Mobile Networks and Applications 8, 209–221, 2003

A Unified Approach for Improving QoS and Provider Revenue in 3G Mobile Networks

CHRISTOPH LINDEMANN *, MARCO LOHMANN and AXEL THÜMMLER
University of Dortmund, Department of Computer Science, August-Schmidt-Str. 12, 44227 Dortmund, Germany

Abstract. In this paper, we introduce a unified approach for the adaptive control of 3G mobile networks in order to improve both quality of service (QoS) for mobile subscribers and to increase revenue for service providers. The introduced approach constantly monitors QoS measures as packet loss probability and the current number of active mobile users during operation of the network. Based on the values of the QoS measures just observed, the system parameters of the admission controller and packet scheduler are controlled by the adaptive performance management entity. Considering UMTS, we present performance curves showing that handover failure probability is improved by more than one order of magnitude. Moreover, the packet loss probability can be effectively regulated to a predefined level and provider revenue is significantly increased for all pricing policies.

Keywords: performance evaluation of next generation mobile systems, Quality of Service in mobile systems, admission control in mobile system, pricing and revenue optimization

1. Introduction

The third generation (3G) of mobile networks is expected to complete the worldwide globalization process of mobile communication. Since different parts of the worlds emphasize different issues, the global term 3G has regional synonyms: In the US and Japan, 3G often carries the name International Mobile Telephony 2000 (IMT2000). In Europe, 3G has become Universal Mobile Telecommunications System (UMTS) following the ETSI perspective. The European industrial players have created *the 3rd Generation Partnership Project* (3GPP) [1] for the standardization of UMTS. 3G mobile networks provide the foundation for new services with high-rate data not provided by current second generation systems [26]. While the standardization of 3G is still ongoing the discussion of technical issues beyond 3G has already started [23,28]. Recently, Aretz et al. reported a vision for the future of wireless communication systems beyond 3G that consists of a combination of several optimized access systems on a common IP-based medium access and core network platform [5].

Charging and pricing are essential issues for network operations of 3G mobile networks. A primary target of differentiated pricing of Internet services is the prevention of system overload and an optimal resource usage according to different daytimes and different traffic intensities [12]. Among the proposed pricing proposals, flat-rate pricing [11] is the most common mode of payment today for bandwidth services. Flat-rate pricing is popular because of its minimal accounting overhead. A flat-rate encourages usage but does not offer any motivation for users to adjust their demand. Dynamic pricing models that take the state of the network into account in the price determination have been proposed as being more responsive. Usage-based pricing regulates usage by imposing a fee based on the amount of data actually sent, whereas congestion-sensitive pricing uses a fee based on the current state of congestion in the network. Thus, a unified approach considering both dynamic pricing and controlling quality of service (i.e., performance management) provides an effective tool for the operation of 3G mobile networks. However, in previous work [8,13,19,21,25] the improvement of Quality of Service (QoS) in 3G mobile networks and the optimization of mobile service provider revenue has been considered separately.

The Quality of Service (QoS) concept and architecture for UMTS networks specified in [2] provides means for sharing radio resources among different groups of users according to their individual QoS demands. Furthermore, the concept of UMTS management and control functions such as admission controller and resource manager is roughly outlined. Das et al. proposed a framework for QoS provisioning for multimedia services in 3G wireless access networks [8]. They developed an integrated framework by combining various approaches for call admission control, channel reservation, bandwidth degradation, and bandwidth compaction. In [19], we introduced a framework for the adaptive control of UMTS networks, which utilizes online monitoring of QoS measures (e.g., handover failure and call blocking probabilities) in order to adjust system parameters of the admission controller and the packet scheduler. The presented approach is based on a lookup table called the Performance Management Information Base (P-MIB). Entries of the P-MIB have to be determined using extensive off-line simulation experiments to determine optimal parameter configuration for the considered scenarios. Given the entries of the P-MIB, we showed how to improve QoS for mobile users by periodically adjusting system parameters. The practical applicability of this approach is limited if the P-MIB comprises many en-

[*] Corresponding author.

tries (i.e., many scenarios have to be considered) because of the high computational effort for determining these entries by simulation.

This paper introduces a unified approach for the adaptive performance management for 3G mobile networks. As the main result of the paper, the introduced approach is based on a mathematical framework for the proposed update schemes rather than a lookup table. As a consequence, the adaptive control mechanism can be adjusted in an intuitive way and optimal system parameter configuration can efficiently be determined. We effectively utilize adaptive performance management for improving not only QoS for mobile users but also increase revenue earned by service providers. As in [19], controlled system parameters comprise queueing weights for packet scheduling, a threshold value of the access queue for admission of non real-time traffic, and a portion of the overall available bandwidth reserved for handover calls. Beyond [19], we propose a scheme for adjusting the queueing weights for both improving QoS for higher priority users that suffer from a high population of users with lower priority and for increasing the revenue earned by the service provider. For the analysis of the update strategy of the queuing weights, we consider a usage-based and a usage-/throughput-based pricing policy according to [11,12,21]. Furthermore, we introduce a hybrid pricing policy combining the notion of flat-rate and a usage-based pricing according to current policies of GSM networks. Performance curves derived by simulation evidently illustrate the gain of the unified approach for adaptive performance management. In fact, for UMTS networks, simulation results show that handover failure probability can be improved by more than one order of magnitude. Moreover, packet loss probability can be effectively regulated to a predefined level and the provider revenue is significantly increased for all considered pricing policies.

The paper is organized as follows. Section 2 introduces the unified approach for adaptive performance management and describes its embedding in the system architecture of 3G mobile networks. Section 3 introduces strategies for controlling the parameters of an admission controller in order to improve QoS. Section 4 describes the parameter control of a packet scheduler for the combined improvement of both QoS and provider revenue. In section 5, we present simulation results that illustrate the benefit of employing the proposed approach for adaptive performance management. Finally, concluding remarks are given.

2. Adaptive performance management for 3G mobile networks

2.1. Description of the unified approach

This section introduces the unified approach for regularly adjusting system parameters to changing traffic load, packet arrival pattern or population of users, etc. We consider a cellular mobile network in which a different transceiver station serves each cell. The purpose of the transceiver station is the modulation of carrier frequencies and demodulation of signals. Furthermore, a base station controller (BSC) is considered that is responsible for a cluster of cells, i.e., several transceiver stations. The BSC manages the radio resources, i.e., schedules data packets, and controls handovers inside the cell cluster as well as handovers towards and from neighboring cell clusters.

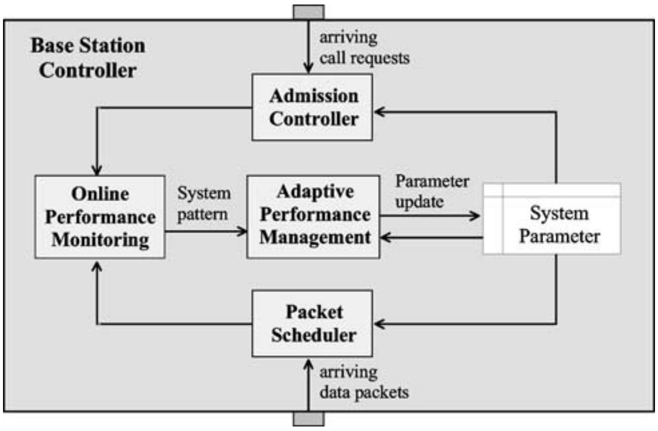


Figure 1. System architecture for adaptive performance management.

To improve QoS for mobile users as well as to increase revenue earned by service providers, an entity for *Adaptive Performance Management (APM)* is included in a BSC. Furthermore, a BSC has to be extended by an online performance monitoring component that derives QoS measures in a certain time window (e.g., handover failure probabilities of mobile users or packet loss probabilities). These QoS measures form a *system pattern* that is submitted in fixed time intervals (i.e., a control period) to the APM entity, which subsequently updates corresponding system parameters (i.e., parameters of traffic controlling components like the admission controller and packet scheduler). Thus, the proposed approach closes the loop between network operation and network control. Figure 1 shows the system architecture for performance management embedded in a BSC.

2.1.1. Online performance monitoring

System parameters of a BSC can be effectively updated by monitoring QoS measures, which are immediately affected by these parameters. A current value for a QoS measure is determined online based on a set of *relevant events* corresponding to this QoS measure (e.g., packet arrivals are relevant events for computing packet loss probabilities). The online monitoring of QoS measures is done by a sliding window technique as introduced in [19]. The width of the sliding window over time depends on the number of relevant events that are occurred according to a QoS measure. Upon arrival of a new relevant event the sliding window moves in time. At the end of a control period the QoS measures are derived for each sliding window (e.g., packet loss probability can be derived from number of lost packets divided by number of all packet arrivals in the sliding window). These QoS measures and the number of events occurred in the last control period form the system pattern that is transferred to the adaptive performance management entity (see figure 1).

Note that an accurate online monitoring of QoS measures requires a specific width for the sliding window. A certain number of events representing the history of the QoS measure have to be considered to get an expressive measure. On the other hand considering a big sliding window prevents the APM entity from fast reaction on changing traffic conditions. A bigger sliding window contains more history and, thus, more events have to be collected to cause a significant change in the online monitored QoS measure. This tradeoff between accurate online monitoring and fast reaction of the APM to changing traffic conditions has to be studied carefully in several experiments to get the optimal width of the sliding window for each QoS measure.

2.1.2. Adaptive performance management

Whenever a system pattern $S = \{(P_1, n_1), \dots, (P_m, n_m)\}$, consisting of online monitored QoS measures $P_1, \dots, P_m$ and the numbers of relevant events $n_1, \dots, n_m$ occurred in the last control period is transmitted to the APM an update of the system parameters can be performed. In general, an update of a system parameter σ is made according to a function f depending on a subset of the QoS measures $P_1, \dots, P_m$ and the previous value $\sigma^{(\mathrm{old})}$ of the system parameter. Let $P_{\tau(1)}, \dots, P_{\tau(k)}$, $k \leqslant m$, be the QoS measures corresponding to system parameter σ, then the update is made if a certain minimum number $n(\sigma)$ of relevant events occurred in the last control period. That is:

$$\sigma^{(\mathrm{new})} = f\big(P_{\tau(1)}, \dots, P_{\tau(k)}, \sigma^{(\mathrm{old})}\big), \quad \text{if } \min\{n_{\tau(1)}, \dots, n_{\tau(k)}\} \geqslant n(\sigma). \tag{1}$$

We classify update functions in *relative* functions, that perform a parameter update relative to the old parameter value and *absolute* functions that set the new parameter value independent of the old value, i.e., f is independent of $\sigma^{(\mathrm{old})}$ in (1). With relative update functions strong fluctuations of the corresponding system parameter in one update step can be avoided. In section 3, we study a special class of relative update functions in order to set the parameters of an admission controller. Furthermore, we develop in section 4 an absolute update function for adjusting the weights of a weighted fair queueing packet scheduler.

2.2. Economics and pricing policies in 3G mobile networks

There are multiple requirements, which should be fulfilled for any viable pricing mechanism in multi-service class data communication networks [12]. A primary target of differentiated pricing of Internet services is the prevention of system overload and an optimal resource usage according to different daytimes and different traffic intensities. Furthermore, the pricing scheme should be implemented in a completely decentralized manner and there should be multiple priorities in order to take into account the different QoS required by different applications and users.

In general, pricing policies can be partitioned into usage-based (pay-as-you-go) pricing, flat-rate (all-you-can-eat) pricing, and dynamic pricing. In usage-based pricing policies a user is charged according to a connection time or traffic volume. Whereas connection based calls (e.g., in GSM) are charged by connection time, packet-switched services (e.g., in UMTS) are charging the transferred data volume. Dynamic pricing models take into account the state of the mobile radio network for determining the current price of a service. Congestion-sensitive pricing as a particular dynamic pricing model has been shown to be more responsive. MacKie-Mason and Varian introduced the concept of congestion-sensitive pricing in their smart market scheme [21]. Under this model, the actual price for each packet is determined based on the current state of network congestion. In [25], Rao and Petersen discussed the optimal pricing of priority services. Analogously to the smart market approach, Gupta et al. presented a pricing scheme that uses priorities on the packet-level [13]. They proposed to differentiate Internet traffic according to delay and loss requirements.

For the analysis of the update strategy of the queuing weights, we consider in section 4 a usage-based and a usage-/throughput-based pricing policy according to [11,12,21]. Furthermore, we introduce a hybrid pricing policy combining the notion of flat-rate and a usage-based pricing according to current policies of GSM networks.

3. Strategies for improving Quality of Service

3.1. Admission controller

The proposed approach distinguishes three different types of services: circuit-switched services, packet-switched real-time services (RT), and packet-switched non real-time services (NRT). Typically, circuit-switched services are voice calls from a GSM mobile station. As proposed by 3GPP, RT services belong to the conversational and streaming classes and NRT services fall into the interactive and background classes [2]. The bandwidth available in a cell must be shared by calls of these different service classes and the different service requirements have to be met. Before a mobile session begins, the user needs to specify its traffic characteristics and desired performance requirements by a *QoS profile*. Then, an admission controller decides to accept or reject the users request based on the QoS profile and the current network state as, e.g., given by queueing length. The purpose of the admission controller is to guarantee the QoS requirements of the user who requested admission while not violating the QoS profiles of already admitted users. The call admission criteria will be different for each service class. The QoS profile for RT sessions specifies a guaranteed bandwidth to be provided for the application in order to meet its QoS requirements. If the network cannot satisfy the desired bandwidth, the corresponding admission request is rejected.

Data packets arriving at the BSC are queued until they are scheduled to be transmitted over the radio link. For NRT sessions, we consider an admission controller taking into account free buffer space in the NRT queue [8]. In order to prevent buffer overflow once a call is admitted, the current queueing length is set against certain buffer availability threshold of the capacity, denoted by η. The admission criteria for voice and RT handovers are the same as for new voice calls and RT sessions except that additional handover bandwidth can be utilized. The analysis of several admission control schemes for cellular systems presented in [24] showed that the simple reservation scheme (i.e., reserving bandwidth for handover calls) performs remarkably well. For simple cellular networks, the optimal amount of bandwidth reserved for handover calls can be determined by analytical models [14]. In the model presented here, we denote with b_h the portion of the overall bandwidth that is exclusively reserved for handover calls from neighboring cells. The considered admission controller does not prioritize NRT handovers over new NRT sessions. Further details of the admission controller are given in [19].

3.2. *Adjusting the admission controller for QoS improvement*

In this section, we show how to utilize equation (1) for setting the parameters η and b_h of the admission controller in order to reduce packet loss probability and handover failure probability. For updating the system parameters, we split the general function introduced in section 2.1 into separate functions each depending only on one QoS measure. Let $P_1, \ldots, P_k$ be the QoS measures corresponding to a system parameter σ. Then, equation (1) can be simplified to

$$\sigma^{(\mathrm{new})} = \frac{f_1(P_1) + \cdots + f_k(P_k)}{k} \cdot \sigma^{(\mathrm{old})},$$
$$L \leqslant \sigma^{(\mathrm{new})} \leqslant R. \tag{2}$$

The interpretation of (2) is the following. Each update function f_i describes the influence that the QoS measure P_i should have on the system parameter σ. Subsequently, the overall update is performed by computing the arithmetic mean of the functions f_i multiplied with the old value of the system parameter. Note that the value $\sigma^{(\mathrm{new})}$ must be truncated at a certain lower bound L and an upper bound R in order to guarantee that the computation of $\sigma^{(\mathrm{new})}$ results in a valid value of the system parameter. As basic update function we consider a logarithmic linear function of the form:

$$f_i(P_i) = m_i \log P_i + b_i. \tag{3}$$

The reason for this choice is that we want to consider QoS measures like loss probabilities and failure/blocking probabilities, which are in the range of 10^{-5} to 1. Therefore, a logarithmic shape is more suitable. In previous work [19], we have studied update schemes of system parameters of an admission controller and a packet scheduler based on a lookup table. In order to determine the optimal entries of this lookup table extensive off-line simulation experiments have been conducted. Applying regression statistics to the entries of this lookup table shows that these entries are well represented by functions with logarithmic shape. Thus, besides the motivation of the update functions given here, their choice is to a large extend originated from regression statistics conducted in earlier work. The strength of the influence of f_i on $\sigma^{(\mathrm{new})}$ can be adjusted with the gradient m_i. The parameter b_i can be determined by the following interpretation: suppose the *desired level* of the QoS measure P_i is β_i (e.g., the desired packet loss probability is 0.001). That is, if the online measured value of P_i is β_i the system parameter σ should not be changed in the update step from the point of view of measure P_i. Therefore, we chose $f_i(\beta_i) = 1$ and from this relation we get $b_i = 1 - m_i \log \beta_i$. Inserting in equation (3) results in the final form of the update function:

$$f_i(P_i) = m_i \log \frac{P_i}{\beta_i} + 1. \tag{4}$$

For ease of notation, we abbreviate the QoS measures handover failure probability and new call/session blocking probability corresponding to voice calls and RT sessions by HFP and CBP, respectively. The probability of a packet loss due to buffer overflow in the NRT queue is abbreviated by PLP. The update strategy according to equations (2)–(4) is justified by its intuitive understanding and the performance results presented in section 5. The suitability of update functions other than (2)–(4), is subject for further study and out of the scope of this paper.

3.2.1. *Update of non real-time queue threshold*

Recall that a system parameter update is performed each time a system pattern arrives at the APM entity and the minimum number of relevant events corresponding to this system parameter is reached. Determining the update for the system parameter η, i.e., determining $\eta^{(\mathrm{new})}$, is performed corresponding to the old value $\eta^{(\mathrm{old})}$ and the actually observed QoS measure PLP. That is:

$$\eta^{(\mathrm{new})} = f(\mathrm{PLP}) \cdot \eta^{(\mathrm{old})}, \quad 0.001 \leqslant \eta^{(\mathrm{new})} \leqslant 1. \tag{5}$$

The truncation of $\eta^{(\mathrm{new})}$ at the lower bound guaranties that the value does not accumulate near zero for long periods of low traffic load. The minimum number of relevant events required for an update of η is counted in data volume rather than in packet arrivals (in the experiments this number is 5 MB). The setting of the gradient m of the corresponding update function is derived from a couple of experiments for different values of the gradient. We found $m = -0.02$ to be suitable. Choosing a suitable value for the gradient is a similar tradeoff as explained for the sliding window size. A large gradient results in a fast update of the system parameter in a few number of update steps, but also introduces higher fluctuations of the system parameter over time. We demonstrate the speed of the parameter adjustment in an experiment in section 5. Furthermore, several experiments for different desired loss values β are presented.

3.2.2. *Update of fraction of bandwidth reserved for handover*

The update for the system parameter b_h, i.e., determining $b_\mathrm{h}^{(\mathrm{new})}$, is performed based on the old value and the actually observed QoS measures HFP and CBP. That is:

$$b_\mathrm{h}^{(\mathrm{new})} = \frac{f_1(\mathrm{HFP}) + f_2(\mathrm{CBP})}{2} \cdot b_\mathrm{h}^{(\mathrm{old})},$$
$$0.001 \leqslant b_\mathrm{h}^{(\mathrm{new})} \leqslant R. \tag{6}$$

The value $b_{\mathrm{h}}^{(\mathrm{new})}$ is truncated at a lower bound of 0.1% and a certain upper bound R which is a fraction of the overall bandwidth available (in the experiments we fix $R = 0.7$). The truncation at the lower bound is for the same reason as explained above. In fact, for computing $b_{\mathrm{h}}^{(\mathrm{new})}$ two QoS measures corresponding to the actually observed HFP and CBP are taken into account. A high HFP should increase $b_{\mathrm{h}}^{(\mathrm{new})}$ but this obviously also increases the CBP because less bandwidth is available for new voice calls and RT sessions. Therefore, the HFP and the CBP influence the handover bandwidth $b_{\mathrm{h}}^{(\mathrm{new})}$. In fact, $m_1 = -m_2$ holds in the update functions f_1 and f_2. From a couple of experiments for different gradients, we found $m_1 = 0.08$ to be suitable. A common assumption in cellular networks is to prioritize handover calls over new calls. Therefore, the desired handover failure level β_1 should be smaller than the desired call blocking level β_2. According to these values the handover bandwidth is slightly increased, if HFP is equal to CBP.

With the presented strategy the parameters of the update functions can be chosen in an intuitive way and optimal parameter configuration can efficiently be determined. This is the major advantage over the approach based on a Performance Management Information Base introduced in [19] which requires extensive off-line simulation experiments.

4. Strategies for improving both QoS and provider revenue

4.1. Packet scheduler

At a BSC responsible for a cluster of cells, data packets from various connections arrive and are queued until bandwidth for transmission is available. In order to distinguish different priorities for NRT traffic corresponding to the traffic handling priority defined by 3GPP [2], scheduling algorithms like Weighted Round Robin (WRR), Weighted Fair Queueing (WFQ [9]) or Class Based Queueing (CBQ [10]) have to be implemented. An overview of queueing issues for guaranteed performance services can be found in [27]. In WFQ, the weights control the amount of traffic a source may deliver relative to other active sources during some period of time. From the scheduling algorithm's point of view, a source is considered to be active, if it has data queued in the NRT queue. Let B be the overall bandwidth available for NRT sessions at time t. For an active source i with weight w_i, the bandwidth B_i that is allocated to this transfer at time t is given by

$$B_i = \frac{w_i}{\sum_j w_j} \cdot B. \tag{7}$$

In (7) the sum is taken over all active NRT sources j. A class based version of WFQ serves packets of each priority class according to the weights rather than every active source.

4.2. Adjusting the packet scheduler for QoS and revenue improvement

This section utilizes the proposed approach for the adaptive control of the weights of a weighted fair queueing packet scheduler in order to improve QoS as well as to increase the revenue. The strategy for adjusting the weights combined with the introduction of several pricing policies constitutes a further contribution of the paper. Recall that the revenue earned by a mobile service provider is determined by the monthly payment of mobile users as well as by the additional usage-based pricing after the monthly amount of data volume is consumed. Note, that the monthly subscription rate is only relevant for monthly revenue calculations. In this section, we consider the revenue improvement in a certain small time period regardless the monthly subscription rates. In section 5, we briefly discuss monthly revenue calculation. Let P denote the number of different priority classes, i.e., weights of the weighted fair queueing scheduler. Define by $b_i(t)$ the transferred data volume in time t of users of priority i and by $r_i(t)$ the payment of users of priority i at time t, i.e., the user pays for the transferred data volume. We distinguish a pure usage-based and a usage-/throughput-based pricing policy:

(a) A user of priority i has a fixed payment p_i per kbit during his session, i.e., $r_i(t) = p_i$.

(b) The payment of a user of priority i consists of a fixed part p_i that is increased proportional to the additional throughput $\varphi_i(t)$ he received due to the update of the queueing weights, i.e., $r_i(t) = p_i\varphi_i(t)$.

According to the proposed data volume based pricing with respect to different priority classes the revenue function $\Phi(t)$ is given by

$$\Phi(t) = \sum_{i=1}^{P} r_i(t) b_i(t). \tag{8}$$

The revenue function of equation (8) is utilized in section 5 for evaluating the strategies for revenue improvement presented below.

4.2.1. Update of WFQ weights

Recall that packets of NRT users arriving at the BSC are first queued until they are scheduled for transfer by a weighted fair queueing discipline. Let $w_i \geqslant w_{i+1},\ i = 1, \dots, P-1$, be the basic weights of the WFQ scheduler. The update of the queueing weights, i.e., determining $w_i^{(\mathrm{new})}$ is made according to an absolute update function depending on the basic weights w_i and the current number of NRT sessions belonging to priority i. Therefore, every system pattern that is transmitted from the online monitoring component to the adaptive performance management entity contains the current number of active NRT sessions with priority i in the cell. For ease of notation, the number of active non real-time sessions with priority i is abbreviated by NRT_i.

The idea behind the strategy for revenue improvement is to shift the overall utilization of bandwidth for NRT traffic towards higher priority users, which pay more for the transferred data volume. Note that the update strategy should be conservative in a way that the transfer of packets of low priority is not simply blocked if packets of higher priorities are arriving, i.e., priority queueing. Assuming that the majority of users will buy a cheaper low priority service class, priority queueing will leave most users unsatisfied. Therefore, the update strategy also considers the QoS aspect. The update strategy concerning the queueing weights is developed according to the following premises:

(i) If the number of active NRT users in the cell is the same for each priority class, i.e., $\text{NRT}_i = \text{NRT}_j,\ i \neq j$, the weights $w_i^{(\text{new})}$ should be set according to the basic weights w_i for $i = 1, \ldots, P$.

(ii) Priority classes with low population of users compared to other classes should be prioritized, i.e., the corresponding weights should be increased.

(iii) The relative ordering of the weights should be preserved in a strong way, i.e., $w_i^{(\text{new})} \cdot (w_i/w_{i+1}) \cdot w_{i+1}^{(\text{new})}$ for $i = 1, \ldots, P-1$.

Premise (i) constitutes the key of the update strategy. If all priority classes have the same population of users the scheduling should work as in the case without adaptive control of the weights. The rationale behind premise (ii) is to prioritize users that are consuming less bandwidth (relative to their weights) than users belonging to other classes, i.e., users of low population should be made more independent from the influence of user classes with higher population. This premise constitutes the basic idea for QoS improvement and is demonstrated by the following example that considers two priority classes, i.e., a high and low priority class. In WFQ the available bandwidth is shared among *all* active users according to their weights. That is, if the minority are high priority users, the overall bandwidth consumed by these users will suffer from a strong influence of low priority users that hold the majority. Therefore, increasing the weights for high priority users will result in a higher QoS for this user class. Updating the weights according to this strategy will result in a scheduling algorithm somewhere between a WFQ and a class based queueing scheduler. In fact, the benefit of both is utilized: the fair sharing of the bandwidth of WFQ and the higher bandwidth guarantees for each priority class provided by a class based queueing scheduler.

Preserving the relative ordering of the weights (i.e., premise (iii)) guarantees that QoS for higher priority users and, therefore, the provider revenue can only be improved due to the adaptive control of the weights. If the intention of the update strategy is not primary on improving provider revenue the weights can be also set in a weak relation, i.e., $w_i^{(\text{new})} \geqslant w_{i+1}^{(\text{new})}$. This might be useful to increase QoS for users of low population independent of their priority class. With the following algorithm the computation of the weights $w_1^{(\text{new})}, \ldots, w_P^{(\text{new})}$ can be performed iteratively in $P-1$ minimum calculations. The iteration is given by

$$w_1^{(\text{new})} = w_1 \cdot (\text{NRT}_1)^{-\alpha}, \tag{9}$$

$$w_i^{(\text{new})} = \min\left(\frac{w_i}{w_{i-1}} \cdot w_{i-1}^{(\text{new})},\ w_i \cdot (\text{NRT}_i)^{-\alpha}\right), \quad i = 2, \ldots, P. \tag{10}$$

In order to smooth the influence of the number of NRT users on the queueing weights, an exponent $\alpha \geqslant 0$ is considered (e.g., $\alpha = 1/2$). It is easy to show that premises (i), (ii) and (iii) hold for the weights set according to equations (9) and (10). The iteration starts with setting $w_1^{(\text{new})}$ according to NRT_1 and continues up to $w_P^{(\text{new})}$. Note that this is only one possibility to set the new weights. Any other starting position for the iteration is possible and results in a slightly different update of the weights. Nevertheless, the algorithms work in a similar way, and therefore, we consider only the iteration of (9) and (10). If currently no users of priority i are in the cell, i.e., $\text{NRT}_i = 0$, the algorithm skips the setting of the corresponding weight $w_i^{(\text{new})}$ and the next iteration step $i+1$ is related to step $i-1$. Subsequently, these weights were set to zero. For other scheduling disciplines like weighted round robin or a class based queueing corresponding update strategies can be derived in a similar way.

4.2.2. *Considering advanced pricing policies*

In pricing policy (b) introduced above, users have to pay an additional fee depending on the throughput improvement due to the update of the queueing weights. This concept of pricing indicates strong similarities to the congestion-sensitive pricing of the smart market scheme [21], where the actual price for each packet is determined based on the current state of network congestion. Similarly, in our throughput-based pricing policy the throughput of users is determined by their willingness-to-pay additional costs (according to their choice of priority class) for transmission of packets in a congested network. The additional payment is justified because the throughput for users of higher priority will be maintained, even if more and more users of lower priority attend the cell, i.e., the network is currently congested. We describe the relative throughput increase of priority class i with the function

$$\varphi_i(t) = \varphi_i^{(\text{new})} = \left(\frac{w_P \cdot \text{THR}_i}{w_i \cdot \text{THR}_P}\right)^{\gamma}. \tag{11}$$

In equation (11), THR_i is the current throughput of class i derived from the corresponding sliding window and $0 \leqslant \gamma \leqslant 1$ is a scaling exponent (e.g., $\gamma = 1/4$) that has to be adjusted by the service provider for appropriate revenue dimensioning. In order to guarantee that revenue will be only improved, $\varphi_i(t)$ has to be truncated, i.e., $\varphi_i(t) \geqslant 1$.

Next, we adjust the weights according to an advanced pricing policy that adopts ideas, which have been successful in existing GSM networks. In GSM networks, the pricing of a provided service is as follows: the proposed service is offered based on a monthly payment for a dedicated amount of call time. If a user has consumed this amount of time before the end of the month, he has to pay for any further use of this service based on a time-dependent accounting. This idea can be generalized and extended towards packet-switched services in 3G networks. Analogously, a user has to pay a monthly charge for a dedicated amount of data volume, which can be transferred without further pricing. After using up this monthly amount of data, the user has to pay for the desired services according to the transferred data volume (byte-based). Moreover, analogous to GSM networks a user can utilize "unused" data volume, i.e., the unused fraction of the prepaid monthly amount of data volume, in subsequent months. If the monthly amount of data is unrestricted, this pricing would become a flat-rate pricing and if there is no monthly payment, the pricing follows a usage-based policy. Thus, our pricing policy constitutes a hybrid approach of flat-rate and usage-based pricing.

The update of the queueing weights can now be extended in a way that users consuming their monthly amount of data are served with a lower priority than users currently paying for their data transfer. Therefore, we introduce a new weight w' corresponding to the not paying users. The weight w' must be sorted in the weights $w_1, \dots, w_P$ and the iterative update algorithm (9)–(10) can be applied to the $P + 1$ weights as described above. In order to distinguish not paying users with different priorities these users are served by the WFQ scheduler with weights $w_1, \dots, w_P$ relative to w'. That is, WFQ is applied to $2 \cdot P$ weights, i.e., $w_1, \dots, w_P$ and $(w'/w) \cdot w_1, \dots, (w'/w) \cdot w_P$ with $w = w_1 + \cdots + w_P$.

4.3. Implementation issues

As outlined in section 2.1, the controlled system parameters for QoS and revenue improvement, i.e., η, b_{h}, w_i, φ_i, constitute an integral component of the proposed extension to a BSC. The adjustment of system parameters is only based on implicit information that is directly measured by the online monitoring component. Therefore, no additional signaling with other BSCs is necessary for updating system parameters. The online monitored QoS measures, i.e., PLP, HFP, CBP, NRT_i, and THR_i, can easily be derived and stored within the BSC (see figure 1). The PLP can directly be determined by counting the number of IP packets, which are lost due to buffer overflow in the NRT queue. HFP as well as the CBP is determined by the non-admitted handover calls and new calls in the admission controller, respectively. Admission, termination, and handover of NRT calls enable the profiling of NRT_i, the number of non real-time sessions with priority i. Moreover, the packet scheduler allows the throughput computation of NRT users according to their individual priorities. Furthermore, no time consuming signaling is needed to transfer the system pattern inside the BSC because the online performance monitoring component and the performance management entity both reside in the BSC.

The question arises how call charging can be accomplished for the considered pricing policies in 3G mobile networks. For pricing policy (a), i.e., a fixed payment per kbit, call charging can easily be processed by the subscription management component of the operation subsystem (OSS) by means of the call charging mechanism using the home location register (HLR) [1]. Similarly, the hybrid pricing scheme can be realized except that the remaining amount of prepaid data volume has to be stored in the HLR for charging the transferred data volume. Utilizing these existing charging mechanisms, no additionally signaling overhead arises for charging data services. The throughput-based pricing policy (pricing policy (b)) just slightly changes the situation and can easily be implemented within the BSC using a local copy of the user's HLR charging data fields. This local data minimizes signaling overhead of individual user charging. According to the transferred data volume and current throughput of the user's bandwidth class, this local charging profile is continuously updated. Handovers with changing BSC of response induce the transfer of this local charging profile to the new BSC of response. Subsequently, these local data have to be updated in the HLR for individual user accounting after termination of the call. Note, that this transfer of local charging profiles can naturally be embedded in the OSS functionality.

5. Evaluation of the adaptive performance management strategies

5.1. Traffic characterization for 3G networks

For traffic modeling of RT applications we utilize the approach proposed in [18], where variable bit rate video traffic is modeled in terms of time-discrete M/G/∞ input processes. This model is based on measured video streams and efficiently captures the correlation structure of the considered video traffic applying the time-discrete M/G/∞ input process. The generated traffic is transformed utilizing a hybrid Gamma/Pareto numerical transformation in order to capture the marginal distribution of the measured traffic stream. Subsequently, the synthetically generated traffic is broken down to IP packets of a maximum size of 1500 bytes, which are uniformly distributed within a given frame-duration of the MPEG video sequence comprising of $1/30$ s. Note that this traffic model does not propose information for modeling RT session durations. Therefore, we assume session durations to be exponentially distributed (see section 5.2).

Recent recommendations for modeling NRT traffic and analytical traffic models for 3G mobile networks are proposed in [15,16], respectively. The traffic model is based on real measurements conducted at an Internet service provider dial-in link, which comprises comparable characteristics of future mobile networks [17], i.e., different access speeds, influence of the user behavior due to different tariff limits, as well as asymmetric up- and downlink traffic. Based on these measurements a NRT traffic model is conducted, applying the idea of the single user traffic model, which describes traffic characteristics on session-level, connection-level, i.e., application-level, and packet-level, respectively. The key insight of this modeling approach lies in an appropriate scaling procedure of

Table 1
Characteristics for different UMTS session types.

	Circuit switched	Streaming real time (RT)		Interactive non real time (NRT)		
	voice service	Audio	Video	high priority	normal priority	low priority
Portion of arriving requests	25%	12%	3%	6%	18%	36%
Session duration	120 s	180 s		determined by session volume distribution		
Session dwell time	60 s	120 s		120 s		

the measured trace data towards typical bandwidth classes of 3G mobile networks, i.e., 64 kbps, 144 kbps, and 384 kbps. In this context, a bandwidth class denotes the maximum bandwidth capability of future handheld devices. We refer to [15] for details of the NRT traffic model, especially for the parameterization of the traffic characteristics.

5.2. The simulation environment

In order to evaluate the proposed approach for adaptive control, we developed a simulation environment for a UMTS access network, i.e., a *UMTS Terrestrial Radio Access Network* (UTRAN [3]). The simulator considers a cell cluster comprising of seven hexagonal cells with corresponding transceiver stations (i.e., *Node B* elements), that are managed by a base station controller (i.e., a *Radio Network Controller*, RNC). We assume that a mobile user requests a new *session* in a cell according to a Poisson process. When a mobile user starts a new session, the session is classified as voice, RT, or NRT session, i.e., with the session the user utilizes voice, RT, or NRT services mutually exclusive. RT sessions consist of streaming downlink traffic corresponding to the UMTS streaming class specified by 3GPP [2] and NRT sessions consist of elastic traffic and correspond to the UMTS interactive class or background class, respectively. For the year 2010 an amount of about 50% voice calls is anticipated [26]. We assume that one half of the voice calls are served over the frequency spectrum for traditional GSM services (i.e., 890–915 and 935–960 MHz) and the second half is served over the new frequency spectrum allocated for UMTS. Nevertheless, the simulator considers only the new frequency spectrum. Therefore, we assume that 25% of the call requests are voice calls whereas RT and NRT sessions constitute 15% and 60% of the overall arriving requests (see table 1).

Subsequently, we have to specify the QoS profile for RT and NRT sessions. For RT sessions the simulator considers two QoS profiles, i.e., a low bandwidth profile comprising of a guaranteed bit rate of 64 kbps corresponding to streaming audio and a high bandwidth profile comprising of a guaranteed bit rate of 192 kbps corresponding to streaming video. According to the RT traffic model presented in section 5.1, we assume that 80% of the RT sessions utilize the low bandwidth profile whereas the remaining 20% utilize the high bandwidth profile. Following the single user traffic model, NRT sessions are partitioned according to different bandwidth classes as follows: 60% for 64 kbps, 30% for 144 kbps, and 10% for 384 kbps, comprising of different priorities (see table 1), respectively.

The amount of time that a mobile user with an ongoing session remains within the cell is called *dwell time*. If the session is still active after the dwell time, a handover toward an adjacent cell takes place. The *call/session duration* is defined as the amount of time that the call will be active, assuming it completes without being forced to terminate due to handover failure. We assume the duration of voice calls and RT sessions to be exponentially distributed. As proposed in [6], the dwell time is modeled by a lognormal distribution. All corresponding mean values are shown in table 1. A NRT session remains active until a specific data volume drawn according to a bandwidth-dependent lognormal distribution is transferred. To distinguish between NRT traffic classes, the UMTS simulator implements a WFQ scheduler with three packet priorities: 1 (high), 2 (normal), and 3 (low) with weights $w_1 = 4$, $w_2 = 2$, and $w_3 = 1$. These priorities correspond to the traffic handling priority specified by 3GPP. To model the user behavior in the cell, the simulator considers the handover flow of active mobile users from adjacent cells. The iterative procedure introduced in [4] is employed for balancing the incoming and outgoing handover rates. The iteration is based on the assumption that the incoming handover rate of a user class at step $i + 1$ is equal to the corresponding outgoing handover rate computed at step i.

5.2.1. UMTS system model assumptions

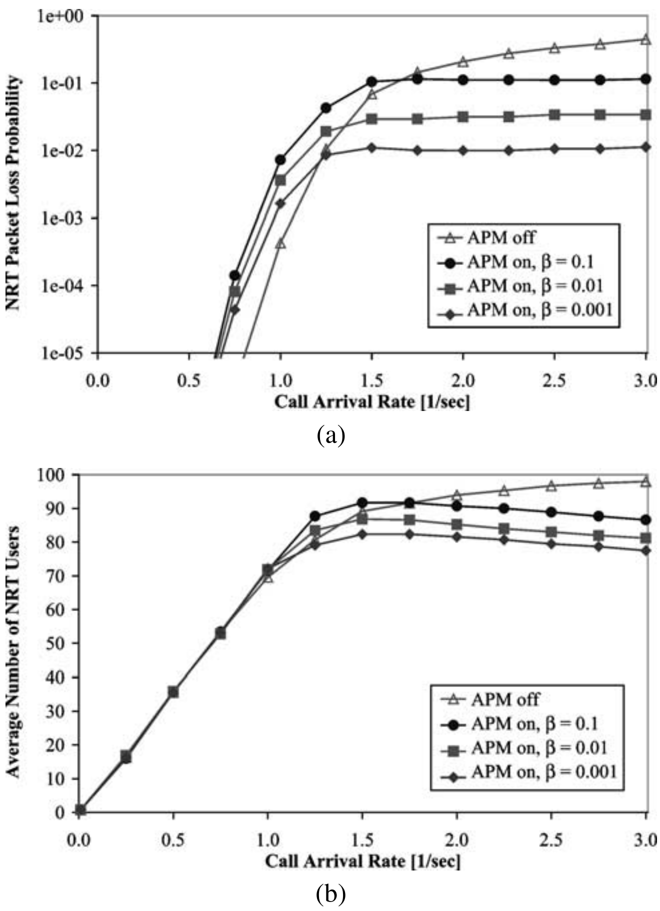
The simulator exactly mimics UMTS system behavior on the IP level. The focus is not on studying link level dynamics. Therefore, we assume a reliable link layer as provided by the automatic repeat request (ARQ) mechanism of the Radio Link Control (RLC) protocol. As shown in [22] for the General Packet Radio Service (GPRS), the ARQ mechanism is fast enough to recover from packet losses before reliable protocols on higher layers (e.g., TCP) recognize these losses due to timer expiration. Thus, a reliable link level can be assumed when considering higher layer protocol actions (see, e.g., [20]). To accurately model the UMTS radio access network, the simulator represents the functionality of one radio network controller and seven Node B transceiver stations, one for each of the considered cells. Since in the end-to-end path, the wireless link is typically the bottleneck, and given the anticipated traffic asymmetry, the simulator focuses on resource contention in the downlink (i.e., the path RNC → Node B → MS) of the radio interface.

The simulator considers the UTRAN access scheme based on Wideband-Code Division Multiple Access (W-CDMA) in Frequency Division Duplex mode (FDD) proposed by 3GPP [1]. In FDD downlink, a division of the radio frequencies into four physical code channels with data rates of 1,920 kbps each up to 512 physical code channels with 15 kbps data rates each is possible. Therefore, the overall bandwidth that is available in one cell is 7,680 kbps. For the channel coding, we assume a convolution-coding scheme with coding factor 2. In the experiments without adaptive control the handover bandwidth portion b_{h} is 5% and the NRT queue threshold η is set to 95%. The simulation environment was implemented using the simulation library CSIM [7]. In a presimulation run the handover flow is balanced, for each cell at the boundary of the seven-cell cluster. All simulation results are derived with confidence level of 95% using the batch means method. The execution of a single simulation run requires about 40–60 min of CPU time (depending on the call arrival rate) on a dual processor Sun Sparc Enterprise with one GByte main memory.

5.2.2. Implementation of the hybrid pricing policy in the simulator

According to the hybrid pricing policy as introduced in section 4.2.2, the user's overall remaining amount of prepaid data volume d out of the user's monthly data volume D is determined at the beginning of a session. Moreover, the remaining amount of data volume of previous months r is determined. For simulation study purposes, this is accomplished by choosing the random value d uniformly out of the interval $[0, kD]$. kD captures the monthly amount of data a user typically transfers, i.e., a user typically transfers a multiple k of the data volume D that is available for a fixed monthly payment. The random value r, is sampled according to a uniform distribution out of the interval $[0, 0.1D]$, where $0.1D$ measures the maximum amount of "unused" data volume of previous months. If d exceeds $D + r$ the user has no remaining prepaid data volume, including the data volume of the current and the previous months. Otherwise, there is a remaining amount of prepaid data volume $D + r - d$ for the considered user and additional pricing arises only, if the transferred data volume of the user session exceeds $D + r - d$. Thus, during the user session the remaining data volume has to be updated according to the actually transferred data.

In the simulation studies we utilize the proposed hybrid-pricing scheme with a prepaid monthly data volume of 150 MB. According to the different priority classes 1, 2, and 3, the volume-based pricing for transferred data exceeding the prepaid monthly data volume comprises of 20, 15, and 10 cost-units per MB, respectively. Considering the changing traffic loads according to the daytime, this approach can be refined, by the notion of different pricing for daily periods of time. For the parameterization of the typically monthly transferred data volume, we assume $k = 2$. Note that the parameterization of the pricing scheme is chosen for demonstration purposes only. Due to the high flexibility of the hybrid pricing scheme, it can be easily extended towards multiple, concurrent pricing schemes comprising of, e.g., different monthly amounts of prepaid data volumes, different payments for the individual priority classes, or a pure usage-based pricing as well as pure flat-rate pricing.



(a)

(b)

Figure 2. Impact of adaptive performance management on non real-time traffic.

5.3. Performance results

Using simulation experiments, we illustrate the benefit of the proposed unified approach for adaptive performance management of UMTS systems. In particular, we show the improvement of QoS measures and the increase in revenue earned by service providers. The presented curves plot the mean values of the confidence intervals for the considered QoS measures. In almost all figures, the overall call/session arrival rate of new mobile users is varied to study the cell under increasing load conditions. For ease of notation, results with and without adaptive performance management (APM) are abbreviated by *APM on* and *APM off*, respectively.

In a first experiment, we investigate the effect of adaptive control on the threshold for the buffer size of the NRT queue denoted by η. Figure 2 shows the NRT packet loss probability (a) and the average number of NRT users in the cell (b) for the UMTS system with and without adaptive control. Furthermore, the figures distinguish between different desired loss levels β as introduced in section 3.2. We observe that the APM achieves a substantially decrease in packet loss probability. Moreover, the packet loss probability can be kept below a constant level for increasing arrival rates of mobile users. Note, that this level slightly differs from the desired level of the QoS measure. This is due to the fact that the update function only decreases the NRT threshold if the online

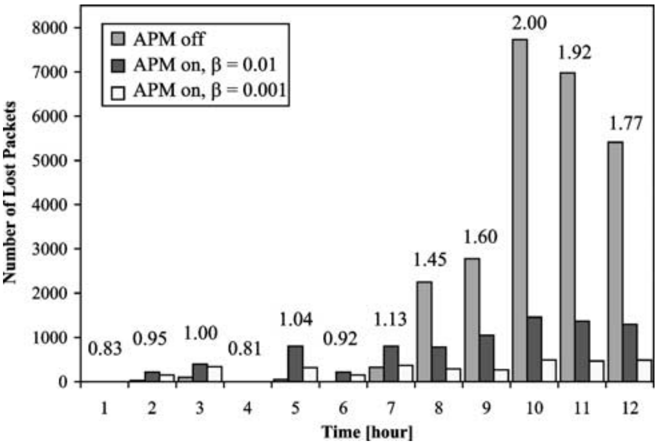
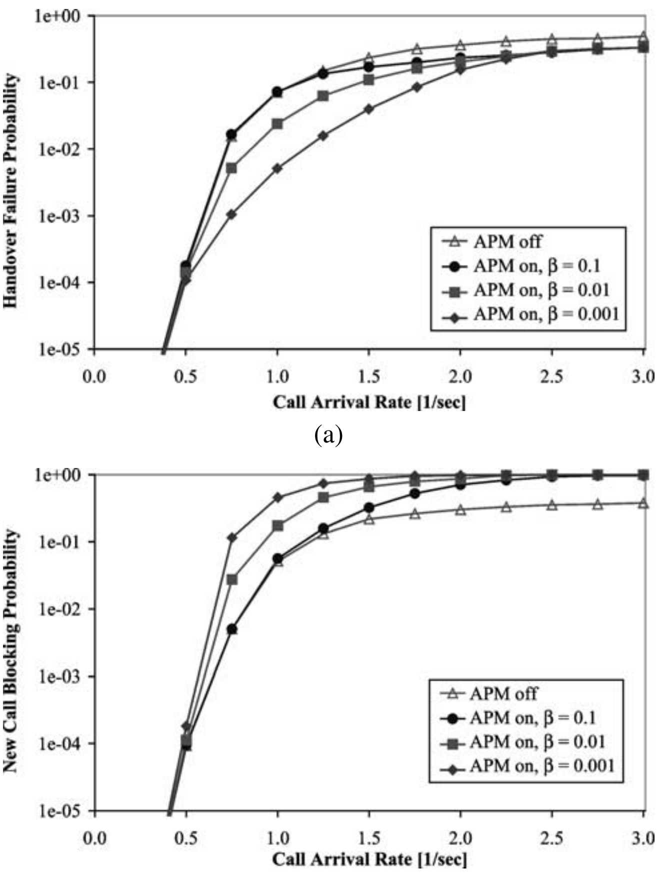


Figure 3. Number of packet losses for a half day window of a weekly usage pattern.

measured packet loss probability is greater than β. Therefore, the packet loss probability is in steady state also slightly greater than β. Nevertheless, figure 2 shows that the resulting packet loss probability can be adjusted quite well. For very low arrival rates, the packet loss probability is increased compared to the case without adaptive control. This is because the packet loss probability is below the desired level and η is adjusted towards 100%.

Figure 2(b) shows the average number of NRT users admitted in the cell. For all curves, the number of NRT users in the cell first increases up to about 70 users for an arrival rate of 1.0 arrivals per second. For higher arrival rates the admission controller decides to reject requests depending on the choice of the NRT threshold. In the case without APM the number of NRT users approaches 100 whereas in the cases with adaptive control less users are admitted in the cell because the threshold parameter η is decreased (e.g., about 80 users for $\beta = 0.001$). For high arrival rates a slightly decrease of the average number of NRT users can be observed. This is due to the fact that with increasing arrival rate the competition between voice, RT and NRT traffic decreases the bandwidth capacity available for NRT traffic. Therefore, less NRT users are admitted.

In the experiment presented in figure 3, we study the absolute number of packet losses observed in one hour for a transient scenario, i.e., the arrival rate of new calls is changing every hour according to a half day window of a weekly usage pattern [15]. The purpose of this experiment is to show that the adaptive performance management is fast enough to react on changing traffic conditions, i.e., to effectively adjust the NRT threshold in order to reduce packet losses. The bars shown in figure 3 correspond to the number of packet losses for experiments with and without adaptive control. Furthermore, the figure distinguishes between a desired loss level β of 0.01 and 0.001, respectively. The new call arrival rates considered in one hour are depicted above the bars. We conclude from figure 3 that for a real-life pattern of changing arrival rates the packet losses can be effectively controlled by the APM. This justifies the choice of the gradient $m = -0.02$ in the update function for the NRT threshold.



(a)

(b)

Figure 4. Impact of adaptive performance management on handover traffic.

Next, we study the effect of the APM on the handover traffic. Figure 4 shows the handover failure probability (a) and the new call blocking probability (b) for the UMTS system with and without APM. Similar to figure 2, we distinguish between different desired levels β for the handover failure probability. The desired level for new call blocking is fixed to 0.1. Note, that for controlling the handover bandwidth the desired level β can be used only to adjust the degree of prioritization of handover failure over new call blocking. Distinct from the packet loss probability, it cannot be expected to keep the handover failure probability at a constant level for increasing traffic load. That is for two reasons: (1) the handover bandwidth is adjusted according to two QoS measures that have a contrary influence and (2) the increase of the handover bandwidth must be limited by a certain portion of the overall available bandwidth (see section 3.2). If this limit is reached handover failures occur more frequently for further increasing call arrival rate. These two effects can be observed in the curves of figure 4. Nevertheless, the handover failure probability is improved more than one order of magnitude for call arrival rates between 0.75 and 1.25 call requests per second and a desired loss level $\beta = 0.001$. When studying the blocking probability of new voice calls and RT sessions (see figure 4(b)), we surly observe a higher blocking probability of new calls in the case with adaptive control and high arrival rate. In

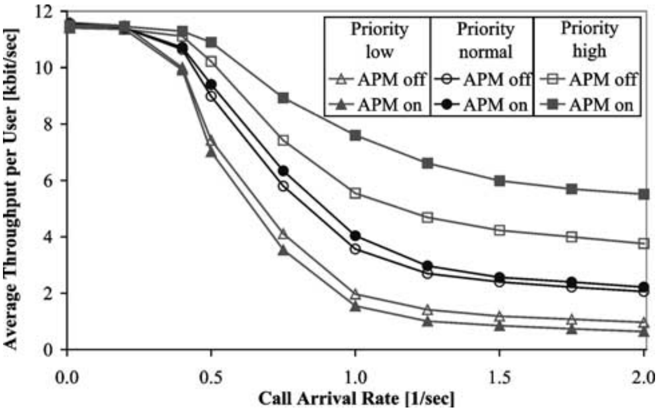


Figure 5. Improving QoS for high priority non real-time users.

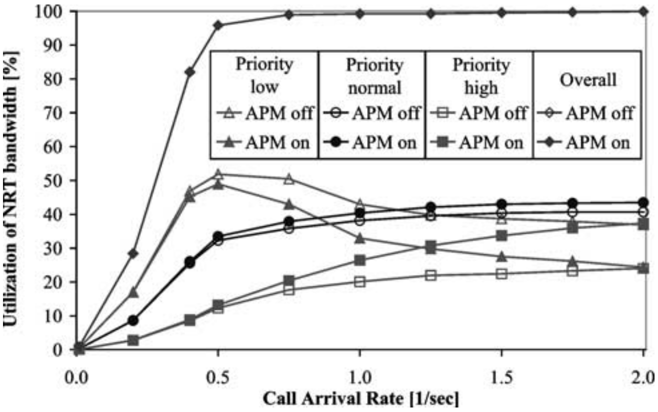
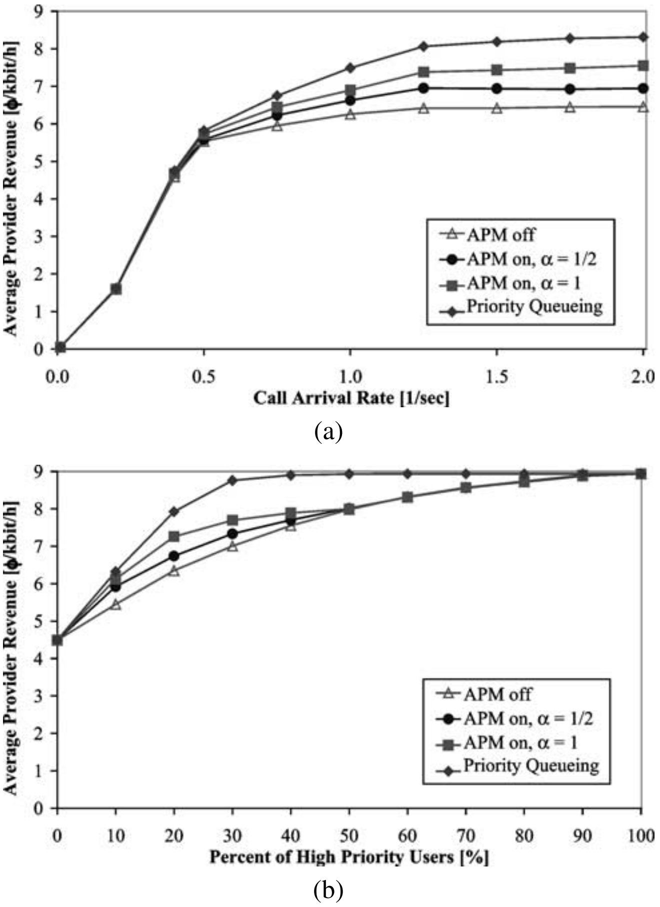


Figure 6. Effect of adjusting WFQ weights on bandwidth utilization of NRT traffic.



(a)

(b)

Figure 7. Revenue improvement for usage-based pricing policy.

fact, almost all call requests are blocked if system load is high.

In a next experiment, we study the impact of NRT users on QoS by the adaptive control of the queueing weights as introduced in section 4.2. Figure 5 plots the average throughput per user for each priority class of NRT traffic. As shown in table 1, we assume 10% NRT users with high priority, 30% with normal priority, and 60% with low priority. Recall, that higher priority service is more expensive and, hence, more users choose low priority service. If the overall load in the cell is very low (i.e., less than 0.3 call arrivals per second) each NRT user receives the maximal throughput independent of the priority class. However, when the cell load is further increased (arrival rates of more than 0.5 arrivals per second), throughput for users of all priority classes decreases. The intention of adaptively controlling the queueing weights is to reduce heavy throughput degradation of high priority users in this case. The performance increase of high priority users and the decrease of low priority users are shown in figure 5. Figure 6 plots the bandwidth portion utilized for each priority class of NRT traffic. For low arrival rate (i.e., less than 0.5 call arrivals per second) NRT users with low priority utilize the greatest portion of the NRT bandwidth because most NRT users have priority low. When the cell load is increased (arrival rates of more than 0.5 arrivals per second), the bandwidth will be utilized more and more by high priority users. The adaptive control of the WFQ weights decides to intensify this effect because users belonging to priority high suffer from the high population of low priority users. Figures 5 and 6 are derived from simulation runs with $\alpha = 1/2$ (see section 4.2).

In the following experiments, we study the impact of controlling the queueing weights on the revenue function (see equation (8)) for the three proposed pricing policies, i.e., usage-based, usage-/throughput-based, and the hybrid pricing policy. From the revenue function the average (steady state) provider revenue ϕ in the considered cell can be derived. Recall that the available bandwidth for NRT traffic is variable for different call arrival rates. Therefore, we consider the revenue earned by the provider in one hour per available bandwidth unit, i.e., per available kbit, for NRT traffic. Figure 7 shows the provider revenue for the usage-based pricing policy (i.e., $\gamma = 0$) and different values of the exponent α. As discussed in section 4.2, the best revenue improvement will be achieved with priority queueing. From the curves we conclude that the update strategy increases the revenue in one cell successfully for the considered traffic assumptions. Recall that the revenue improvement stems from a shift in bandwidth utilization towards higher priority users (see figure 6) if the population of high priority users is low compared to users of lower priority.

Figure 7(b) shows the revenue improvement for different user populations. In the experiment the percentage of high

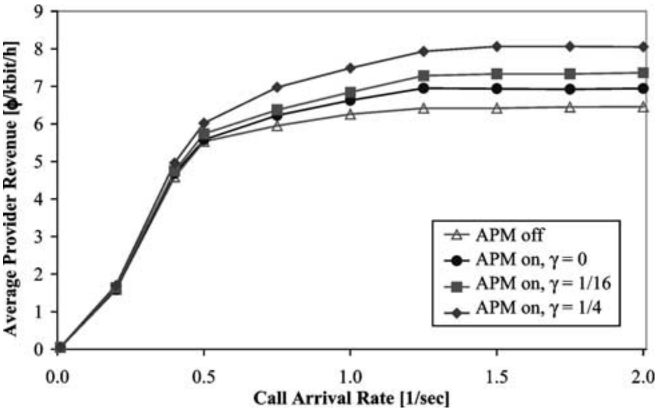


Figure 8. Revenue improvement for usage-/throughput-based pricing policy.

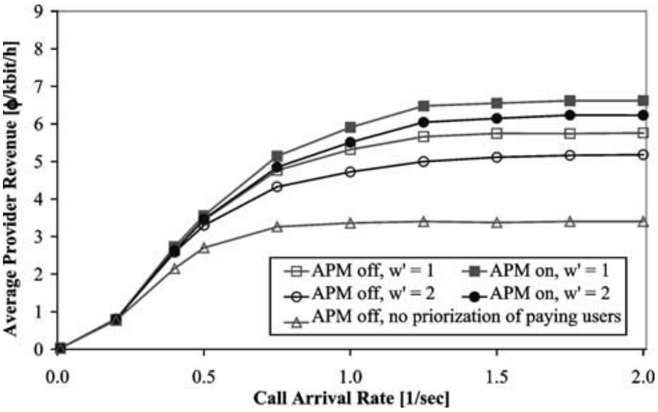


Figure 9. Revenue improvement for hybrid pricing policy.

priority users among the arriving user requests is varied. The remaining users are assumed to be low priority users. Normal priority users are not considered in this experiment (i.e., 0% normal priority users). This figure shows how the adaptive control of the queueing weights works. As expected, for a low percentage of high priority users the corresponding weight is increased. Therefore, QoS for high priority users and the provider revenue is also increased. For more than 50% high priority users the revenue is the same as in the case without adaptive control. No further revenue improvement is allowed because degradation of QoS for low priority users would be unacceptable. Considering a weak relation among the weights as introduced in section 4.2 would decrease the revenue compared to the case without adaptive control for more than 50% high priority users. This might be useful to increase QoS for users of low population independent of their priority class.

Figure 8 shows the revenue improvement for the usage-/throughput-based pricing policy and scaling exponents $\gamma = 1/4$ and $\gamma = 1/16$. In the last experiment we studied the revenue improvement for the hybrid pricing policy (see figure 9). We assume that half of the arriving users start their session in non-paying mode (i.e., $k = 2$). The curves distinguish between weights $w' = 1$ and $w' = 2$ for the non-paying users. Furthermore, the revenue for the case with and without adaptive control is compared. The curves are derived from simulations with $\gamma = 0$ and $\alpha = 1/2$. From the revenue curves of figures 7–9 the average monthly revenue can be computed considering a daily/weekly usage-pattern and different splits of call arrival rates of users requesting different services (i.e., voice, RT, NRT with different priorities). Comparing the monthly revenue for the pricing policies used in figures 7–9 with the monthly revenue for the hybrid pricing policy a provider can determine values such as the monthly free data volume and monthly payment per user.

6. Conclusions

We introduced a unified approach based on a mathematical framework for the adaptive performance management of 3G mobile networks. Opposed to previous work [8,13,19,21,25], the improvement of quality of service (QoS) and the optimization of mobile service provider revenue was considered in an integrated way. The unified approach aims at improving both QoS for mobile subscribers and increasing revenue earned by service providers. System parameters controlled by adaptive performance management constitute the portion of bandwidth reserved for handovers, the buffer threshold of the queue for non real-time traffic, and the weights of a weighted fair queueing packet scheduler.

Using the UMTS traffic model of [15] and a simulator on the IP level for the UMTS system, we presented performance curves for various QoS measures to illustrate the benefit of the unified approach for adaptive performance management. We introduced update functions that effectively control the packet loss probability and the handover failure probability. Considering usage-based, usage-/throughput-based, and hybrid pricing policies, we showed that the provider revenue in one cell can be significantly increased by the adaptive control of the queueing weights.

Throughout the paper, we considered the services and QoS profiles standardized for UMTS. Thus, the proposed approach for adaptive control is tailored to UMTS networks. However, by considering other services and QoS profiles, the basic ideas underlying the unified approach for adaptive performance management can also be applied for the adaptive control of other kinds of multi-service IP networks.

References

[1] 3GPP, `http://www.3gpp.org`

[2] 3GPP, QoS concept and architecture, Technical Specification TS 23.107 (September 2001).

[3] 3GPP, UTRAN overall description, Technical Specification TS 25.401 (September 2001).

[4] M. Ajmone Marsan, S. Marano, C. Mastroianni and M. Meo, Performance analysis of cellular mobile communication networks supporting multimedia services, Mobile Networks and Applications 5 (2000) 167–177.

[5] K. Aretz, M. Haardt, W. Konhäuser and W. Mohr, The future of wireless communications beyond the third generation, Computer Networks 37 (2001) 83–92.

[6] F. Barceló and J. Jordán, Channel holding time distribution in public cellular telephony, in: *Proceedings of the 16th International Teletraffic Congress*, Edinburgh, Scotland (1999) pp. 107–116.

[7] CSIM18 – The Simulation Engine, `http://www.mesquite.com`

[8] S.K. Das, R. Jayaram, N.K. Kakani and S.K. Sen, A call admission and control scheme for Quality-of-Service provisioning in next generation wireless networks, Wireless Networks 6 (2000) 17–30.

[9] A. Demers, S. Keshav and S. Shenker, Analysis and simulation of a fair queueing algorithm, in: *Proceedings of the International Symposium on Communications Architectures and Protocols (SIGCOMM)*, Austin, TX (1989) pp. 1–12.

[10] S. Floyd and V. Jacobson, Link-sharing and resource management models for packet networks, IEEE/ACM Transactions on Networking 3 (1995) 365–386.

[11] X. Geng and A.B. Whinston, Profiting from value-added wireless services, IEEE Computer 34 (August 2001) 87–89.

[12] A. Gupta, D.O. Stahl and A.B. Whinston, The economics of network management, Communications of the ACM 42 (1999) 57–63.

[13] A. Gupta, D.O. Stahl and A.B. Whinston, Priority pricing of integrated services networks, in: *Internet Economics*, eds. L. McKnight and J. Bailey (MIT Press, 1995) pp. 323–378.

[14] G. Haring, R. Marie and K.S. Trivedi, Loss formulas and their application to optimization for cellular networks, IEEE Transactions on Vehicular Technology 50 (2001) 664–673.

[15] A. Klemm, C. Lindemann and M. Lohmann, Traffic modeling and characterization for UMTS networks, in: *Proceedings of GLOBECOM 2001*, San Antonio, TX (November 2001) pp. 1741–1746.

[16] A. Klemm, C. Lindemann and M. Lohmann, Traffic modeling of IP networks using the batch Markovian arrival process, in: *Proceedings of Tools 2002*, London, Great Britain (April 2002) pp. 92–110.

[17] J. Kilpi and I. Norros, Call level traffic analysis of a large ISP, in: *Proceedings of the 13th ITC Specialist Seminar on Measurement and Modeling of IP Traffic*, Monterey, CA (2000) pp. 6.1–6.9.

[18] M. Krunz and A. Makowski, A source model for VBR video traffic based on $M/G/\infty$ input processes, in: *Proceedings of the 17th Conference on Computer Communications (IEEE INFOCOM)*, San Francisco, CA (1998) pp. 1441–1449.

[19] C. Lindemann, M. Lohmann and A. Thümmler, Adaptive performance management for UMTS networks, Computer Networks 38 (2002) 477–496.

[20] R. Ludwig, A. Konrad and A.D. Joseph, Optimizing the end-to-end performance of reliable flows over wireless links, in: *Proceedings of the 5th Conference on Mobile Computing and Networking (ACM MobiCom)*, Seattle, WA (1999) pp. 113–119.

[21] J.K. MacKie-Mason and H.R. Varian, Pricing the Internet, in: *Public Access to the Internet*, eds. B. Kahin and J. Keller (MIT Press, 1995) pp. 269–314.

[22] M. Meyer, TCP performance over GPRS, in: *Proceedings of the First Wireless Communications and Networking Conference (IEEE WCNC)*, New Orleans, MS (1999) pp. 1248–1252.

[23] Mobile Wireless Internet Forum (MWIF), OpenRAN architecture in 3rd generation mobile systems, Technical report MTR-007 (September 2001) `http://www.mwif.org`

[24] J.M. Peha and A. Sutivong, Admission control algorithms for cellular systems, Wireless Networks 7 (2001) 117–125.

[25] S. Rao and E.R. Petersen, Optimal pricing of priority services, Operations Research 46 (1998) 46–56.

[26] UMTS-Forum, UMTS/IMT-2000 Spectrum, Report No. 6 (1999).

[27] H. Zhang, Service disciplines for guaranteed performance service in packet-switched networks, Proceedings of the IEEE 83 (1995) 1374–1396.

[28] Wireless World Research Forum (WWRF), `http://www.wireless-world-research.org`

Christoph Lindemann is an Associate Professor in the Department of Computer Science at the University of Dortmund and leads the Computer Systems and Performance Evaluation group. From 1994 to 1997, he was a Senior Research Scientist at the GMD Institute for Computer Architecture and Software Technology (GMD FIRST) in Berlin. In the summer 1993 and during the academic year 1994/1995, he was a Visiting Scientist at the IBM Almaden Research Center, San Jose, CA. Christoph Lindemann is a Senior Member of the IEEE. He is author of the monograph *Performance Modelling with Deterministic and Stochastic Petri Nets* (Wiley, 1998). Moreover, he co-authored the survey text *Performance Evaluation – Origins and Directions* (Springer-Verlag, 2000). He served on the program committees of various well-known international conferences. His current research interests include mobile computing, communication networks, Internet search technology, and performance evaluation.
E-mail: cl@cs.uni-dortmund.de
WWW: http://www4.cs.uni-dortmund.de/~Lindemann/

Marco Lohmann received the degree Diplom-Informatiker (M.S. in computer science) with honors from the University of Dortmund in March 2000. Presently, he is a Ph.D. student in the Computer Systems and Performance Evaluation group at the University of Dortmund. He is a student member of the IEEE and the ACM. His research interests include mobile computing, Internet search technology, and stochastic modeling.
E-mail: ml@ls4.cs.uni-dortmund.de

Axel Thümmler received the degree Diplom-Informatiker (M.S. in computer science) from the University of Dortmund in April 1998. Presently, he is a Ph.D. student in the Computer Systems and Performance Evaluation group at the University of Dortmund. His research interests include mobile computing, communication networks, and performance evaluation.
E-mail: at@ls4.cs.uni-dortmund.de