Speculative Resource Allocation for Packet-Switched Wireless Networks

Magnus Lindström*, Leonardo Badia†, Jens Zander*, Michele Zorzi† ‡
{lindstrm,jensz}@radio.kth.se, {lbadia,zorzi}@ing.unife.it
* Wireless@KTH, Royal Institute of Technology, 16440 KISTA, Sweden
† Department of Engineering, University of Ferrara, 44100 Ferrara, Italy
‡ Department of Information Engineering, University of Padova, 35131 Padova, Italy

Abstract— **In this paper we study benefits of speculative scheduling in wireless networks with elastic services. Common goals of Radio Resource Management (RRM) have traditionally been maximisation of throughput and provisioning of QoS guarantees. Providers of wireless access, however, need to acquire adequate revenue to sustain their business. While throughput maximisation and QoS guarantees increase the quantity or quality, respectively, of the chargeable goods, it has also been shown that by accounting for users' perceptions of what is acceptable from QoS and pricing perspective, provider revenue may increase. We propose and evaluate a packet scheduler for realtime streaming services with the aim to maximise user satisfaction and, thus, improve operator revenue. We also extend an existing service *acceptance* probability model with a service *continuation* probability model. The proposed scheduling scheme is compared to a proportional fair scheduler. Preliminary results show that revenue can be increased significantly, while keeping user satisfaction approximately constant. Alternatively, the user satisfaction can be correspondingly improved for fixed revenue.**

I. BACKGROUND

In this paper we study benefits of speculative scheduling in wireless networks with elastic services [1], [2].

Common goals of Radio Resource Management (RRM) have traditionally been maximisation of throughput and provisioning of QoS guarantees. Providers of wireless access, however, need to acquire adequate revenue to sustain their business models and while throughput maximisation maximises the chargeable goods and QoS guarantees increase the attractiveness of the goods, the users' appreciation of a service also depend on the price of the service. That is, overpricing is as unattractive as substandard quality. It has been shown that by accounting for users' perceptions of reasonable pricing of services, provider revenue may be increased [3], [4]. Also, the introduction of appropriate pricing schemes makes it possible for the service provider to improve the efficiency and robustness of the radio resource management [5], [6]. In fact, when the resource to allocate is scarce and users compete for it, the pricing strategy might be aimed at increasing cooperation among users and preventing allocation to users who have low satisfaction from the service.

Awareness of and information about users' service appreciation can be exploited further. It can be used to improve the match between the requested service quality and the supplied service quality.

In a packet switched system, it is the task of a scheduler to distribute the radio resource among users so as to make the "best use" of the resource under the constraint that QoS requirements and/or some fairness criteria are met. The "best use" of the resource is generally taken to be equivalent to maximising throughput, revenue or some fairness criterion. To this end, numerous scheduling algorithms have been presented in the literature, a few of the most important being: Round Robin (RR), Equal Throughput (ET), Fractional Fair (FF), Proportionally Fair (PF) and Relatively Best (RB) scheduling.

The RR scheduler cycles through the users, giving them equal time on the channel. ET scheduling, as the name implies, seeks to provide users with equal throughput and can be viewed as a weighted RR scheduler with (per user) weights that are inversely proportional to the achievable rates. An FF scheduler tries to improve total throughput by giving priority to users with higher achievable rates and can also be described as a weighted RR scheme, with weights that are proportional to the achievable rates. More intricate schemes try to combine the objectives of these simple schedulers. The PF scheduler seeks to improve system throughput by exploiting multiuser diversity, while at the same time providing some degree of (long term) throughput fairness [7]. Similarly, RB scheduling exploits multiuser diversity, but maintains time fraction fairness [8]. There are also schemes that consider delay fairness, for instance: the Largest Weighted Delay First (LWDF) scheduler and its derivatives.

The underlying assumption of all of these algorithms is that higher system throughput translates to greater revenue and/or that fairness is important (or even matters). This is, however, arguable. First because services might not be priced directly proportional to their rate requirements. Second, and maybe more important, users are generally not aware of the situations of other users and so can't compare. Thus, they have no notion of fairness. A user is likely to judge the system based on the perception of receiving decent service, rather than on fairness from a time or throughput or even price point of view.

In this paper, we focus on users' service appreciation and take fairness to mean that users perceive the combination of QoS and price they are offered to be decent. Then, from this fairness point of view and assuming that disappointed users may shorten their sessions, the "best use" of the resource could be to maximise the willingness of the users to stay connected.

0-7803-8887-9/05/$20.00 (c)2005 IEEE

The provider may also allow a speculative allocation of resource to some users at the expense of lowering the service level of other users. This can be beneficial if the expected marginal revenue associated with the variation is positive. However, lowering of the service level of a user is risky. If the user no longer perceives the service as usable and affordable he may leave the system, resulting in a loss of revenue. Thus, to enable the use of speculative over-assignment, service appreciation of and resources allocated to already admitted users must be carefully monitored and managed to maximise users willingness to stay in the system.

To highlight the importance of considering users' overall perception, i.e. price and quality, of the service, we propose and evaluate a packet scheduler for realtime streaming services with the aim to maximise user satisfaction for the set of admitted users [9], [10]. In particular, we focus on the forward direction of a high speed packet access system and realtime streaming services [11]–[13]. It is further assumed that the provider can estimate the impact of his decisions, quantified with a degree of satisfaction perceived by the users.

The remainder of this paper is organised as follows: in Section II a model for describing users' willingness to accept a service offer is presented and extended to suit the needs of a scheduler evaluation. In Section III the user satisfaction or service perception aware scheduler is introduced followed by a proportionally fair reference scheduler for comparison. Finally, a numerical evaluation is presented in Section IV, and in Section V some conclusions are drawn.

II. THE MEDUSA MODEL

The speculative allocation requires a framework describing users' willingness to accept a service. In this work, we refer to the MEDUSA model presented in [14]. The MEDUSA framework describes users' acceptance of service offers from the provider with a probability A which accounts for the trade-off between utility (or perceived QoS), u, and price, p; that is $A = A(u, p)$.

The values of u and p are determined as functions of a parameter r which describes the allocated amount of resource. In our context it is reasonable to identify r with the achieved data rate. The utility functions must satisfy certain properties. In particular, as every user is willing to have as much resource as possible,

$$\frac{\mathrm{d}u(r)}{\mathrm{d}r} \geq 0\,, \tag{1}$$

The law of diminishing marginal utilities from economics, stating that:

$$\lim_{r\to\infty} \frac{\mathrm{d}u(r)}{\mathrm{d}r} = 0\,. \tag{2}$$

should also apply. In fact, there are intrinsic limitations which prevent the users from experiencing QoS beyond a certain limit; i.e. there is an upper bound to the appreciation of a service. Thus, we replace Eq. (2) with the stricter requirement:

$$\lim_{r\to\infty} u(r) = l. \tag{3}$$

In particular, in this paper we model the utilities as sigmoid curves which are well-known functions often used to describe QoS perception [1], [2]. These curves can be represented for example by the following expression:

$$u(r) \triangleq \frac{(r/K)^{\zeta}}{1+(r/K)^{\zeta}}. \tag{4}$$

Also the price is represented by a function $p(r)$ (in general, dependent on the rate) for which no particular assumptions are made, even though it seems reasonable to require that $p'(r) \geq 0$. For this reason in the following we consider a linear pricing [4] model, though other similar expressions can be used as well, without changing substantially the framework.

A suitable expression for the MEDUSA model [14] is:

$$A(u,p) \triangleq 1 - e^{-C\cdot u^{\mu}\cdot p^{-\epsilon}} \tag{5}$$

where C, μ, ϵ, are appropriate positive constants. The value of C is simply chosen according to the normalisation of utility and price, whereas μ and ϵ describe the users sensitivity to changes in utility and price, respectively.

The original proposal of the MEDUSA framework only treated *initial* service offerings and suggested a service *acceptance probability*. Here we are, however, concerned not only with users' initial acceptance of a service contract at the time of admission, but also with users' continuing willingness to hang on to the contract once they have entered. Hence we extend the Acceptance Probability model to an *Accept and Stay Probability* model defined on the service interval $[t_{\text{in}}, t_{\text{out}}]$. To do this, assume that the user achieves a decent level of satisfaction with the initial request r_{req}. It is reasonable to assume that, if the rate actually supplied determine a higher satisfaction of the user, the service will be even more appreciated and the session kept active. However, would r change in a way that would reduce the satisfaction, we assume that the user only sticks to the service with a probability conditioned on the previously being satisfied. Thus, we define the probability that a user hangs on to the service at time t_2, given that the contract at time t_1 has been accepted, to be:

$$S(t_1,t_2) = \begin{cases} \min\left(1, \frac{\min\limits_{t_1\leq t\leq t_2} A(t)}{H(t_1)}\right) & \text{if } t_{\text{in}} \leq t_2 < t_{\text{out}}, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

where $t_{\text{in}} \leq t_1 \leq t_2 \leq t_{\text{out}}$, $A(t)$ is $A(u,p)$ evaluated at time t and

$$H(t_1) = \begin{cases} A(u(r_{\text{req}}), p(r_{\text{req}})) & \text{if } t_1 = t_{\text{in}} \\ \min\limits_{t_{\text{in}}\leq t\leq t_1} A(t) & \text{if } t_{\text{in}} < t_1 \leq t_{\text{out}} \end{cases} \tag{7}$$

represents the lowest satisfaction level that has already been accepted (Fig. 1).

The probability of accepting the *session* for the wanted interval $[t_{\text{in}}, t_{\text{done}}]$, i.e. the probability of a *successful exit*, then becomes:

$$A^S(t_{\text{in}}, t_{\text{done}}) = \min\left(1, \frac{\min\limits_{t_{\text{in}}\leq t\leq t_{\text{done}}} A(t)}{A(u(r_{\text{req}}), p(r_{\text{req}}))}\right), \tag{8}$$

where $t_{\text{done}} \geq t_{\text{out}}$.

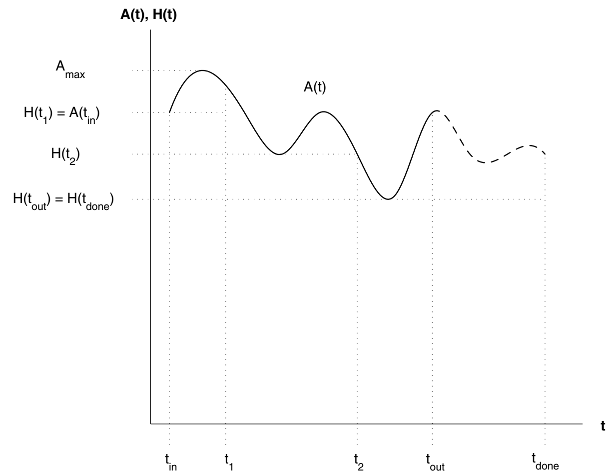

Fig. 1. Illustration of the MEDUSA extension.

If we further assume that the user always requests the rate, $\hat{r}$, that yields the best compromise between the given utility and price, and therefore maximises $A(u, p)$, Eq. (8) simplifies to:

$$A^S(t_{\text{in}}, t_{\text{done}}) = \frac{\min_{t_{\text{in}} \le t \le t_{\text{done}}} A(t)}{A(u(\hat{r}), p(\hat{r}))}. \quad (9)$$

That is, the probability of successful exit equals the MEDUSA acceptance probability for the worst utility-price combination of the session, given that the requested combination is satisfactory. The probability of successful exit is difficult to determine a priori. In fact, if the session is prematurely aborted ($t_{\text{out}} < t_{\text{done}}$), determining it a posteriori is equally difficult.

The above model does not explicitly account for the fact that user decisions are neither instantaneous nor based on instantaneous values. This is, however, partially compensated for by the moving-average procedures of the schedulers described in Sec. III. More involved modelling is outside the scope of this paper and left for further research.

III. Service Perception Aware Scheduling

The scheduling of packets can be seen as a particular case of resource assignment, the goal of which is to obtain a satisfactory provider's revenue. To this end the scheduler should, according to our previous discussion, give priority to assignments which leave the users satisfied with respect to both QoS and price paid. Unsatisfied users are expected to be likely to leave the service, thus, deteriorating revenue.

Rather than equalising the data rates or on-air times we want to maintain a high service appreciation. Time-slots are, therefore, assigned to the users that would end up the least satisfied would they *not* get the resource. More formally, assume that users initially request the rates, $\hat{r}_i$, that maximise their service acceptance probabilities, that is:

$$\hat{r}_i = \arg\max_r A(u_i(r), p_i(r)), \quad (10)$$

resulting in the initial acceptance probabilities

$$\hat{A}_i = A(u_i(\hat{r}_i), p_i(\hat{r}_i)). \quad (11)$$

Let $\bar{r}_i(n)$ be a moving average of the achieved data rate by the time of slot n:

$$\bar{r}_i(n+1) = \begin{cases} [1 - \frac{1}{\tau}]\bar{r}_i(n) + \frac{r_i(n)}{\tau} & \text{if } i \text{ served in slot n} \\ [1 - \frac{1}{\tau}]\bar{r}_i(n) & \text{otherwise} \end{cases}, \quad (12)$$

where $r_i(n)$ is the instantaneous data rate during slot n and τ is the time constant of the smoothing filter. Then, the user to be scheduled for slot n, $j(n)$, is

$$j(n) = \arg\max_{i:\ \bar{r}_i(n)[1-\frac{1}{\tau}]<\hat{r}_i} \frac{\hat{A}_i - A\left(u_i(\bar{r}_i(n)[1-\frac{1}{\tau}]), p_i(\bar{r}_i(n)[1-\frac{1}{\tau}])\right)}{\hat{A}_i}. \quad (13)$$

For comparisons with a scheduler that is more focused on technical performance we choose a PF scheduler. Like Maximum C/I scheduler, it exploits multi user diversity to increase the aggregated throughput, but unlike a Maximum-C/I scheduler it is also concerned with the individual throughputs of the users.

In [7], proportional fair resource allocation is shown to be equivalent with the constrained optimisation problem:

$$\text{maximise} \quad F(\mathrm{E}[\overrightarrow{r}]) \equiv \sum_{i=1}^{k} \log(\mathrm{E}[r_i]) \quad (14)$$

$$\text{subject to} \quad \sum_{i=1}^{k} \mathrm{E}[r_i] < C \quad (15)$$

$$\text{over} \quad \mathrm{E}[r_i] \ge 0,\ 1 \le i \le k, \quad (16)$$

where C is the system capacity and k the number of users competing for the resources. The solution to this problem can in theory be computed explicitly. In practise, however, the solution is not constant because both the system capacity and the number of users vary. It can be shown that the best one can do is to move towards the "instantaneously" best solution. Instead of the mean rate one would consider using a moving average of the rate, Eq. (12) and we get the scheduling rule:

$$\arg\max_i \frac{r_i(n)}{\bar{r}_i(n)}. \quad (17)$$

Since we do not consider buffering in this paper, we shall modify this scheduling rule not to schedule users with average rate higher than the requested and get:

$$j_{\text{PF}}(n) = \arg\max_{i:\ \bar{r}_i(n)[1-\frac{1}{\tau}]<\hat{r}_i} \frac{r_i(n)}{\bar{r}_i(n)}, \quad (18)$$

IV. Numerical Evaluation and Conclusions

To demonstrate the value of taking users' service perception into account in the radio resource management we compare the two schedulers described in Section III by means of computer simulation. To assess performance we define the following four performance measures:

- the *blocking rate* as the fraction of the users that were refused service because the system was either unable or unwilling to service them

- the *Premature Session Termination Rate* (PSTR) as the fraction of the users that decided to leave the system earlier than they intended because of poor service
- the *dissatisfaction rate* as the sum of blocking and PSTR
- the *goodput* or chargeable throughput as the throughput that the user has requested and, thus, is willing to pay for. Any excess throughput the user is supplied with is considered nonchargeable.

The PSTR is an important addition to the traditional system performance measures blocking rate and outage rate. Unlike the outage rate, which is a network-centric measure of service availability, the PSTR is a user-centric measure of the perceived service quality. A high PSTR indicates unreliable or unstable service and may, in the long run, motivate users to change service providers.

The proposed scheduling scheme is compared to the proportional fair scheduler described in Section III in a single cell hexagonal environment. The cell radius is r_{cell} and we assume that the propagation loss L between a base station and a terminal can be written

$$L(t, f, x, y) = L_{1m}\, A(d)\, B(x, y)\, F(t, f), \tag{19}$$

where t, f, x and y are time, frequency and two-dimensional spatial coordinates, respectively. $L_{1\text{m}}$ is a reference loss factor measured at 1 meter from the base station, $A(d)$ is a distance attenuation factor, $B(x, y)$ is a large-scale shadowing factor and $F(t, f)$ is a small-scale frequency selective fading factor. We assume that the distance attenuation $A(d)$ increases with the α-th power of the distance. The shadowing factor $B(x, y)$ is log-normally distributed with expectation 1 and standard deviation σ. Mobility is not modelled and, hence, the shadow fading is assumed to be constant for the duration of a session. The small-scale, or fast fading $F(t, f)$ is assumed to be an exponentially distributed random variable with expectation 1. This corresponds to Rayleigh Fading and is modelled with a Markov model to achieve time-correlation. The Doppler frequency of the channel is f_{doppler}.

The base station transmits at full power, p_{max}, in time-slots of duration t_{slot} and bandwidth W. For simplicity, extra-cell interference is assumed to be constant, lumped with the noise and denoted I_{const}. Depending on the channel conditions the base station may use different data rates from a set of rates $\mathcal{S}$. The C/I estimation is error free and the achievable data rate on the channel is given by Shannon's formula:

$$R = W\, \log_2(1 + \frac{C}{I}) \tag{20}$$

Users and their terminals are assumed to be uniformly distributed over the cell area and session inter-arrival times are exponentially distributed with mean T_{arr}. For simplicity we assume that sessions have fixed, and rather short, duration equal to T_{dur}. Users' behaviour during sessions are based on the sigmoid utility functions defined in Eq. (4), and at session setup, they are assumed to always request the rate $\hat{r}$, defined in Eq. (10), that maximises their utility function.

When a session request arrives at the base station, an admission control mechanism evaluates if there is enough free resources (time-slots) to accommodate the request. We assume that the session is admitted if the data rate corresponding to the *mean* C/I on the link, is sufficient to carry the rate r_{req} in the available time-slots, i.e. if:

$$r_{\text{req}} \leq (\text{rate of free time-slots}) \cdot R(\overline{C/I}), \tag{21}$$

and refused or blocked otherwise. However, with the SPA scheduler, the user has a second chance. If the estimated resource requirement is greater than the available resources, the admission control system may offer the user an admission at *reduced rate* if the expected revenue would thereby increase. The operator may reduce the rate only for the user under consideration or *speculate* in the effects of a global reduction of the QoS and distribute the reduction over some or all of the users. In this paper, we assume that the operator bases his speculation on equal relative rate reduction for all users. A user which rejects the counteroffer is classified as being denied service.

The numerical values of the system parameters and the parameters of the accept-and-stay probability model are summarised in Tables I and II, respectively. The evaluation is performed by means of simulation.

Results show that, with a scheduler that give priority to users that gain the most in terms of willingness to continue service, revenue can be increased significantly, while keeping user satisfaction approximately constant, compared to a proportionally fair scheduler, (Fig. 2). Alternatively, the user satisfaction can be correspondingly improved for fixed revenue. Fig. 3 displays the effective chargeable throughputs (the "goodputs"), and Fig. 4 the service denial rates, both as functions of cell load.

It is highlighted that, especially at high data rates, the SPA scheduler obtains significant performance gains. Importantly, the increase in the goodput, shown in Fig. 3, depends simply on the "best usage" of the allocatable data rates, since also the PF scheduler can be seen to fully utilise the available resource. Thus, our improvement consists in a more efficient,

TABLE I
SUMMARY OF SYSTEM PARAMETERS

Parameter	Symbol	Value
cell radius	r_{cell}	500 m
loss @ 1 m from tx	$L_{1\text{m}}$	28 dB
propagation loss exponent	α	3.5
log-normal fading std.dev.	σ	8 dB
Doppler frequency	f_{doppler}	100 Hz
slot duration	t_{slot}	2 ms
max tx power	p_{max}	33 dBm
bandwidth	W	5 MHz
interference + noise (constant)	I_{const}	-80 dBm
available rates	$\mathcal{S}$	{ 0, 64, 128, 256, 512,1024,2048} kbps
C/I thresholds	$\mathcal{G}$	$2^{R/W} - 1,\ R \in \mathcal{S}$
smoothing filter time constant	τ	25 slots
$P_{\text{tx}} = 33\text{dBm} \rightarrow SINR = -9.4\text{dB}$ @ cell border $\rightarrow$ 770 kbps		

TABLE II
ACCEPT-AND-STAY PROBABILITY PARAMETERS

Parameter	Symbol	Value
utility parameter	ζ	$2 \div 20$
utility parameter	K	$[0.05 \div 1] \cdot 256$ kbps
acceptance prob. parameter	μ	2.0
acceptance prob. parameter	ϵ	4.0
acceptance prob. parameter	C	$-(2048 \cdot 10^3)^4 \cdot \log(0.9)$
price	p	1 unit per bps

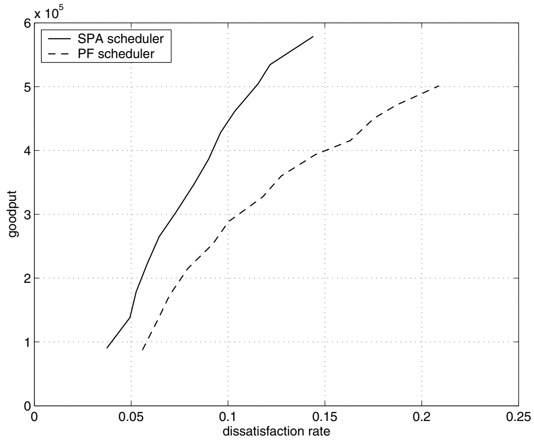

Fig. 2. Goodput plotted versus dissatisfaction rate.

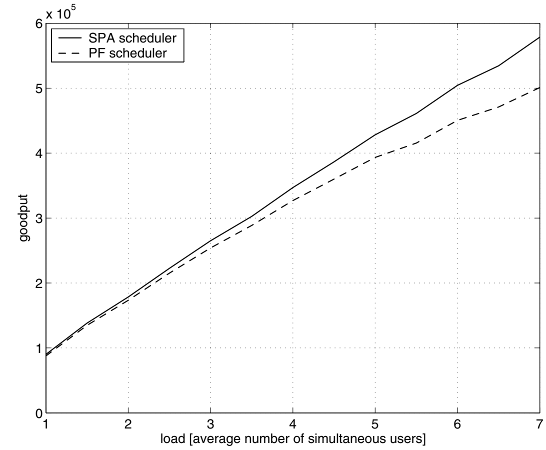

Fig. 3. Effective chargeable data rate in the cell (goodput) versus the load.

rather than higher, resource allocation to the users. This is also visible in Fig. 4, where in particular we can reasonably infer that the strong improvement of the performance in terms of lower PSTR implies a high benefit for the provider, due to the improvement in the user's/customer's satisfaction.

Indeed, a further way to exploit the improvement offered by speculative resource allocation could be to play in the trade-off between blocking and Premature Session Termination rates. In fact, even though their relative weights strongly depend on the kind of service, it is likely that many calls could be saved from being terminated due to poor quality by a less annoying block in the admission phase. This identifies another direction to exploit these results, which corresponds to the design of an appropriate proactive Admission Control (AC) strategy, which can improve the performance at a more general level.

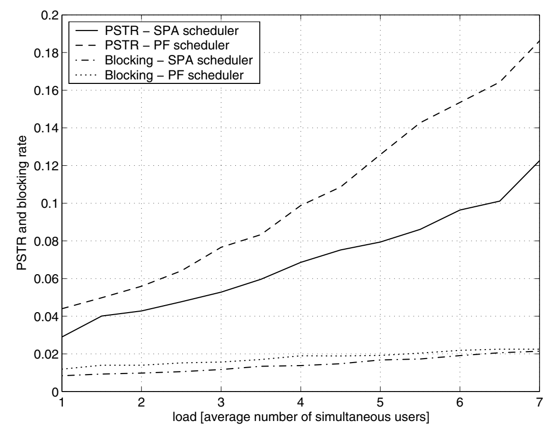

Fig. 4. Blocking and premature session termination rates, i.e., the fraction of users which refuse to access the system or decide to prematurely terminate their sessions due to perceived poor service, respectively, versus the load.

V. Conclusions

In this paper, we developed the application of speculative strategies to the downlink scheduling of a high-speed packet-switched systems. We showed by means of simulation that introducing awareness about how users react to QoS supply and pricing is an important issue, which should be carefully taken into account while determining the suitability of a scheduling strategy. This, in fact, improves both (or alternatively) users' service appreciation and revenue generation.

The importance of the scheduling order varies with the stringency of the quality requirements. With loose requirements, trying to increase users appreciation on a packet basis is less beneficial. Taking user appreciation into account is still important, however, for improving fairness with respect to user location. Results also show that, while a good scheduling policy is essential, proper admission control is important too.

References

[1] D. Famolari, N. Mandayam, D. Goodman, and V. Shah, "Wireless multimedia network technologies,", ch. 1: A New Framework for Power Control in Wireless Data Networks: Games, Utility and Pricing, pp. 289–310. Kluwer Academic Publishers, 1999.

[2] V. Siris, "Resource control for elastic traffic in CDMA networks," in *Proceedings ACM MobiCom 2002*, pp. 193–204, 2002.

[3] C. Lindemann, M. Lohmann, and A. Thümmler, "A unified approach for improving QoS and provider revenue in 3G mobile networks," *ACM Journal on Special Topics in Mobile Networks and Applications (MONET)*, 2003.

[4] R. Cocchi, S. Shenker, D. Estrin, and L. Zhang, "Pricing in computer networks: Motivation, formulation and example," *IEEE/ACM Transactions on Networking*, vol. 1, no. 6, pp. 614–627, 1993.

[5] R. J. Gibbens and F. P. Kelly, "Resource pricing and the evolution of congestion control," *Automatica*, vol. 35, pp. 1969–1985, 1999.

[6] C. Saraydar, N. Mandayam, and D. J. Goodman, "Efficient power control via pricing in wireless data networks," *IEEE Transactions on Communications*, vol. 50, no. 2, pp. 291–303, 2002.

[7] F. Kelly, "Charging and rate control for elastic traffic," *European Transactions on Telecommunications*, vol. 8, no. 3, pp. 33–37, 1997.

[8] X. Liu, E. Chong, and N. Shroff, "Opportunistic transmission scheduling with resource-sharing constraints in wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 19, pp. 2053–2064, 2001.

[9] P. Liu, R. Berry, and M. L. Honig, "Design and analysis of downlink utility-based schedulers," in *Proceedings Allerton Conference on Communication, Control and Computing*, 2002.

[10] L. Tassiulas and S. Sarkar, "Maxmin fair scheduling in wireless networks," in *Proceedings of INFOCOM 2002*, pp. 763–772, 2002.

[11] A. Bedekar, S. Borst, K. Ramanan, P. Whiting, and E. Yeh, "Downlink scheduling in CDMA data networks," in *Proceedings GLOBECOM '99*, vol. 5, pp. 2653–2657, 1999.

[12] P. A. Hosein, "QoS control for WCDMA high speed packet data," in *Proceedings MWCN 2002*, pp. 169–173, 2002.

[13] T. Moulsley, "Performance of UMTS HSDPA for data streaming," in *Proceedings Third International Conference on 3G Mobile Communication Technologies*, 2002.

[14] L. Badia, M. Lindström, J. Zander, and M. Zorzi, "Demand and pricing effects on the radio resource allocation of multimedia communication systems," in *Proceedings Globecom 2003*, 2003.