Rate Allocation for Multi-User Video Streaming over Heterogenous Access Networks

Xiaoqing Zhu*, Piyush Agrawal†, Jatinder Pal Singh‡, Tansu Alpcan‡ and Bernd Girod*

*Information Systems Laboratory, Stanford University Stanford, CA 94305, U.S.A.
†Department of Computer Science and Engineering, Indian Institute of Technology, Kanpur, India
‡Deutsche Telekom Laboratories, Technische Universität Berlin Ernst-Reuter Platz 7, 10587, Germany

*{zhuxq,bgirod}@stanford.edu †piyushag@iitk.ac.in,
‡{jatinder.singh,tansu.alpcan}@telekom.de

ABSTRACT

Contemporary wireless devices integrate multiple networking technologies, such as cellular, WiMax and IEEE 802.11a/b/g, as alternative means of accessing the Internet. Efficient utilization of available bandwidth over heterogeneous access networks is important, especially for media streaming applications with high data rates and stringent delay requirements. In this work we consider the problem of rate allocation among multiple video streaming sessions sharing multiple access networks. We develop and evaluate an analytical framework for optimal video rate allocation, based on observed available bit rate (ABR) and round trip time (RTT) over each access network, as well as the video distortion-rate (DR) characteristics. The rate allocation is formulated as a convex optimization problem that minimizes the sum of expected distortion of all video streams. We then present a distributed approximation of the optimization, which enables autonomous rate allocation at each device in a media- and network-aware fashion. Performance of the proposed allocation scheme is compared against robust rate control based on H^∞ optimal control and two heuristic schemes employing TCP-style additive-increase-multiplicative-decrease (AIMD) principles. We simulate in NS-2 [1] simultaneous streaming of multiple high-definition (HD) video streams over multiple access networks, using ABR and RTT traces collected on Ethernet, IEEE 802.11g, and IEEE 802.11b networks deployed in a corporate environment. In comparison with heuristic AIMD-based schemes, rate allocation from both the media-aware convex optimization scheme and H^∞ optimal control benefit from proactive avoidance of network congestion, and can reduce the average packet loss ratio from 27% to below 2%, while improving the average received video quality by 3.3 - 4.5 dB in PSNR.

MM'07, September 23–28, 2007, Augsburg, Bavaria, Germany.

Categories and Subject Descriptors

C.2 [**Computer-Communication Networks**]: Distributed Systems

General Terms

Design, performance

Keywords

distributed rate allocation, video streaming, heterogeneous access network

1. INTRODUCTION

The widespread acceptance and deployment of infrastructure for fixed-line, wireless, and mobile access to the Internet enables opportunistic Internet connectivity via a multitude of access technologies. The resulting aggregate transmission capabilities of the multi-homed end-user devices can be utilized for better quality of service (QoS) provisioning for otherwise bandwidth constrained media applications.

Recent years have witnessed increasing efforts towards standardization of architectures for convergence of heterogeneous access networks. Integration of heterogeneous networks is part of the 4G network design [2]. IEEE 802.21 [3] is delineating a framework to enable handovers and interoperability between heterogeneous wireless and wireline networks. The IP Multimedia Subsystems (IMS) platform [4] has defined an overlay architecture for providing multimedia services on top of heterogeneous networks.

While platforms and architectures supporting convergence can allow high-bandwidth video streaming applications (e.g., HDTV over Internet) to benefit from simultaneous connectivity to multiple access networks, distributed rate allocation policies have to be designed for suitable application metrics and efficient network utilization. Access networks differ in their attributes such as available bit rates (ABRs) and round trip times (RTTs), which also vary with time. On the other hand, video streaming applications differ in their latency requirements and distortion-rate (DR) characteristics. For instance, streaming a high-definition (HD) video sequence containing dynamic scenes from an action movie would require much higher data rate to achieve the same quality as that needed for streaming a static head-and-shoulder news clip for a mobile device with a low-resolution display. Video

streaming applications also require timely delivery of each packet to ensure continuous media playout. Late packets are typically discarded at the receiver, causing drastic quality degradation of the received video due to error propagation at the decoder.

In this work, we propose, evaluate, and compare distributed rate allocation policies for video streaming over heterogeneous networks, with and without awareness of media and network characteristics. For the case where devices have access to both the video DR characteristics, and network ABRs and RTTs, we formulate the rate allocation problem in a convex optimization framework to minimize the sum of expected distortions of all participating video streams. A distributed approximation to the optimization is presented, to enable autonomous rate allocation at each device in a media- and network-aware fashion. To address the scenario where media-specific information is not accessible by the devices, we propose a scheme based on H^∞ optimal control. The scheme achieves optimal bandwidth utilization on the access networks by guaranteing a lower bound on a cost function that models the deviation of rate allocated to a stream from the rate available on a network. We compare the above policies with simple heuristic-based rate allocation schemes that assign rates to streams on different networks in accordance with the available bit rate on the networks, with the total rate of each stream following the TCP-style additive-increase-multiplicative-decrease (AIMD) principle [5].

We evaluate the performance of the above rate allocation policies with NS-2 [1], using ABR and RTT traces collected from Ethernet, IEEE 802.11b and IEEE 802.11g networks. Simulation results are presented for the application scenario of simultaneous streaming of multiple high-definition (HD) video sequences over multiple access networks. We demonstrate that rate allocation based on media-aware convex optimization and H^∞ optimal control can achieve significantly lower packet delays and loss rates (less than 0.1 % for the media-aware allocation, and between 0.5 % and 1.9 % for H^∞ optimal control), whereas heuristic AIMD-based schemes incur packet losses of up to 27 %. As a consequence, the media-aware allocation can improve average received video quality by 3.3 - 4.5 dB in PSNR, and tends to assign higher rates for the more demanding video sequence by reducing allocation to easier-to-encode sequences. It therefore achieves more balanced video quality among the streams, and tends to allocate resource more evenly among the available access networks.

A review of related work in rate control and multi-flow, multi-network resource allocation is provided in the next section. We then present our system model of the access networks and expected video distortion in Section 3, followed by descriptions of the rate allocation schemes (media-aware convex optimization, H^∞ optimal control, and AIMD-based heuristics) in Section 4. Performance evaluation and comparison of the schemes are discussed in Section 5, for simulations of three HD video sequences simultaneously streaming over three access networks.

2. RELATED WORK

Rate allocation among multiple traffic flows over shared network resources is an important and well-studied problem. Internet applications typically use the TCP Congestion Control mechanism for regulating the outgoing rate [5] [6]. For media streaming applications over UDP, TCP-Friendly Rate Control (TFRC) is a popular choice [7] [8], and several modifications have been proposed to improve its media-friendliness [9]. Rate allocation to flows with different utilities has also been studied in the mathematical framework proposed by [10]. The authors also present distributed rate allocation algorithms based on a pricing mechanism between the source and relaying agents. In our work, the notion of utility of each traffic flow corresponds to its expected received video quality, measured in terms of mean-squared-error (MSE) distortion from the original uncompressed video signals. The mathematical framework has also been extended, to consider rate allocation over multiple networks simultaneously.

The problem of efficient utilization of multiple networks via suitable allocation of traffic flows has also been explored from different perspectives. A game-theoretic framework to allocate bandwidth for elastic services in networks with fixed capacities has been addressed in [11–13]. Our work, in comparison, attempts to address the time-varying nature of wireless networks as well, by dynamically tracking the available bit rate and delay over each network, and updating the allocation results accordingly. In [14], a cost price mechanism is proposed, to enable a mobile device to split its traffic amongst several IEEE 802.11 access points based on throughput obtained and price charged. However, the work does not take into account the existence of heterogeneous networks or the characteristics of traffic, nor does it specify an operational method to split the traffic.

Rate adaptation of multimedia streams has been studied in the context of heterogeneous networks in [15], where the authors propose an architecture to allow online measurement of network characteristics and video rate adaptation via transcoding. The rate control algorithm, on the other hand, is based on TFRC and unaware of the media content. In [16], media-aware rate allocation is achieved, by taking into account the impact of both packet loss rates and available bandwidth over each link, on the end-to-end video quality of a single stream, whereas in [17], the rate allocation problem has been formulated for multiple streams sharing one wireless network. Unlike our recent work, where the multi-stream multi-network rate allocation problem is addressed from the perspective of stochastic control of Markov Decision Processes [18] and robust H^∞ optimal control of linear dynamic systems [19], in this paper we stay within the convex optimization framework for media-aware optimal rate allocation, and compare the performance of the scheme with prior approaches.

3. SYSTEM MODEL

In this section, we introduce the mathematical notations used for modeling the characteristics of the access networks, and for estimating expected received distortion of each video stream. We envision a middleware functionality as depicted in Fig. 1, which collects characteristic parameters of both the access networks and video streams, and performs the optimal rate allocation according to one of the schemes described in Section 4.

3.1 Network Model

Consider a set of access networks $\mathcal{N} = \{1, 2, \ldots, N\}$, simultaneously available to multiple devices. Each access network n is characterized by its available bit rate ABR_n and

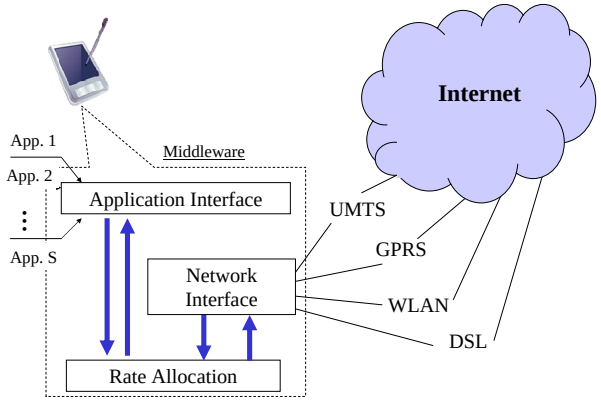


Figure 1: Middleware functionality in a device. The rate allocation module collects the observed media statistics and network characteristics (e.g., ABR and RTT), and dictates the rate allocation among application streams, over each network interface.

round trip time RTT_n, which are measured and updated periodically. For each device, the set of video streams is denoted as $\mathcal{S} = \{1, 2, \ldots, S\}$. Traffic allocation can be expressed in matrix form: $\mathbf{R} = \{R_n^s\}_{S \times N}$, where each element R_n^s corresponds to the allocated rate of Stream s to Network n. Consequently, the total allocated rate over Network n is $R_n = \sum_s R_n^s$, and the total allocated rate for Stream s is $R^s = \sum_n R_n^s$. We denote the *residual bandwidth* over Network n as:

$$B_n = ABR_n - \sum_{s \in \mathcal{S}} R_n^{s'} = ABR_n - R_n. \quad (1)$$

From the perspective of Stream s, the observed available bandwidth is:

$$ABR_n^s = ABR_n - \sum_{s' \neq s} R_n^{s'}. \quad (2)$$

Note that $B_n = ABR_n - R_n = ABR_n^s - R_n^s$.

As the allocated rate on each network approaches the maximum achievable rate, average packet delay typically increases due to network congestion. We use a simple fractional function to approximate the non-linear increase of packet delay with traffic rate over each network, as:

$$T_n = \frac{\alpha_n}{B_n} = \frac{\alpha_n}{ABR_n - R_n} = \frac{\alpha_n}{ABR_n^s - R_n^s}, \quad (3)$$

which is reminiscent of the classical M/M/1 queuing model [20]. The value of α_n is estimated from past observations of RTT_n and B_n:

$$\alpha_n = \frac{B_n RTT_n}{2}, \quad (4)$$

assuming equal delay on both directions.

3.2 Video Distortion Model

Expected video distortion at the decoder comprises of two terms:

$$D_{dec} = D_{enc} + D_{loss}, \quad (5)$$

where D_{enc} denotes the distortion introduced by quantization at the encoder, and D_{loss} represents the additional distortion caused by packet loss [21].

Typically, the distortion-rate (DR) characteristic of the encoded video stream can be fit to a parametric model [21]:

$$D^s(R^s) = D_0^s + \frac{\theta^s}{(R^s - R_0^s)}, \quad (6)$$

where the parameters D_0^s, θ^s and R_0^s depend on the coding scheme and the content of the video. They can be estimated from three or more trial encodings using non-linear regression techniques. To allow fast adaptation of the rate allocation to abrupt changes in the video content, these parameters need to be updated for each group of pictures (GOP) in the encoded video sequence, typically once every 0.5 second.

The distortion introduced by packet loss due to transmission errors and network congestion, on the other hand, can be derived from [22] as:

$$D_{loss}^s = \kappa^s P_{loss}^s, \quad (7)$$

where the sensitivity factor κ^s reflects the impact of packet losses P_{loss}^s, and depends on both the video content and its encoding structure. For simplicity, we assume in the rest of the paper that random packet losses due to transmission errors are remedied at the lower layers (e.g., MAC-layer retransmissions and PHY-layer channel coding). In this case, P_{loss}^s comprises solely of packet late losses due to network congestion.

4. DISTRIBUTED RATE ALLOCATION

In this section, we address the problem of rate allocation among multiple streams over multiple access networks with several alternative approaches. We first present a convex optimization formulation of the problem in Section 4.1, and explain how to approximate the media- and network-aware optimal solution with decentralized calculations. In the case that video DR characteristics are unavailable, we resort to a formulation of H^∞-optimal control in Section 4.2, which dynamically adjusts the allocated rate to each stream according to fluctuations in observed network available bandwidth. As a basis of comparison, we investigate in Section 4.3 two heuristic allocation schemes following the TCP-style additive-increase-multiplicative-decrease (AIMD) principle.

4.1 Media-Aware Allocation

We seek to minimize the total expected distortion of all video streams sharing multiple access networks:

$$\min_{\mathbf{R}} \quad \sum_s D_{dec}^s(R^s, P_{loss}^s) \quad (8)$$

$$s.t. \quad R^s = \sum_n R_n^s, \quad \forall s \in \mathcal{S} \quad (9)$$

$$R_n = \sum_s R_n^s < ABR_n, \quad \forall n \in \mathcal{N} \quad (10)$$

$$R_n^s = \rho_n R^s, \quad \forall n \in \mathcal{N}. \quad (11)$$

In (8), the expected distortion D_{dec}^s is a function of the allocated rate R^s and average packet loss P_{loss}^s according to (5). The constraints in (11) are introduced to impose uniqueness of the optimal solution. We choose $\rho = ABR_n / \sum_{n'} ABR_{n'}$ to ensure balanced utilization over each interface:

$$\frac{R_n}{ABR_n} = \frac{\sum_s R_n^s}{ABR_n} = \frac{\rho_n \sum_s R^s}{ABR_n} = \frac{\sum_{n'} R_{n'}}{\sum_{n'} ABR_{n'}}, \forall n \in \mathcal{N}. \quad (12)$$

It can also be shown that $\rho_n = ABR_n^s / \sum_n ABR_n^s, \forall s \in \mathcal{S}$. Each stream can therefore calculate ρ_n independently, based on its own observation of ABR_n^s for each network n.

The average packet loss P^s_{loss} for each stream is the weighted sum of packet losses over all networks:

$$P^s_{loss} = \sum_n \rho_n e^{-T^s_0/T_n}, \tag{13}$$

Following the derivations in [22], the percentage of late packets is estimated as $e^{-T^s_0/T_n}$ based on an exponential approximation of the packet delay distributions, with average delay of T_n over Network n and playout deadline T^s_0 for Stream s. Given (3), one can express P^s_{loss} as:

$$P^s_{loss} = \sum_n \rho_n e^{-T^s_0(ABR_n - R_n)/\alpha_n}. \tag{14}$$

Combining (5)-(14), it can be verified that the optimization objective is a convex function of the variable matrix $\mathbf{R}$. If all the observations and parameters were available in one place, the solution could be found by any suitable convex optimization method [23].

We desire to minimize the objective (8) in a distributed manner, with as little exchange of information among the devices as possible. One approach is to consider the impact of network congestion on one stream at a time, and alternate between streams until convergence. From the perspective of a single stream s, its contribution to (8) can be rewritten as:

$$\begin{aligned} \min_{R^s} \quad & D^s(R^s) + \kappa^s \textstyle\sum_n \rho_n e^{-T^s_0(ABR^s_n - R^s_n)/\alpha_n} \\ & + \textstyle\sum_n \sum_{s' \neq s} \rho_n \kappa^{s'} e^{-T^{s'}_0(ABR^s_n - R^s_n)/\alpha_n} \\ s.t. \quad & R^s_n = \rho_n R^s, \forall n \in \mathcal{N} \\ & R^s_n < ABR^s_n, \forall n \in \mathcal{N}. \end{aligned} \tag{15}$$

In (15), optimization of rate allocation for Stream s would require knowledge of its own distortion-rate function $D^s(R^s)$, its own loss sensitivity κ^s, but also its impact on the late loss of other streams via $\kappa^{s'}$ and $T^{s'}_0$. While each video stream can locally obtain information regarding its own loss sensitivity and playout deadline, exchange of such information among different video streams is undesirable for a distributed solution.

We therefore further simplify the optimization to:

$$\begin{aligned} \min_{R^s} \quad & D^s(R^s) + \textstyle\sum_n \kappa' \rho_n e^{-T^s_0(ABR^s_n - R^s_n)/\alpha_n} \\ s.t. \quad & R^s_n = \rho_n R^s, \forall n \in \mathcal{N} \\ & R^s_n < ABR^s_n, \forall n \in \mathcal{N}, \end{aligned} \tag{16}$$

where κ' is empirically chosen to control the level of aggressiveness of the allocation. Even though (16) does not necessarily lead to a optimal solution to (8), it nevertheless incorporates the considerations of both network congestion and encoder video distortion in the choice of allocated rates. Influence to the performance of other streams are included implicitly, since congestion introduced over each network is captured in the second term in (16), which would impact all streams traversing that network.

Note that in essence, optimization of (16) involves a one-dimensional search of R^s, thus can be solved efficiently using various numerical methods. In practice, implementation of the distributed allocation scheme would also require each stream to track its observations of ABR^s_n's and RTT_n's over all available access networks, to periodically update its estimate of α_n according to (4), and to determine its rate allocation by finding the optimal R^s for (16) and allocate the rate in proportion to ρ_n over respective networks.

4.2 H$^\infty$-Optimal Allocation

The problem of optimal rate allocation among streams sharing multiple networks with time-varying characteristics can also be addressed using H$^\infty$ optimal control [19]. In this approach, we use a linear state-space system to keep track of current and past observations on available bandwidth of each network, and model the variations as unknown disturbances. We then consider a "worst-case" formulation and let each video stream update its rate using H$^\infty$-optimal control [24]. This allows to treat the dynamics of each stream as independent of the others, thereby decoupling the interaction between different streams. Unlike the previous approach, this scheme does not require RTT observations and is unaware of media-specific knowledge, such as the video DR characteristics.

Each stream can estimate via various online measurement tools [25] the quantity w_n, which is approximately proportional to the residual bandwidth B_n defined in (1). Without loss of generality, we use the following formulation:

$$w_n = \begin{cases} B_n, & \text{if } B_n \geq 0 \\ \kappa\,(t_f - t_i), & \text{if } B_n < 0 \end{cases}, \tag{17}$$

where κ is a negative constant. Here, t_i and t_f denote the initial and final time instance when B_n is negative.

We next define a system from the perspective of a single stream s keeping track of a single network n. The system state x^s_n reflects roughly the ABR on this network. We first focus on the single network case and drop the subscript n to simplify analysis. The case involving multiple access networks is discussed in the Appendix. Since each stream is independent of others in the H$^\infty$ control formulation, the analysis also generalizes immediately to the case with multiple streams. We refer the readers to [19] for further details in the derivations.

The system equation for stream s is:

$$\dot{x}^s = a\,x^s + b\,u^s + w, \tag{18}$$

where u^s represents the control action of the device. The parameters $a < 0$ and $b < 0$ adjust the memory horizon (the smaller a the longer the memory) and the "expected" effectiveness of control actions, respectively, on the system state x^s. The stream s bases its control actions on its state, which not only takes as input the current measured ABR, but also accumulates past observations. One can also interpret the system (18) as a low pass filter with input w and output x.

Let us introduce a rate update scheme which is approximately proportional to the control actions:

$$\dot{r}^s = -\phi r^s + u^s, \tag{19}$$

where $\phi > 0$ is sufficiently small. Since w is a function of the available bandwidth B according to (17), which, in turn, is a function of the aggregate rates from all video streams, the systems (18) and (19) are connected via a feedback loop. For simplicity, the coefficient of u^s is chosen to be unity in (19). Since increases in allocated rate will reduce the residual bandwidth, the parameter b in (18) needs to be negative. Notice that we resort here to a "bandwidth probing scheme" similar to the additive-increase multiplicative-decrease (AIMD) principle in TCP Congestion Control [5].

We now consider the *controlled output*, z^s, as a two dimensional vector:

$$z^s := [hx^s \;\; gu^s]^T, \tag{20}$$

where g and h are positive weighting parameters. In H^∞ analysis, the cost of stream s is defined as the ratio of the L^2-norm of z^s to that of w:

$$L^s(x^s, u^s, w) = \frac{\|z^s\|}{\|w^s\|}, \tag{21}$$

where $\|z^s\|^2 := \int_0^\infty |z^s|^2 \, d\tau$, and, $\|w\|^2 := \int_0^\infty |w|^2 \, d\tau$. Note that L^s captures the proportional changes in z^s due to changes in w. If $\|w\|$ is very large, the cost L^s should be low even if $\|z^s\|$ is large as well. A large $\|z^s\|$ indicates that the state $|x^s|$ and the control $|u^s|$ have high values reflecting and reacting to the situation, respectively. However, they should not grow unbounded, which is imposed by minimizing the cost L^s.

The performance factor γ is defined as the worst possible value for the cost L^s. H^∞-optimal control theory allows us to find an optimal controller given γ:

$$u^s = -\left(\frac{b}{g^2}\,\sigma_\gamma\right)\, x^s, \tag{22}$$

with $\sigma_\gamma = (-a \pm \sqrt{a^2 - \lambda h^2})/\lambda$ and $\lambda = 1/\gamma^2 - b^2/g^2$. Note that the optimal solution (22) is a linear feedback controller operating on the system state x. The gain can be calculated offline given a set of system (a, b) and preference (h, g) parameters, and target "worse-case" cost γ. Furthermore, it guarantees a lower bound of γ^*:

$$\gamma^* = \left[\sqrt{\frac{a^2}{h^2} + \frac{b^2}{g^2}}\right]^{-1}. \tag{23}$$

In practice, the H^∞-optimal rate control scheme is implemented over discrete time instants as follows: each stream s keeps track of the ABR of each access network n via the respective state equation (18), with w_n as input. The linear feedback control u^s is computed from (22) for each network separately, given a set of system (a, b) and preference (h, g) parameters. Finally, the stream updates its rate allocation to each network according to (19).

4.3 AIMD-Based Heuristics

As a basis for comparison, we introduce in this section two heuristic rate allocations schemes based on the additive-increase-multiplicative-decrease (AIMD) principle used by TCP congestion control [5]. Instead of performing proactive rate allocations by optimizing a chosen objective according to observed network and video characteristics, the AIMD-based schemes are reactive in nature, probing the network for available bandwidth and reducing rate allocation *after* congestion occurs.

Each stream s initiates its rate at a specified rate R^s_{min} corresponding to the minimum acceptable video quality, and increases its allocation by ΔR^s every Δt seconds unless network congestion is perceived, in which case the allocated rate is dropped by $(R^s_n - R^s_{min})/2$ over the congested network n.

We consider two variations of the AIMD-based schemes. They differ in their manners of allocation over the multiple access networks during the additive-increase phase:

- *Greedy AIMD*: The increase in rate allocation ΔR^s is allocated to the network interface offering the maximum instantaneous available bit rate ABR^s_n.
- *Rate Proportional AIMD*: The increase in rate allocation ΔR^s is allocated to all available networks in proportion to the average ABR of each.

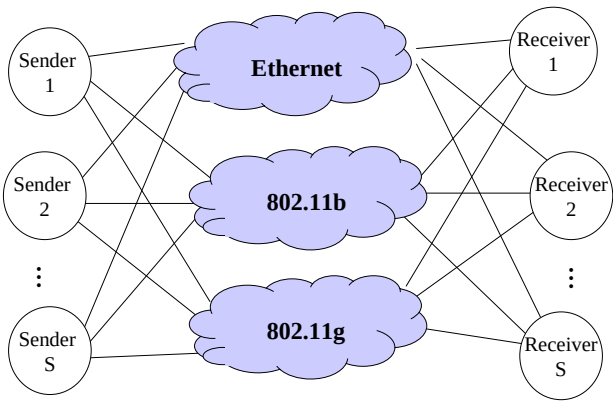


Figure 2: Topology for network simulations

		ABR(Mbps)	RTT(ms)
Ethernet	Avg.	31.5	190.1
	Std. Dev.	1.7	0.03
802.11g	Avg.	15.1	193.0
	Std. Dev.	3.6	3.2
802.11b	Avg.	4.2	195.7
	Std. Dev.	0.3	0.3

Table 1: Statistics of measured Available Bit Rate (ABR) and round-trip-time (RTT) from Deutsche Telekom Laboratories to Stanford University.

In both schemes, congestion over Network n is indicated upon detection of a lost packet, or when the observed RTT exceeds a specified threshold, based on the playout deadline of the video stream.

5. PERFORMANCE EVALUATION

5.1 Simulation Methodology

We simulate all four rate allocation policies in NS-2 [1], for an example network topology shown in Fig. 2. Each sender streams one HD video sequence via all three access networks (Ethernet, 802.11b and 802.11g) to its receiver, using the middleware functionality depicted in Fig. 1 for determining its total rate and allocation over each network.

Each network is simulated as a link with varying available bandwidth and delay, according to the traces collected from several actual access networks using the ABR and RTT measurement tool [25] [1]. Table 1 summarizes the statistics of the collected ABR ad RTT trace between Deutsche Telekom Laboratories in Berlin and Stanford University in California, over a two-hour duration. Details of the trace collection procedures and online bandwidth and delay measurements are reported in [18].

Three HD video sequences: *Bigships*, *Cyclists*, *Harbor* are streamed by three senders, respectively. The sequences have spatial resolution of 1280×720 pixels, and frame rate of 60 fps. Each stream is encoded using a fast implementation of the H.264/AVC codec [26] [27] at various quantization step sizes, with GOP length of 30 and IBBP... structure similar to

[1]Forward and backward trip delays are both simulated as half of the measured RTTs.

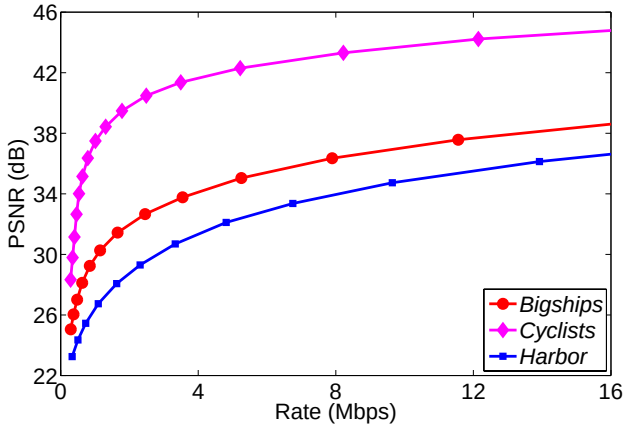


Figure 3: Rate-PSNR performance of 3 HD video sequences used in the experiments: *Bigships*, *Cyclists* and *Harbor*, all encoded using the H.264/AVC codec at 60 frames per second, GOP length of 30.

that often used in MPEG-2 bitstreams. The trade-off curves between average encoded video quality in PSNR and average bit rates over the entire sequence durations are plotted in Fig. 3. Encoded video frames are segmented into packets with maximum size of 1500 bytes, and the transmission intervals of each packet in the entire GOP are spread out evenly, so as to avoid unnecessary queuing delay due to the large sizes of intra coded frames.

In addition to the video streaming sessions, additional background traffic is included over each interface, using the exponential traffic generator in NS-2. The background traffic rate is varied between 10% and 50% of the total ABR of each network. We first compare the performance of various allocation schemes with a background traffic load of 20% and fixed playout deadline of 300 ms over each network in Section 5.2. The impact of background traffic load on the allocation results obtained from different schemes is studied in Section 5.3. The effect of different video streaming playout deadlines is investigated in Section 5.4.

5.2 Comparison of Allocation Traces

The traces of aggregate rate allocated over the Ethernet interface are plotted for all four allocation schemes, together with the available bit rate over that network. It can be observed in Fig. 4 (a) that the media-aware allocation avoids much of the fluctuations in the two AIMD-based heuristics. Fig. 4 (b) shows that it achieves higher network utilization than H^∞ optimal rate allocation, which is designed to optimize for the worst-case scenario. Similar observations also hold for the traces of aggregate allocated rate over the other two interfaces.

In Fig. 5, we compare the traces of total allocated rate for each video stream, resulting from the various allocation schemes. In greedy AIMD allocation, the total rate of each stream increases until multiplicative decrease is triggered by either packet losses or increase in the observed RTTs from one of the interfaces. Therefore traces of the allocated rates bear a saw-tooth pattern. Behavior of the rate proportional AIMD scheme is similar, except that rate drops tend to occur around the same time. The allocations from H^∞ optimal control yields less fluctuations. In both the rate

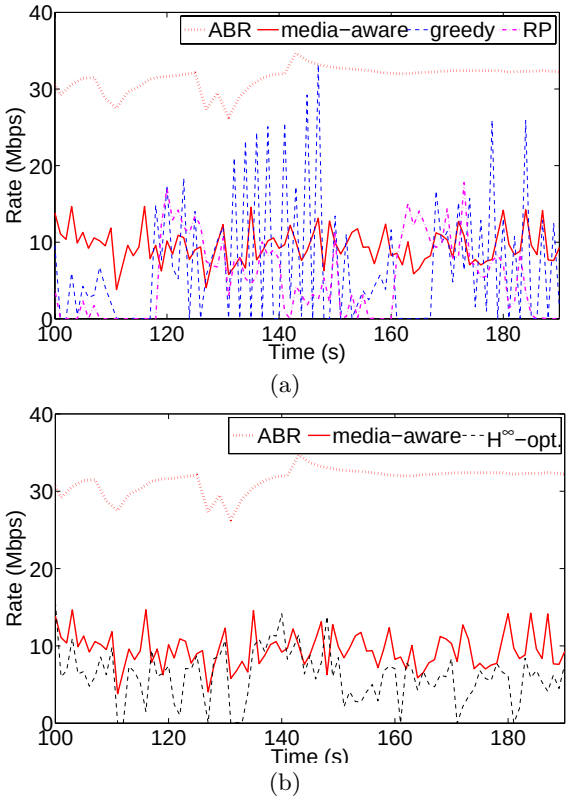


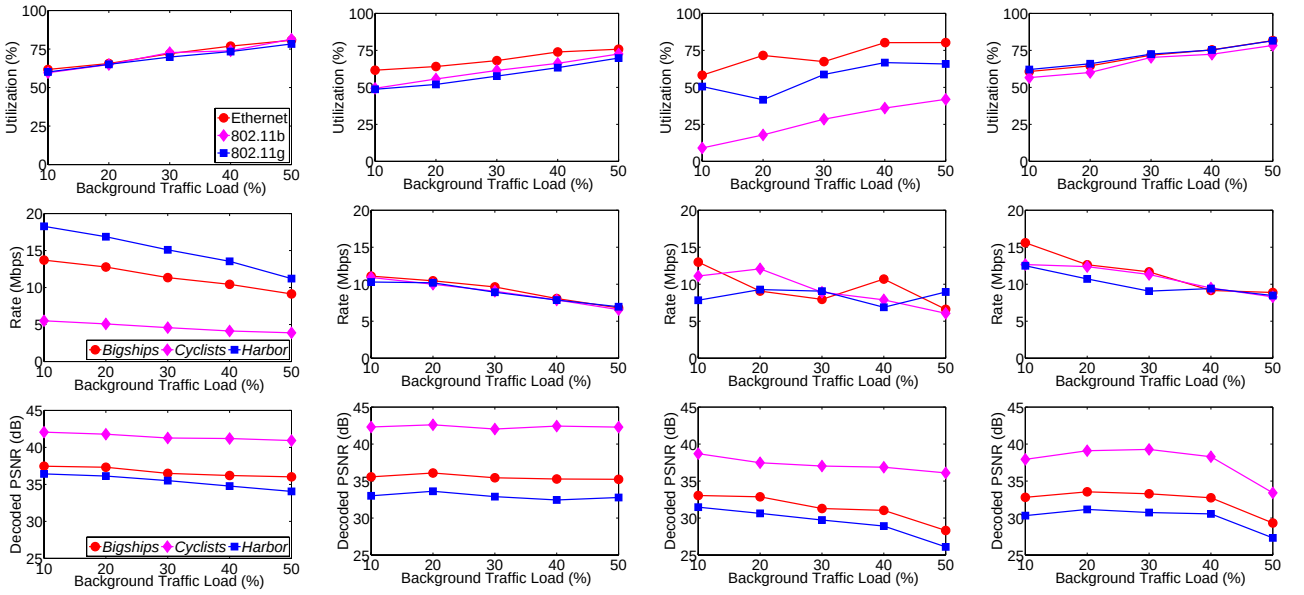
Figure 4: Trace of aggregated rate over the Ethernet interface. (a) Media-aware allocation versus AIMD-based heuristics; (b) Media-aware allocation versus H^∞ optimal control. In this experiment, background traffic load is 20% and the playout deadline is 300 ms. The network available bit rate is also plotted as a reference.

proportional AIMD allocation and the H^∞ optimal control schemes, allocated rates are almost identical to each video stream, since all flows are treated with equal importance. The media-aware convex optimization scheme, in contrast, consistently allocates higher rate for the more demanding *Harbor* stream, with reduced allocation for *Cyclists* with less complex contents.

5.3 Impact of Background Traffic Load

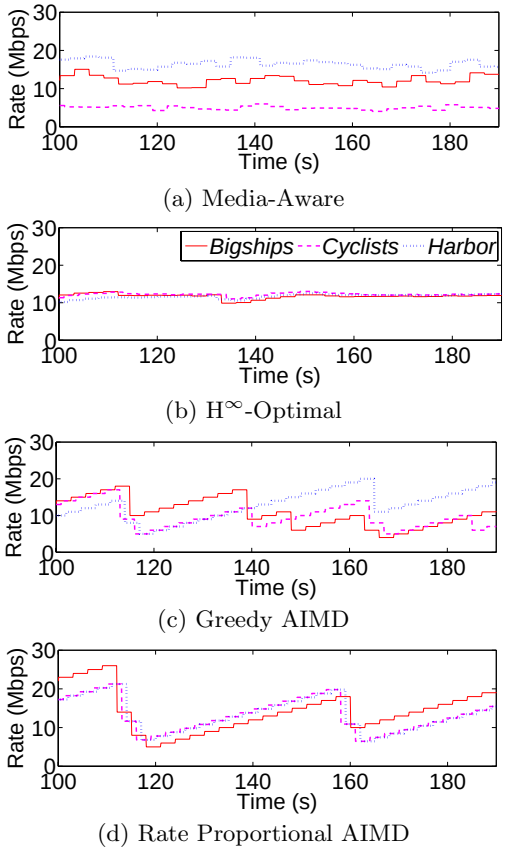
Next, we vary the percentage of background traffic over each network from 10% to 50%. Figure 6 compares the average utilization over each interface, allocated rate to each stream, and corresponding received video quality achieved by the four allocation schemes, for background traffic load of 30%. The impact of the background traffic load on the allocation results is shown in Fig. 7. It can be observed that utilization over each interface increases with the background traffic load. For the media-aware, H^∞ optimal and rate proportional AIMD schemes, utilization varies between 60% to 90%, whereas for the greedy AIMD scheme, the 802.11b interface is underutilized. [2] Note that utilization over all three interfaces are balanced by the media-aware convex optimization allocation, as predicted by (12).

[2]In fact, since 802.11b has significantly lower ABR than the other two interfaces, it is never chosen by the greedy allocation.



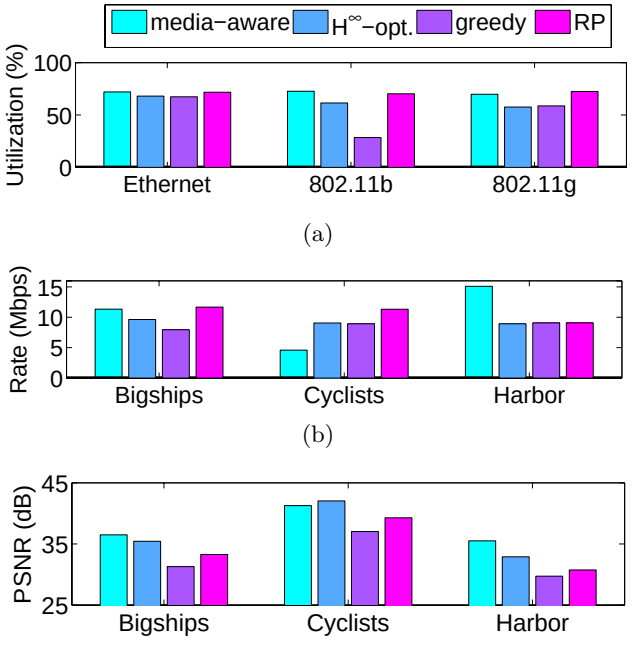
(a) Media-Aware (b) H$^\infty$-Optimal (c) Greedy AIMD (d) Rate Proportional AIMD

Figure 7: Comparison of allocation results from different schemes, as the background traffic load increases.



(a) Media-Aware (b) H$^\infty$-Optimal (c) Greedy AIMD (d) Rate Proportional AIMD

Figure 5: Trace of allocated rate to each video stream, aggregated over three interfaces. Background traffic load is 20% and the playout deadline is 300 ms.



(a) (b) (c)

Figure 6: Comparison of allocation results from different schemes, with background traffic load of 30%, and playout deadline chosen at 300 ms. (a) Aggregated network utilization over each interface; (b) Allocated video rate for each stream; (c) Corresponding received video quality in PSNR.

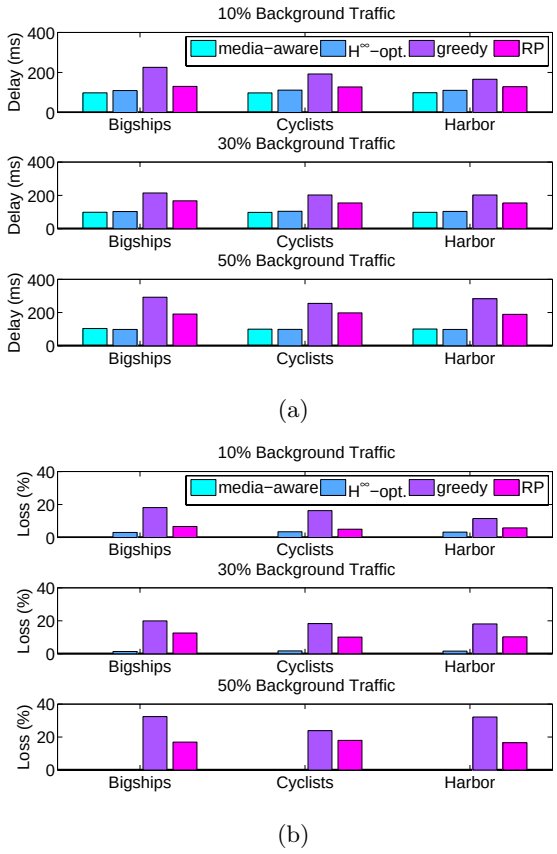


Figure 8: Average packet delivery delay for each video stream (a), and packet loss rates at the decoder (b), with playout deadline of 300 ms and background traffic load at 10%, 30% and 50%, respectively.

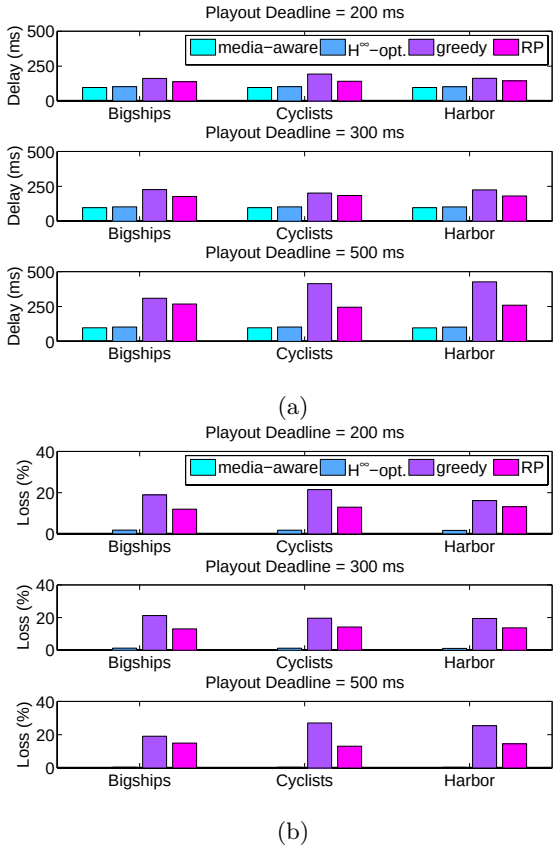


Figure 10: Average packet delivery delay (a) for each video stream, and packet loss rates (b) at the decoder, for playout deadline of 200 ms, 300 ms and 500 ms respectively. Background traffic load is chosen at 20 %.

In Fig. 6 (b), it can be noticed that the media-aware allocation leads to much lower rate for *Cyclists* and much higher rate for *Harbor*, compared to the other schemes. This improves the video quality of *Harbor*, the stream with the lowest PSNR among the three, at the cost of reducing the quality of *Cyclists*. As a consequence, the video quality is more balanced among the streams (see Fig. 6 (c)).

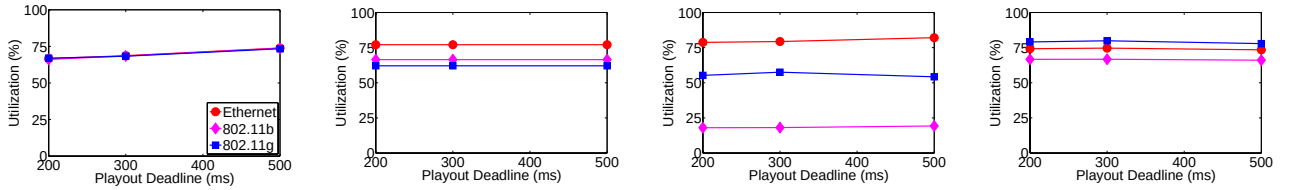
Fig. 7 (b) demonstrates that, as the background traffic load increases, allocated rate of each stream decrease accordingly. While the other three schemes treat the three flows with equal importance, the media-aware allocation consistently favors the more demanding *Harbor*, thereby reducing the quality gap between the three sequences.

Figure 8 compares the average packet delivery delay and packet loss ratios due to late arrivals. In the two AIMD-based schemes, allocated rates are reduced only *after* congestion has been detected. The media-aware convex optimization and the H^{∞} optimal allocation schemes, on the other hand, try to avoid over-congesting the network in a proactive manner in their problem formulations, therefore can achieve significantly lower packet loss ratios and delays and improved received video quality.

5.4 Varying Playout Deadline

In the next set of experiments, we vary the playout deadline for each video streams from 200 ms to 500 ms, fixing the background traffic load to 20%. As the playout deadline increases, higher packet delay can be tolerated for each video stream. The media-aware convex optimization scheme therefore increases its allocation accordingly, as shown in Fig. 9. Allocation from the other three media-unaware schemes, in comparison, tend to remain the same regardless of playout deadlines of the video streams.

Comparison of the average packet delivery delay and packet loss ratios due to late arrivals are shown in Fig. 10. Similar to the results in Fig. 8, the average packet delay achieved by the media-aware and H^{∞} optimal allocations are much lower than those achieved by the two AIMD-based heuristics, especially in the case of higher playout deadlines. The packet loss rates are almost negligible (less than 0.1 %) from the media-aware allocation, and very small (between 0.5 % and 1.9 %) from H^{∞} optimal control. In comparison, the two AIMD-based heuristics lead to packet loss rate in the range of 16 - 27 % and 12 - 15%, respectively. As a consequence, while the average received video quality of *Bigships* at playout deadline 300 ms is 34.0 dB and 32.8 dB from the greedy and rate proportional AIMD schemes, respectively, they are improved to 37.3 dB with the media-aware convex optimization, and to 36.0 dB with H^{∞} optimal control. Similar results are observed for other sequences, and other playout deadlines, as shown in Fig. 11.



(a) Media-Aware (b) H$^\infty$-Optimal (c) Greedy AIMD (d) Rate Proportional AIMD

Figure 9: Aggregated network utilization over each interface, as the playout deadline increases from 200 ms to 500 ms.

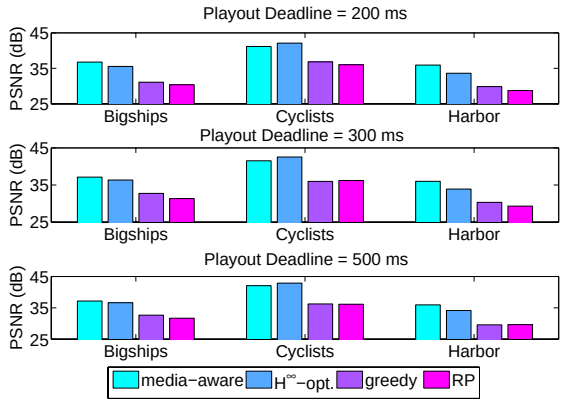


Figure 11: Received video quality in PSNR for *Bigships*, *Cyclists* and *Harbor*, for playout deadline of 200 ms, 300 ms and 500 ms respectively. The background traffic load is chosen at 20 %.

6. CONCLUSIONS

We consider the problem of rate allocation for multiple video streaming sessions sharing multiple access networks. We provide an analytical framework for optimal rate allocation based on observed network attributes (e.g., available bit rates and round trip times) and video distortion-rate (DR) characteristics. The proposed media-aware rate allocation scheme is compared against several alternatives, including a robust flow rate allocation scheme based on H$^\infty$-optimal control and two heuristic AIMD-based schemes with proportional or greedy allocation over each network.

All four schemes are evaluated in NS-2 simulations, where three different high-definition (HD) video sequences are simultaneously streamed over three heterogeneous access networks. Our results demonstrate that both the media-aware convex optimization scheme and the robust allocation with H$^\infty$ optimal control lead to smaller rate fluctuations, lower delays and significantly reduced packet losses than the two AIMD-based heuristics: the former benefit from proactive avoidance of network congestion, while the later adjusts the allocated rates reactively, for instance after detection of packet drops or excessive delays. The media-aware approach further takes advantage of explicit knowledge of the video distortion-rate (DR) characteristics, and can achieve more balanced video quality than the other schemes. While allocation from the other schemes are oblivious to the video streaming playout deadlines, the media-aware scheme adjusts its level of aggressiveness in the allocation accordingly, and achieves higher network utilization in case of a more relaxed deadline.

APPENDIX

We now provide the H$^\infty$-optimal control formulation for the general case of multiple access networks for a single stream $s \in S$ and drop the superscript s for ease of notation.

Let us define $\mathbf{x} := [x_n]$, $\mathbf{r} := [r_n]$, and $\mathbf{u} := [u_n]$ for all $n \in \mathcal{N}$. Then, the counterpart of the system (18) and (19) is given by

$$\begin{aligned} \dot{\mathbf{x}} &= A\,\mathbf{x} + B\,\mathbf{u} + D\,\mathbf{w} \\ \dot{\mathbf{r}} &= -\Phi\,\mathbf{r} + \mathbf{u}, \end{aligned} \tag{24}$$

where $\mathbf{w} := [w_n] \quad \forall n$. Here, the matrices A, B, and Φ are obtained simply by multiplying the identity matrix by a, b, and ϕ, respectively.

The counterpart of the *controlled output* in (20) is:

$$\mathbf{z} := H\mathbf{x} + G\mathbf{u}, \tag{25}$$

where we assume that G^TG is positive definite, and that no cost is placed on the product of control actions and states: $H^TG = 0$. The matrix H represents a cost on variation from zero state, i.e. full capacity usage.

The cost function is defined as:

$$L(\mathbf{x}, \mathbf{u}, \mathbf{w}) = \frac{\|\mathbf{z}\|}{\|\mathbf{w}\|}, \tag{26}$$

where $\|\mathbf{z}\|^2 := \int_0^\infty |\mathbf{z}(t)|^2 dt$ and $\|\mathbf{w}\|^2 := \int_0^\infty |\mathbf{w}(t)|^2 dt$. The corresponding differential game is parameterized by γ:

$$J_\gamma(\mathbf{u}, \mathbf{w}) = \|\mathbf{z}\|^2 - \gamma^2 \|\mathbf{w}\|^2. \tag{27}$$

Here γ is larger than the γ^* defined in (23).

The corresponding game algebraic Ricatti equation (GARE)

$$A^TZ + ZA - Z(B(G^TG)^{-1}B^T - \gamma^{-2}DD^T)Z + Q = 0 \tag{28}$$

admits a unique minimal nonnegative definite solution $\bar{Z}_\gamma$, for $\gamma > \gamma^*$, if (A, B) is stabilizable and (A, H) is detectable [24].

Similar to the solutions for the scalar system, we obtain the H$^\infty$-optimal linear feedback controller for the multiple network case:

$$\mu_\gamma(\mathbf{x}) = -(G^TG)^{-1}B^T\bar{Z}_\gamma\mathbf{x}, \tag{29}$$

for each $\gamma > \gamma^*$, which is also stabilizing.

A. REFERENCES

[1] "NS-2," `http://www.isi.edu/nsnam/ns/`.

[2] P. Vidales, J. Baliosion, J. Serrat, G. Mapp, F. Stejano, and A. Hopper, "Autonomic sytem for mobility support in 4G networks," in *IEEE Journal on Selcted Areas in Communications*, Dec. 2005, vol. 23, pp. 2288–2304.

[3] "IEEE 802.21," http://www.ieee802.org/21/.

[4] A. Cuevas, J. I. Moreno, P. Vidales, and H. Einsiedler, "The IMS platform: A solution for next generation network operators to be more than bit pipes," in *IEEE Communications Magazine, Issue on Advances of Service Platform Technologies*, Aug. 2006, vol. 44, pp. 75–81.

[5] V. Jacobson, "Congestion avoidence and control," in *Proc. SIGCOMM'88*, Aug. 1988, vol. 18, pp. 314–329.

[6] M. Allman, V. Paxson, and W. R. Stevens, *TCP Congestion Control, RFC 2581*, Apr. 1999.

[7] S. Floyd and K. Fall, "Promoting the use of end-to-end congestion control in the Internet," *IEEE/ACM Trans. on Networking*, vol. 7, no. 4, pp. 458–472, Aug. 1999.

[8] M. Handley, S. Floyd, J. Pahdye, and J. Widmer, *TCP Friendly Rate Control (TFRC): Protocol Specification, RFC 3448*, Jan. 2003.

[9] Z. Wang, S. Banerjee, and S. Jamin, "Media-friendliness of a slowly-responsive congestion control protocol," in *Proc. 14th International Workshop on Network and Operating Systems Support for Digital Audio and Video*, Cork, Ireland, 2004, pp. 82–87.

[10] F. Kelly, A. Maulloo, and D. Tan, "Rate control for communication networks: Shadow prices, proportional fairness and stability," *Journal of Operations Research Society*, vol. 49, no. 3, pp. 237–252, 1998.

[11] H. Yaiche, R. Mazumdar, and C. Rosenburg, "A game theoretic framework for bandwidth allocation and pricing in broadband networks," *IEEE/ACM Trans. on Networking*, vol. 8, no. 5, pp. 667–678, Oct. 2000.

[12] T. Alpcan and T. Başar, "A utility-based congestion control scheme for Internet-style networks with delay," *IEEE Trans. on Networking*, vol. 13, no. 6, pp. 1261–1274, December 2005.

[13] T. Alpcan and T. Başar, "Global stability analysis of an end-to-end congestion control scheme for general topology networks with delay," in *Proc. 42nd IEEE Conference on Decision and Control (CDC'03)*, Maui, HI, U.S.A., Dec. 2003, pp. 1092–1097.

[14] S. Shakkottai, E. Altman, and A. Kumar, "The case for non-cooperative multihoming of users to access points in IEEE 802.11 WLANs," in *Proc. IEEE INFOCOM'06*, Barcelona, Spain, Apr. 2006, pp. 1–12.

[15] A. Szwabe, A. Schorr, F. J. Hauck, and A. J. Kassler, "Dynamic multimedia stream adaptation and rate control for heterogeneous networks," in *Proc. 15th International Packet Video Workshop, (PV'06)*, Hangzhou, China, May 2006, vol. 7, pp. 63–69.

[16] D. Jurca and P. Frossard, "Media-specific rate allocation in heterogeneous wireless networks," in *Proc. 15th International Packet Video Workshop, (PV'06)*, Hangzhou, China, May 2006, vol. 7, pp. 713–726.

[17] X. Zhu, J. P. Singh, and B. Girod, "Joint routing and rate allocation for multiple video streams in ad hoc wireless networks," in *Proc. 15th International Packet Video Workshop, (PV'06)*, Hangzhou, China, May 2006, vol. 7, pp. 727–736.

[18] J. P. Singh, T. Alpcan, P. Agrawal, and V. Sharma, "An optimal flow assignment framework for heterogeneous network access," in *Proc. IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks*, Helsinki, Finland, Apr. 2007.

[19] T. Alpcan, J. P. Singh, and T. Basar, "A robust flow control framework for heterogenous network access," in *Proc. 5th Intl. Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks*, Limassol, Cyprus, June 2007.

[20] L. Kleinrock, *Queuing Systems, Volume II: Computer Applications*, Wiley Interscience, New York, USA, 1976.

[21] K. Stuhlmüller, N. Färber, M. Link, and B. Girod, "Analysis of video transmission over lossy channels," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 6, pp. 1012–32, June 2000.

[22] X. Zhu, E. Setton, and B. Girod, "Congestion-distortion optimized video transmission over ad hoc networks," *EURASIP Journal of Signal Processing: Image Communications*, vol. 20, no. 8, pp. 773–783, Sept. 2005.

[23] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, United Kindom, 2004.

[24] T. Basar and P. Bernhard, *H^∞-Optimal Control and Related Minimax Design Problmes: A Dynamic Game Approach*, Birkhaüser, Boston, MA, 1995.

[25] Jiri Navratil and R. Les. Cottrell, "Abing," http://www-iepm.slac.stanford.edu/tools/abing/.

[26] ITU-T and ISO/IEC JTC 1, *Advanced Video Coding for Generic Audiovisual services, ITU-T Recommendation H.264 - ISO/IEC 14496-10(AVC)*, 2003.

[27] "x.264," http://developers.videolan.org/x264.html.