Scheduling Support for Mixed VoIP and Web Traffic over HSDPA

Mats Folke*†, Sara Landström†, Ulf Bodin†, Stefan Wänstedt*
mats.folke@ericsson.com, sara.landstrom@ltu.se, ulf.bodin@ltu.se, stefan.wanstedt@ericsson.com
†Department of Computer Science and Electrical Engineering, Luleå University of Technology, Sweden
*Ericsson Research, Luleå, Sweden

***Abstract*—HSDPA (High-Speed Downlink Packet Access), introduced in WCDMA release 5, provides a high-bandwidth shared channel with short transmission time interval (TTI). The short TTI together with appropriate scheduling enable HSDPA to support efficient multiplexing of traffic.**

We explain the performance of four scheduling algorithms when transmitting a traffic mix consisting of both conversational (VoIP) traffic and background (web) traffic over the high-speed downlink shared channel (HS-DSCH) of HSDPA. We consider both cell throughput and user satisfaction. The proportional fair (PF), the maximum rate (MR) scheduler and two extended versions of MR, are tested for different VoIP scheduling delay budgets and varying load.

To understand the behaviour of the schedulers, we use the ns-2 simulator extended with a model of HS-DSCH to simulate a mixed traffic scenario. Our results show that a scheduler that gradually increases the VoIP priority and considers the user's current possible rate, performs well. A more drastic increase in VoIP priority is however needed when the delay budget is short. Furthermore, attempting to uphold quality for both VoIP and web traffic makes the system sensitive to overload situations.

I. Introduction

In evolved 3G systems, such as High Speed Downlink Packet Access (HSDPA) [1] and Evolution Data Only (EvDO) [2], time-shared channels are introduced to improve capacity for best-effort traffic using TCP [3]. The latency is also reduced in HSDPA by moving some of the radio resource management responsibilities to the base station from the radio network controller. Among these responsibilities are link adaptation and channel aware scheduling of the short TTIs. Higher order modulation and Hybrid ARQ are also important features.

TCP tends to have a bursty sending pattern, making time multiplexing efficient. Recent studies [4], [5] show that it is feasible to efficiently transport also Voice over IP (VoIP) and Push-To-Talk (PTT) over time-shared channels with sufficient quality.

Ultimately, all services should be able to coexist over a few shared channels. This simplifies the infrastructure, which creates opportunities for savings in investments as well as operational costs. Therefore, techniques for service differentiation are an important area of research. In particular, the scheduling algorithm that assigns time slots, power and codes to the users is crucial in a time-shared and service differentiating system.

So far most of the scheduling algorithms proposed for time-shared wireless systems, such as HSDPA and EvDO, focus on supporting either best-effort, streaming or conversational traffic. In this paper we consider a scenario where simultaneous conversational and best-effort traffic compete for time slots. Previous studies on best-effort traffic and streaming separately, show that these services can be supported well by HSDPA [6].

In a mixed best-effort and streaming traffic scenario it has been shown that a reasonably fair scheduler can provide sufficient quality without service differentiation [7] as long as the load is not high. At higher loads, service differentiation is however needed to protect the quality of service of the streaming users. Service differentiation was provided by a scheduler giving strict priority to the streaming flows and a Proportional Fair (PF) scheduler with the barrier function from [8]. The barrier function induces a penalty when the user receives less than the desired minimum bit rate. It was found that strict priority comes at the expense of the web traffic, whereas the minimum bit rate aware scheduler achieves a better balance between the two services.

Conversational traffic has even stricter delay requirements than streaming traffic and generally lower bit rate requirements. These characteristics make service differentation more important for conversational traffic than streaming when sharing a channel with best-effort traffic.

In this paper we study four scheduling algorithms regarding their ability to maintain a high cell throughput while keeping a mix of VoIP and web users satisfied. The aim is to identify a set of desirable features for a scheduling algorithm in this type of scenario, that will give us the understanding to develop better algorithms.

The investigated schedulers are the well-known Proportional Fair (PF) and Max-rate (MR) schedulers. In addition, we study two variations of MR scheduling. In [9], a number of different schedulers are studied for a mix of best-effort and VoIP traffic. We provide a more in-depth analysis of the reasons for the observed system throughput and user satisfaction levels. Furthermore, our TCP model is dynamic and adjusts to the current network conditions.

In the next section, the models and metrics used in our simulation study are described. The results are reported in Section III and discussed in Section IV.

1550-2252/$25.00 ©2007 IEEE

II. Methodology

In this section, we describe the simulation model, the scheduling algorithms and our evaluation criteria.

A. Metrics

A VoIP user is satisfied if its Frame Loss Rate (FLR) is below 1%, which is elaborated upon in [10]. This value corresponds to a good perceived speech quality for the AMR codec. For web traffic, the target bit rate is set to 64 kbps.

The ITU E-model [11] states that when the one-way mouth-to-ear delay exceeds 250 ms the voice quality rating rapidly deteriorates. When delay induced by the voice encoder/decoder and other nodes in the system is subtracted, about 80 to 150 ms remains for Node-B processing and UE reception [4]. The exact value depends on whether both users are mobile or not. We define the term *delay budget* to be the time available for scheduling.

B. HS-DSCH model

Our simulation model of HS-DSCH extends the *Network Simulator* version 2.28, ns-2 [12] and is available at [13]. The model is further described in [14].

In our simulations, the cell plan consists of seven sites with three sectors each. Wrap-around for mobility and interference is used to improve the reliability of the results.

The radio model includes lognormal shadow fading with a standard deviation of 8.9 dB and exponential path loss with a propagation constant of 3.5. The multi-path model is Typical Urban and is dependent on the speed of the user. In the simulations the fading is modelled at a user speed of 3 km/h, but the users are stationary. Thus no hand-overs will occur.

We assume that the Channel Quality Information (CQI) is reported by the receivers every TTI. It determines the modulation and coding scheme that achieves the highest efficiency given a certain residual Block Error Rate (BLER). We use the model presented in [5] to perform the mapping, but we choose from the finite set of Transport Block (TB) sizes given in Annex A of [15] instead of assuming that any rate is possible. In total 23 different TB sizes are represented in the simulation module. Code multiplexing for up to four users in the same TTI in each cell is supported. The optimisation criteria is system throughput and user ranking. The highest prioritised user is assigned the power and codes necessary to support its highest possible bitrate, thereafter the user ranked second will be scheduled if there are sufficient resources left. We adjust the assignments to higher prioritised users if more users are scheduled.

When the Signal to Interference Ratio (SIR) is less than the requirement for the smallest TB size, the BLER is 50%. For better SIR conditions the BLER is 20%. This relatively high percentage compensates for the assumption of perfect channel estimation and is similar to the error rate observed in [4].

If a TB is corrupted the user's queue will be blocked for 12 ms, before data for that user are eligible for transmission. 12 ms corresponds to a configuration with six Hybrid ARQ (HARQ) processes in a real system. The corrupted data are given higher priority than the other data in the user's queue, but not higher priority than data from other users. The user has to be scheduled according to the scheduling algorithm to perform retransmissions.

In this simplified HARQ model, a block is retransmitted until the transmission succeeds or the delay budget is exceeded after which the block is dropped. For web traffic, the delay budget is set high enough to avoid all packet losses due to delay.

C. Application model

We consider a mix of conversational (VoIP) traffic and interactive (web) traffic.

1) VoIP traffic: VoIP traffic is assumed to have exponentially distributed on and off times, both with an average duration of 7 s. A VoIP frame is sent every 20 ms during the on periods, which yields a bit rate of 12.2 kbps, comparable to one of the AMR codec bitrates [16]. The compressed IP/UDP/RTP header increases the bitrate to 13.6 kbps [17].

2) Web traffic: Web traffic yields mostly short flows. The file sizes are drawn from a Pareto distribution with a mean of 30458 bytes and the shape parameter set to 1.7584 [18]. When a user has finished a transfer there is an exponentially distributed waiting time with an average of 0.5 s before the next transfer begins.

For TCP, we use the tcp-sack1 agent in ns-2. It supports Limited Transmit [19] and a variant of SACK loss recovery, as specified in [20]. As in real TCP implementations, acknowledgements are used to clock out new segments. This creates a dependency between received throughput and sending rate. The segment size is set to 1460 bytes.

3) Load model: The load is varied by changing the number of users. 50% of the users generate VoIP streams and the other half web transfers. The mobiles are spread according to a uniform distribution and the same number of users are placed in each cell.

It is not possible to compare the capacity in terms of users acheived in this study against a study with only VoIP traffic, since the web traffic is much more demanding in terms of bit rate.

D. Schedulers

The Proportional Fair (PF) scheduler will pick the user for which

$$i* = \arg\max_i \left\{ \frac{r_i}{\mu_i} \right\} \tag{1}$$

where μ_i is the average throughput of user i and r_i the instantaneous rate considering the data buffered for the user and its radio conditions. The average throughput is continously updated even if there is no data to be scheduled. TCP has a bursty sending pattern and therefore updating the average throughput continously will better reflect the performance. For VoIP users, data sent after a period of silence will be prioritised. The start of each talk spurt is important for the intelligibility of speech.

MR also considers the data buffered as well as the radio conditions. When combinded with the minimum bit rate requirement, we get $\mathrm{MR_{min}}$. $\mathrm{MR_{min}}$ selects the user that satisfies the following condition:

$$i* = \arg\max_i \{ r_i (1 + \beta e^{-\beta(\mu_i - \mu_{min})}) \}. \qquad (2)$$

The parameter β determines the rate at which the penalty for violating the constraint increases and it should be based on the service class of the user. In the simulations, the $\mathrm{MR_{min}}$ scheduler has had a μ_{min} of 13.6 kbit/s for VoIP traffic and 64 kbit/s for web traffic.

Both PF and $\mathrm{MR_{min}}$ include the average throughput in their metrics. For a web transfer, it is computed from the start of the transfer, whereas we use an Exponentially Weighted Moving Average (EWMA) filter for VoIP that puts 90% of the weight on the throughput achieved during the delay budget.

The strict delay scheduler, $\mathrm{MR_{delay}}$, starts to prioritise VoIP when 40 ms remain of the delay budget. Users are otherwise ranked according to their instantaneous rate. When the delay budget is exceeded the priority of the VoIP users' packets is set based on their delay.

Focusing on VoIP, the chosen schedulers represent different types of prioritisation functions. $\mathrm{MR_{delay}}$ drastically increases the priority of the VoIP flows when the delay reaches a certain value. We call this *strict prioritisation*. The longer the user has waited since it last transmitted, the higher it will be prioritised by PF since the average throughput drops. This increase in user prioritisation is gradual and we will refer to it as *soft prioritisation*. Although a VoIP user will assemble more and more data the increase in priority over time with MR is relatively minor and foremost has the potential to settle the ranking between VoIP users. It will not significantly improve the priority of VoIP relative web traffic.

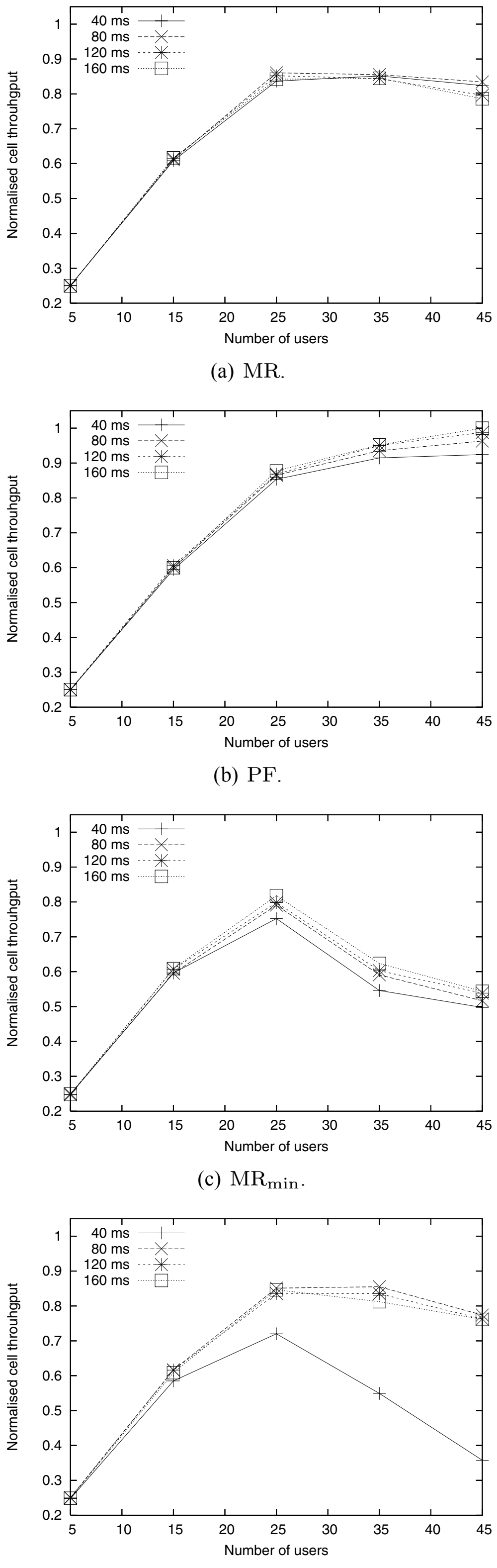
III. Results

Each simulation has run for 50 simulated seconds. We study the average throughput per second to determine when steady state is reached. The system was not in steady state during the first 10 seconds and therefore this part of the simulations was disregarded. The remaining 40 seconds were divided into four blocks of 10 seconds each. Each block has been viewed as independent from the others. Every setup was repeated twice with different seeds. This means that the values in Figures 1 and 4 have been averaged over eight independent samples. In the simulations, the delay budget is set to 40, 80, 120, and 160 ms, respectively.

A. Cell throughput

In Fig. 1 the average cell throughput is plotted for various loads and delay budgets. The load is determined by the number of users in each cell. We normalise the average cell throughput with respect to the maximum value of the PF scheduler, because our focus is on the relative performance of the schedulers.

For 5 up to 25 users in each cell, the average cell throughput is approximately the same for all schedulers. Beyond this



(a) MR.

(b) PF.

(c) $\mathrm{MR_{min}}$.

(d) $\mathrm{MR_{delay}}$.

Fig. 1. The average cell throughput for the evaluated schedulers and delay budgets. All values have been normalised to the maximum value produced by the PF scheduler.

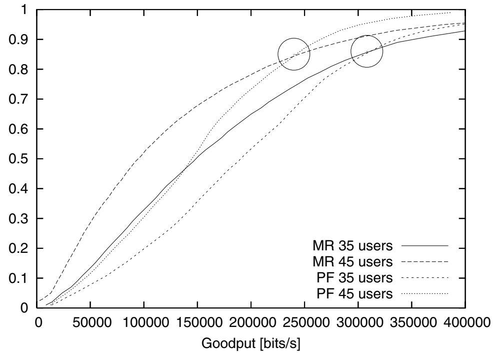


Fig. 2. The CDF of per-flow goodput for MR and PF with a delay budget of 120 ms.

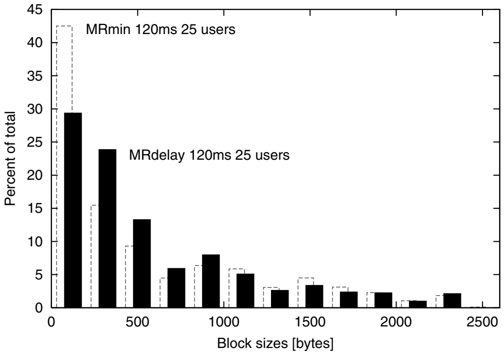


Fig. 3. The transport block sizes of $\mathrm{MR}_{\mathrm{delay}}$ and $\mathrm{MR}_{\mathrm{min}}$ for a representative scenario.

load, PF still increases the average cell throughput. MR and $\mathrm{MR}_{\mathrm{delay}}$ do not.

TCP is the main contributor to the cell throughput, whereas VoIP is a low bit rate service. We therefore study the CDF of the throughput received per flow for TCP closer in Fig. 2. About 10% of the flows get a higher throughput with MR than with PF for the same number of users. PF has fewer flows with low throughput. The intersection of the curves is emphasised through a circle, beyond this point MR provides higher bit rates to the majority of the users. The offered load depends on when flows are completed and therefore does not increase proportionally to the number of users in the cell. An investigation of the offered load verifies that it is lower with MR than with PF, because the majority of the flows experience a lower throughput. The same observation holds for $\mathrm{MR}_{\mathrm{delay}}$.

With $\mathrm{MR}_{\mathrm{min}}$ the average cell throughput drops when exceeding 25 users. By comparing the transport block sizes for $\mathrm{MR}_{\mathrm{delay}}$ and $\mathrm{MR}_{\mathrm{min}}$ in Fig. 3, we draw the conclusion that the barrier function for $\mathrm{MR}_{\mathrm{min}}$ is too aggressive, preventing good radio conditions from being exploited by the scheduler. Except when the delay budget is 40 ms, both VoIP and TCP generally achieve larger block sizes with $\mathrm{MR}_{\mathrm{delay}}$. For the smallest delay budget, $\mathrm{MR}_{\mathrm{delay}}$ sends small VoIP blocks because of the strict priority and there is little capacity left to serve the TCP users. This results in the decay in Fig. 1(d).

B. User perceived quality

Pure MR is unable to serve the VoIP users, as shown in Fig. 4(a). With larger delay budgets, the VoIP users are able to get a higher rate and can thereby compete better with the web users, but not well enough to reach the system quality constraint. When the load is increased it is foremost the VoIP users that suffer. This is because MR is biased towards web traffic.

$\mathrm{MR}_{\mathrm{delay}}$ on the other hand, prioritises VoIP traffic higher than web traffic. As the load increases, the VoIP satisfaction level therefore remains constant at the expense of web traffic quality as long as the resources are sufficient for VoIP alone.

In its attempt to uphold the quality for all users, $\mathrm{MR}_{\mathrm{min}}$ spends a lot of resources on users with bad radio conditions. A CDF of the TCP per flow throughput reveals that all users get a very similar throughput. The satisfaction level therefore drops drastically as resources become scarce.

PF has in previous studies been shown to find a good trade-off between maximising cell throughput and achieving fairness in the sense that the users get similar throughputs. VoIP has a lower average throughput than web transfers and should therefore enjoy a higher priority. The results in Fig. 4(b) support this theory.

However, for the smallest delay budget of 40 ms, only $\mathrm{MR}_{\mathrm{delay}}$ manages to present some quality to the VoIP users. $\mathrm{MR}_{\mathrm{min}}$ and PF do not give strict enough priority. This is an advantage for PF at higher loads, where the softer prioritisation creates opportunities for using larger block sizes for the web traffic than with $\mathrm{MR}_{\mathrm{delay}}$.

C. Web users with throughputs lower than 15 kbps

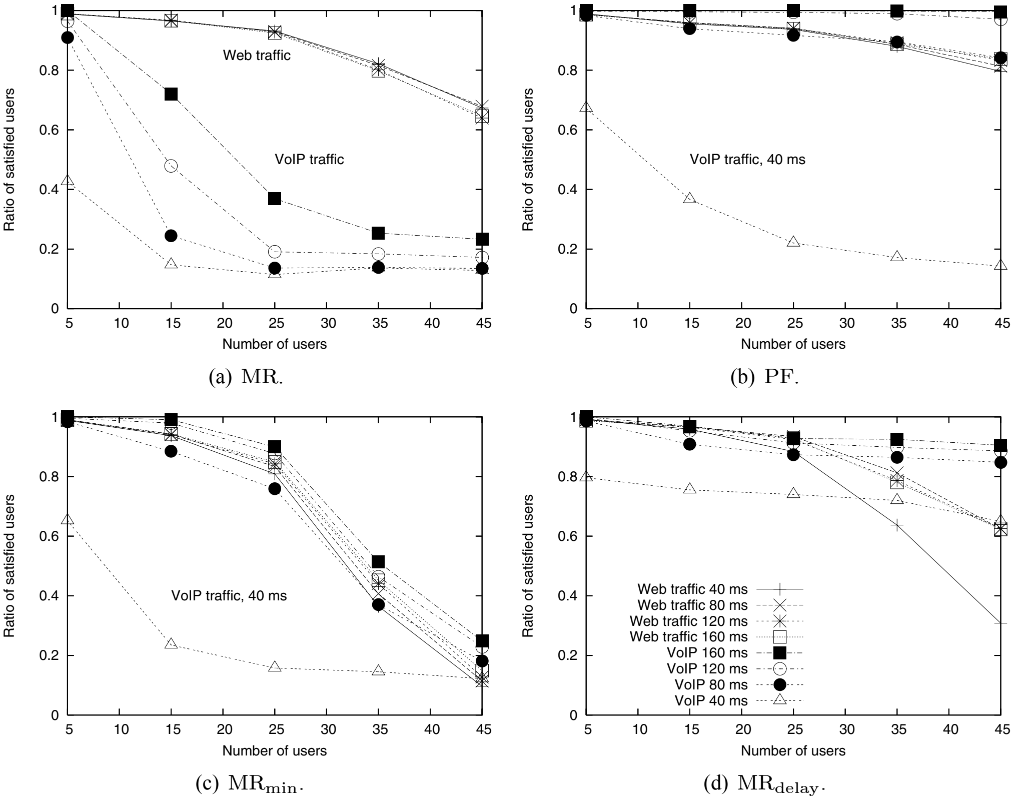
If a web user gets a throughput of less than 15 kbps, we consider that user to be starved. In some cases, a scheduler may gain from not serving all users in terms of cell throughput, because users with poor conditions demand lots of resources.

$\mathrm{MR}_{\mathrm{min}}$ hardly has any starved flows, i.e., less than 0.1%. As stated earlier, this leads to a severe drop in satisfaction when the resources are exhausted. Thereafter PF follows with approximately 1% starved flows at the highest load. For MR 2% of the flows receive less than 15 kbps. $\mathrm{MR}_{\mathrm{delay}}$ has the largest share of starved flows at high loads. With 35 users per cell almost 2.5% of the flows have a throughput of less than 15 kbps.

IV. Discussion and conclusions

The focus of our study was to identify a set of properties a scheduler for mixed interactive and conversational traffic should have. Therefore we have compared scheduling values, transport block sizes, actual offered load and throughput of individual flows for the different schedulers. The understanding that we have gained has been used to explain the high-level results reported in this paper such as average cell throughput and user satisfaction.

For small delay budgets (40 ms), strict priority as represented by $\mathrm{MR}_{\mathrm{delay}}$ must be given to VoIP traffic. This has a minor negative impact on web traffic satisfaction at high



(a) MR.

(b) PF.

(c) $\mathrm{MR}_{\mathrm{min}}$.

(d) $\mathrm{MR}_{\mathrm{delay}}$.

Fig. 4. The ratio of satisfied web traffic and VoIP traffic users for the evaluated schedulers and delay budgets. The legend is the same for all figures.

loads. For larger delay budgets, softer prioritisation as with PF is sufficient and can be favourable, since the scheduler has a larger freedom to consider the radio conditions. Trying to maintain the quality targets for all users is futile, which is shown by $\mathrm{MR}_{\mathrm{min}}$. This scheduler would probably perform better if the barrier function was only used for VoIP. MR is biased towards web traffic and as expected not suitable for VoIP.

A scheduler should thus be designed such that it can sacrifice some users when the resources are not sufficient. Soft prioritisation is desirable, since it enables better exploitation of the radio conditions.

References

[1] TSG-RAN, "UTRA High Speed Downlink Packet Access (HSDPA); Overall description; Stage 2," 3GPP, Tech. Spec. TS 25.308 V6.3.0, Dec. 2004.

[2] TSG-C, "cdma2000 High Rate Packet Data Air Interface Specification," 3GPP2, Tech. Spec. C.S0024-A V2.0, Jul. 2005.

[3] M. Frodigh, S. Parkvall, C. Roobol, P. Johansson, and P. Larsson, "Future-generation wireless networks," *IEEE Personal Communications*, vol. 8, no. 5, pp. 10–17, Oct. 2001.

[4] B. Wang, K. I. Pedersen, T. E. Kolding, and P. E. Mogensen, "Performance of VoIP over HSDPA," in *IEEE VTC-Spring*, vol. 4, May 2005, pp. 2335–2339.

[5] P. A. Hosein, "Capacity of packetized voice services over time-shared wireless packet data channels," in *IEEE INFOCOM 2005*, vol. 3, March 2005, pp. 2032–2043.

[6] P. A. Gutierrez, "Packet Scheduling and Quality of Service in HSDPA," Ph.D. dissertation, Aalborg University, Oct. 2003.

[7] M. Lundevall, B. Olin, J. Olsson, N. Wiberg, S. Wänstedt, J. Eriksson, and F. Eng, "Streaming Applications over HSDPA in Mixed Service Scenarios," in *IEEE VTC-Fall*, vol. 2, Sep. 2004, pp. 841–845.

[8] P. A. Hosein, "QoS Control for WCDMA High Speed Packet Data," in *International Workshop on Mobile and Wireless Communications Network*, Sep. 2002, pp. 169–173.

[9] A. R. Braga, E. B. Rodrigues, and F. R. Cavalcanti, "Packet Scheduling for Voice over IP over HSDPA in Mixed Traffic Scenarios with Different End-to-End Delay Budgets," in *ITS*, Sep. 2006.

[10] TSG-SA, "Performance and characterization of the Adaptive Multi-Rate (AMR) speech codec," 3GPP, Tech. Spec. TS 26.975 V6.0.0, Dec. 2004.

[11] ITU-T, "One-way transmission time," ITU, Tech. Rep. G.114, May 2003.

[12] S. McCanne and S. Floyd, "The Network Simulator – ns-2," http://www.isi.edu/nsnam/ns.

[13] U. Bodin, M. Folke, and S. Landström, "HS-DSCH extension to ns-2," http://www.csee.ltu.se/~folke/luther/.

[14] M. Folke and S. Landström, "An NS Module for Simulation of HSDPA," Luleå University of Technology, Tech. Rep. 2006:03, Feb. 2006.

[15] TSG-RAN, "Medium Access Control (MAC) protocol specification (Release 6)," 3GPP, Tech. Spec. TS 25.321 V6.6.0, Sep. 2005.

[16] TSG-SA, "AMR Speech Codec; General Description," 3GPP, Tech. Spec. TS 26.071 V6.0.0, Dec. 2004.

[17] C. Bormann *et al.*, "RObust Header Compression (ROHC): Framework and four profiles: RTP, UDP, ESP, and uncompressed," IETF, RFC Standards Track 3095, Jul. 2001.

[18] A. Reyes-Lecuona, E. González-Parada, E. Casilari, J. C. Casasola, and A. Díaz-Estrella, "A page-oriented WWW traffic model for wireless simulations," in *16th ITC*, Jun. 1999, pp. 1271–1280.

[19] M. Allman, H. Balakrishnan, and S. Floyd, "Enhancing TCP's Loss Recovery Using Limited Transmit," IETF, RFC Standards track 3042, Jan. 2001.

[20] K. Fall and S. Floyd, "Simulation-based Comparisons of Tahoe, Reno and SACK TCP," *Computer Communications Review*, vol. 26, no. 1, pp. 5–21, Jul. 1996.