application. We observe that the MOS optimization rightfully uses the Q_Net resources for the voice and streaming applications, while the FTP traffic is forwarded through BE_Net. On the other hand, the throughput optimization favors the FTP application, as it forwards part of its traffic over Q_Net (hence increasing the offered rate for the application), at the expense of lower available resources for the voice and streaming applications that share the same network. This explains the lower overall system performance obtained for the throughput metric, compared to MOS (Figure 7). For a total average system throughput varying from 320 to $460kbps$, the MOS optimization outperforms the throughput optimization in most cases by as much as 0.15 MOS points. We also observe that the $Heuristic$ algorithm closely matches the optimal behavior, and the experimental

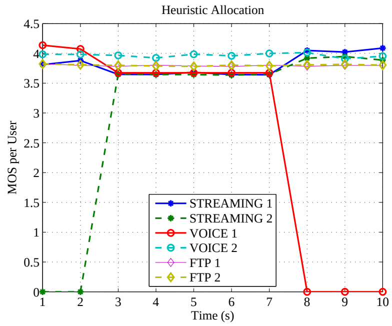

Fig. 10. Client performance when users are added/removed to/from the system: $Heuristic$ algorithm.

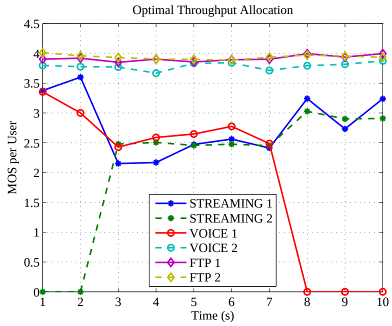

Fig. 11. Client performance when users are added/removed to/from the system: $OptimTh$ algorithm.

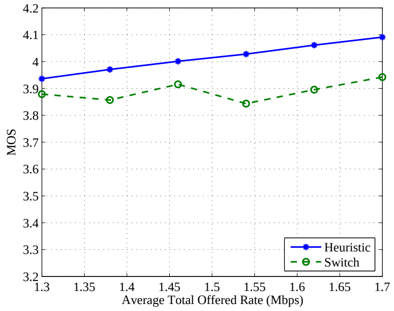

Fig. 12. Average system MOS values: $Heuristic$ vs. $Switch$, 20 users.

results obtained after performing experiments with real video sequences. In the same time, Figure 8 presents the quality performance among the proposed heuristic algorithms. While $Switch$ and $Heuristic$ are quite close to optimum, $SwitchTh$ fails to allocate enough resources to some of the users, hence the important degradation in overall system performance.

TABLE I

TRAFFIC DISTRIBUTION OVER THE TWO NETWORKS (IN %).

Application	MOS Optimization		Throughput Optimization	
	Q_Net	BE_Net	Q_Net	BE_Net
Voice	100	0	100	0
Streaming	88.5	11.5	94	6
FTP	1	99	12	88

Finally, we test the two optimization metrics in dynamic systems where users are allowed to join/leave the networks. We start with 5 clients (2 voice, 1 streaming and 2 FTP users). At time $t = 3s$ we add a streaming user, and at time $t = 8s$ we remove one voice user. Figure 9, Figure 10 and Figure 11 present the average application performance for each user. We observe that in the case of MOS optimization, the system is able to cope with the extra user at the expense of a small quality degradation for the existing users, for both $OptimalMOS$ and $Heuristic$ algorithms. On the other hand, the throughput optimization is unfair, as some of the clients are penalized more than the others, and the overall performance is worse.

C. Large Network Scenarios

In this case we are using a total of 20 clients placed in the coverage area of both networks (7 voice, 6 streaming and 7 FTP clients). The total rate of the system is varied in the interval $[1.3, 1.7]Mbps$ with $R^Q \in [300, 600]kbps$. The loss probabilities for the two networks and the simulation setup are similar as in the previous example.

We are looking at the overall average performance of the $Heuristic$ and $Switch$ algorithms when more active users are present in the system (Figure 12). Intentionally, we omit the performance of the $SwitchTh$ algorithm, due to its very poor results. We observe that while $Switch$ performs quite good, $Heuristic$ still provides a significant improvement in total system quality. This is mainly due to the extra system granularity in allocating the resources of the two networks among the clients, if clients are allowed to connect in parallel to both networks.

Next, we present the average traffic distribution on the two networks, for each type of application, when each of the two algorithms is used to compute the overall rate allocation. Figure 13 and Figure 14 present the distributions obtained by the $Heuristic$ and respectively $Switch$ algorithms. We observe that $Heuristic$ manages to allocate the Q_Net resources mostly to the voice application and as much as possible to the streaming application. The FTP clients are mostly scheduled on BE_Net, which represents an intuitive result. On the other hand, $Switch$ schedules almost half of the voice applications on the BE_Net, at the advantage of streaming applications. While surprising, this result is explained by the fact that voice applications, usually requiring less network resources, are easier to switch on the best-effort network, when the QoS network becomes congested. Such a behavior can however be corrected by applying different weights to the clients, depending on the importance of the accessed application.

Finally, we test our algorithms in dynamic systems. We allow 4 new users to join the system at time $t = 3s$ (2 voice, 1 streaming and 1 FTP clients), while at time $t = 8s$, other 4 users area leaving. Figure 15 and Figure 16 present the results obtained by $Heuristic$ and $Switch$ respectively. In the first case, we observe that the algorithm manages to keep a rather constant application quality for all active clients, by redistributing parts of the network resources to the new users. This way, $Heuristic$ achieves fairness among all users, even if they access different types of applications. On the other hand, $Switch$ copes worse with the system dynamics; we observe that the voice and streaming users are penalized, compared to the FTP users. Again, this is due to the lack of granularity in reallocating network resources, when new users enter the system. This highlights the benefit of resource allocation flexibility given by the multipath network scenario assumed by the proposed algorithm.

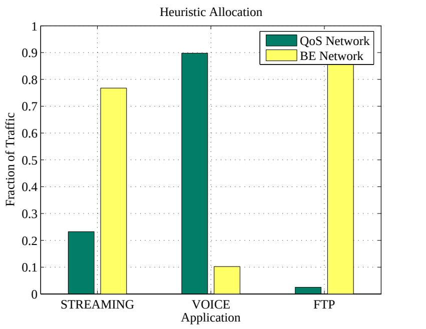

Fig. 13. Average traffic distribution per application type, per network: $Heuristic$ algorithm, 20 users.

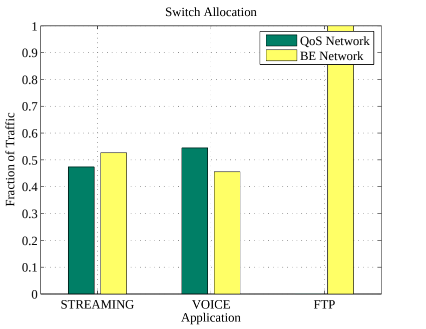

Fig. 14. Average traffic distribution per application type, per network: $Switch$ algorithm, 20 users.

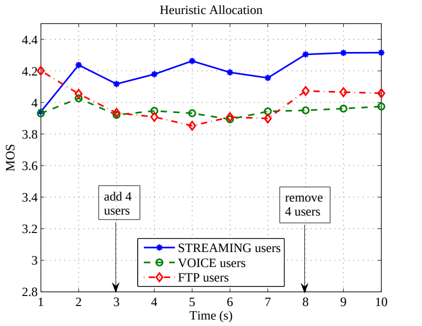

Fig. 15. Average performance per application in case users join/leave the network: $Heuristic$ algorithm.

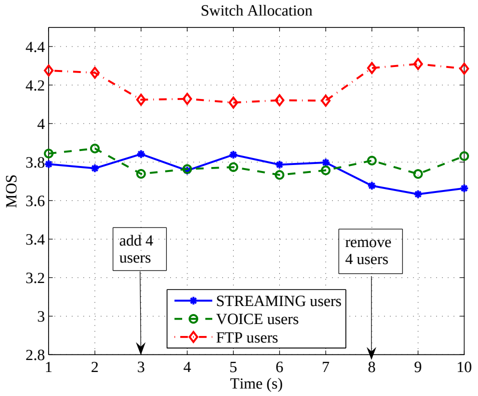

Fig. 16. Average performance per application in case users join/leave the network: $Switch$ algorithm.

VIII. Conclusions

We introduce a new rate allocation and network selection optimization framework for clients accessing multiple applications over parallel networks. In the optimization process we take into account the available network resources and the connection parameters of clients, along with the specific quality requirements of each application. We unify the performance of all applications under a single MOS quality metric, which is later used in the optimization process. Compared to traditional optimization metrics based on throughput, the MOS approach achieves a more fair resource allocation among active clients, and proves to be more scalable in dynamic systems. We finally provide a heuristic algorithm based on utility functions, which achieves a close to optimal resource allocation with low computational resources. Comparing to other heuristic approaches, our algorithm is more stable and adaptable in dynamic situations, emphasizing the benefit of resource aggregation in multipath network scenarios. The obtained results encourage us to further investigate the possibility of multiple wireless networks interconnecting towards the final benefit of the end users.

References

[1] Swisscom Mobile Unlimited UMTS/GPRS/WLAN. http://www.swisscom-mobile.ch/scm/gek_mobile-unlimited-en.aspx.

[2] K. Ahmavaara and H. Haverinen and R. Pichna. Interworking Architecture between 3GPP and WLAN Systems. *IEEE Communications Magazine*, pages 74–81, November 2003.

[3] D. Jurca and P. Frossard. Media-Specific Rate Allocation in Multipath Networks. *IEEE Transactions on Multimedia*, 2006. accepted for publication.

[4] F. Kelly and T. Voice. Stability of End-to-End Algorithms for Joint Routing and Rate Control. *ACM SIGCOMM Computer Communcation Review*, 35(2):5–12, April 2005.

[5] T. Stockhammer, M. Hannuksela, and T. Wiegand. H.264/AVC in Wireless Environments. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(7):657–673, July 2003.

[6] M. van der Schaar, S. Krishnamachari, S. Choi, and X. Xu. Adaptive Cross-Layer Protection Strategies for Robust Scalable Video Transmission over 802.11 WLANs. *IEEE Journal on Selected Areas in Communications*, 21(10):1752–1763, December 2003.

[7] M. Rossi, F. H. P. Fitzek, and M. Zorzi. Error Control Techniques for Efficient Multicast Streaming in UMTS Networks: Proposals and Performance Evaluation. In *Proceedings of SCI*, 2003.

[8] S. A. Malik and D. Zeghlache. Resource Allocation for Multimedia Services on the UMTS Downlink. In *Proceedings of IEEE International Conference on Communication*, volume 5, pages 3076–3080, 2002.

[9] 3GPP 3rd Generation Partnership Project. Feasability study on 3GPP system to Wireless Local Area Network WLAN interworking - Release 6. Technical Report TR 22.934, 2003.

[10] X. G. Wang, J. Mellor, and K. Al-Begain. Towards Providing QoS for Integrated Cellular and WLAN Networks. In *Proceedings of PGNET*, 2003.

[11] M. van der Schaar and S. Shankar. Cross-Layer Wireless Multimedia Transmission: Challanges, Principles, and New Paradigms. *IEEE Wireless Communications*, 12(4):50–58, August 2005.

[12] W. Kellerer, L.-U Choi, and E. Steinbach. Cross-Layer Adaptation for Optimized B3G Service Provisioning. In *Proceedings of the 6th Intl. Symposium on Wireless Personal Multimedia Communications WPMC, Japan*, October 2003.

[13] S. Shakkottai, T. S. Rappaport, and P. C. Karlsson. Cross-Layer Design for Wireless Networks. *IEEE Communications Magazine*, 41(10):74–80, October 2003.

[14] M. Ivrlac and J. Nossek. Cross Layer Design - An Equivalence Class Approach. In *Proc. IEEE International Symposium on Signals, Systems, and Electronics*, 2004.

[15] S. Khan, S. Duhovnikov, E. Steinbach, M. Sgroi, and W. Kellerer. Application-driven cross-layer optimization for mobile multimedia communication using a common application layer quality metric. In *Proceedings of Second International Symposium on Multimedia over Wireless, ISMW*, July 2006.

[16] D. Jurca and P. Frossard. Optimal FEC Rate for Media Streaming in Active Networks. In *Proceedings of IEEE ICME*, July 2004.

[17] A. Jovanovic. Media Aware Rate Allocation and FEC Protection of Streaming Video in Multipath Networks. Master's thesis, EPFL, March 2007.