ELSEVIER

Computer Networks 38 (2002) 477–496

COMPUTER NETWORKS

www.elsevier.com/locate/comnet

Adaptive performance management for universal mobile telecommunications system networks

Christoph Lindemann*, Marco Lohmann, Axel Thümmler

Department of Computer Science, University of Dortmund, August-Schmidt-Strasse 12, 44227 Dortmund, Germany

Received 6 September 2001; accepted 28 September 2001

Responsible Editor: I.F. Akyildiz

Abstract

In this paper, we introduce a framework for the adaptive control of universal mobile telecommunications system (UMTS) networks in order to improve bandwidth utilization of the radio channels. The key contribution of the paper constitutes the introduction of a performance management information base for dynamically adjusting the packet scheduler and admission controller. Thus, the adaptive control framework closes the loop between network operation and network control. Furthermore, the adaptive control framework can effectively deal with the different time scales of packet scheduling and admission control. Moreover, we present a traffic model for non-real-time UMTS traffic based on measured trace data. The analysis and scaling process of the measured trace data with respect to different UMTS bandwidth classes constitutes the basic concept of this traffic characterization. Using this traffic model and simulation on the IP level, the gain of employing the adaptive control framework is illustrated by performance curves for various quality of service measures.

Keywords: Third generation wireless systems; Performance management; Quality of service provisioning; Discrete-event simulation

1. Introduction

Third generation (3G) mobile communication systems like the universal mobile telecommunications system (UMTS) are characterized by a migration from voice-only to integrated services networks with data rates up to 2 Mbps. Thus, applications such as e-mail, Web browsing, and corporate local network access, as well as video conferencing, e-commerce, and multimedia can be supported over wireless data channels. UMTS provides the technical foundation for integrating the currently separate worlds of mobile and fixed telecommunications services into a unified digital data environment. Recently, the global wireless industry has created a global partnership project, the 3rd generation partnership project (3GPP) [1], for standardization of UMTS. This standardization process of UMTS is still ongoing (see e.g., Ref. [16]).

The quality of service (QoS) concept and architecture for UMTS networks has been specified by the 3GPP in Ref. [2]. There, the terms of the UMTS management and control functions

* Corresponding author.
E-mail address: cl@cs.uni-dortmund.de (C. Lindemann).
URL: http://www4.cs.uni-dortmund.de/~Lindemann/.

1389-1286/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S1389-1286(01)00265-1

(e.g., admission controller and resource manager) are defined and their functionality is roughly outlined. However, a detailed technical understanding how network management should effectively be performed for UMTS networks is subject to current industrial and academic research. In a visionary paper, Schwartz posed the engineering challenges in network management and control occurring from the introduction of multimedia services for 3G wireless networks [24]. Due to the scarce and costly radio frequencies for 3G wireless networks, adaptive resource management constitutes an important design issue for such networks. Das et al. proposed a framework for QoS provisioning for multimedia services in 3G wireless access networks [12]. They developed an integrated framework by combining various approaches to call admission control, channel reservation, bandwidth degradation, and bandwidth compaction. Záruba et al. proposed an admission control framework for optimal call mix selection to maximize the revenue earned by the service provider [26]. Their framework includes admission and degradation of calls based on priorities.

Traffic models for wireless IP networks have been addressed in several recent papers. Krunz and Makowski utilized a $M/G/\infty$ process for real-time (RT) video traffic modeling [21] whereas Coutras et al. proposed a Markov-modulated fluid process that serves as RT traffic source [9]. In the UMTS standard [15] recommendations for traffic models are given, which include parameterized distributions for RT and non-real-time (NRT) services, but detailed characteristics are given for WWW traffic only. However, the NRT traffic models are not derived from real measurements, which motivates a characterization of future NRT UMTS traffic based on measurements in network environments comprising of comparable characteristics.

In this paper, we introduce a framework for the adaptive control of UMTS networks. The basic idea lies in dynamically adjusting the packet scheduler and admission controller by means of a performance management information base (P-MIB). The adaptive control framework comprises of two major innovations not considered in previous work [12,14,17,22]. That is: (1) effectively dealing with the different time scales of packet scheduling and admission control; (2) closing the loop between network operation and network control. Thus, the adaptive control framework tackles a major challenge for the effective control of wireless IP networks, i.e., dealing with multiple control loops which operate on different time scales [10]. Packet scheduling operates on a fine time scale in the order of milliseconds whereas admission control operates on a coarse time scale in the order of seconds to minutes. System parameters to be controlled during network operation comprise of queueing weights for packet scheduling, a threshold value of the access queue for admission of elastic traffic (i.e., NRT traffic), and watermarks specifying bandwidth portions of the overall available bandwidth for data, voice, and handover traffic of the UMTS air interface. We introduce a P-MIB to improve overall QoS for different system configurations. The P-MIB contains for each feasible network configuration the optimal setting for the system parameters (i.e., queueing weights, threshold value of the access queue, and the watermarks). Online monitoring of system performance allows an adaptive performance management by quickly reacting to changes in the system performance parameters.

Based on measurements conducted at the Internet service provider (ISP) dial-in modem/ISDN link of the University of Dortmund, we present a NRT traffic model for UMTS networks applying the idea of the single user traffic model [20]. The key insight of this modeling approach lies in an scaling procedure of the measured trace data towards UMTS bandwidth requirements. A simulator on the IP level for the air interface of UMTS has been implemented using the simulation library CSIM [11]. Performance curves derived by simulation evidently illustrate the gain resulting from employing the adaptive control framework proposed in this paper. In fact, for the considered cellular UMTS network, simulation results show that QoS measures such as handover failure probability and packet loss probability are improved significantly.

The remainder of the paper is organized as follows. Section 2 describes the framework for adaptive performance management comprising of an admission controller, packet scheduler, and the P-MIB. Section 3 provides a comprehensive characterization of RT and NRT traffic in 3G mobile networks based on the idea of a single user traffic model. In Section 4, we present simulation results to show the benefit of employing the proposed framework for adaptive performance management. Finally, concluding remarks are given.

2. The framework for adaptive performance management

2.1. Universal mobile telecommunications system network architecture and quality of service classes

The UMTS network architecture standardized by the 3GPP distinguishes the UMTS access network, i.e., the UMTS terrestrial radio access network (UTRAN), and the UMTS core network [3]. The UTRAN consists of a set of radio network controllers (RNC) that are connected to the core network through the logical interface *Iu*. The core network comprises of the same supporting nodes as already introduced in GSM Phase 2+ (e.g., Mobile Switching Center or general packet radio service (GPRS) support nodes). They are responsible for handling circuit switched connections and tunneling packet switched data to public networks (e.g., the Internet). A RNC is connected through the *Iub* interface to a set of *Node B* elements, each of which can serve one or several cells (depending on omni-directional or sector cells). The RNC is responsible for control of the connected Node B elements, i.e., transceiver stations, and the radio link *Uu* to the mobile stations (MS). Inside the UTRAN, the RNC can be interconnected together through the *Iur* interface to support smooth handover for MS leaving the area covered by the serving RNC and entering the area of a drift RNC. The UTRAN architecture is presented in Fig. 1.

The QoS architecture specified for UMTS by ETSI 3GPP [2] distinguishes between four QoS classes: *conversational class*, *streaming class*, *interactive class*, and *background class*. The main distinguishing factor between these QoS classes lies in the delay sensitivity of the traffic. The conversational class is meant for traffic which is very delay

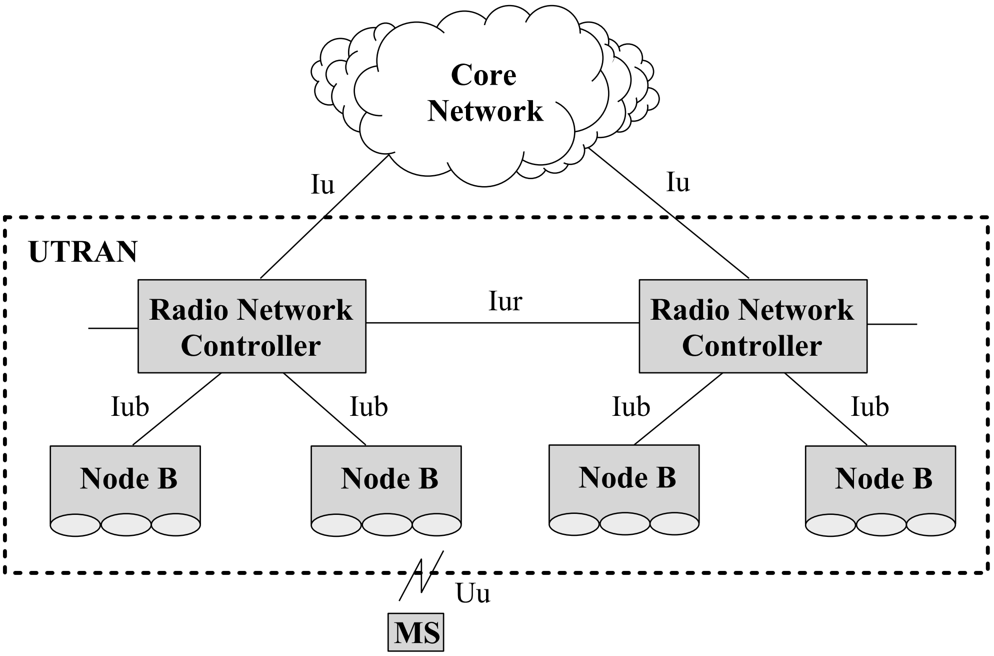


Fig. 1. Architecture of UTRAN.

sensitive while the background class is the most delay insensitive traffic class. Conversational and streaming classes are mainly intended to be used to carry RT traffic flows. A conversational RT traffic stream is characterized by requiring low transfer delay and small delay jitter because of the conversational nature of the stream. The maximum transfer delay is given by the human perception of video and audio conversation (e.g., video conferencing or voice over IP). The streaming traffic class consists of one-way RT traffic streams, e.g., viewing video clips or audio clips. The limit for acceptable transfer delay and delay jitter is not as stringent as in the conversational class. Delay jitter can be reduced by a time alignment function at the receiver that buffers data packets temporarily.

Interactive class and background class are mainly meant to be used by traditional Internet applications like WWW, e-mail, and FTP. The main difference between the interactive and the background class is that interactive class is mainly used by applications as e.g., interactive Web browsing, while background class is meant for e.g., background download of e-mails or background file downloading. Traffic in the interactive class has higher priority than background class traffic. Thus, background applications use transmission resources only when interactive applications do not need them. This is very important in a wireless environment where considerably less bandwidth capacity than in core networks is available.

2.2. Admission control based on adaptive bandwidth partitioning

The proposed framework distinguishes three different types of services: circuit-switched services, and packet-switched RT and NRT services. In general, circuit-switched services are voice calls from a common GSM mobile station. RT services correspond to the UMTS conversational and streaming class and NRT services to the UMTS interactive and background class [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 starts, the mobile 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 queue 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. Admission control of RT sessions is based on a QoS profile that specifies a *guaranteed bitrate* that should be provided to the application to work proper. If the desired bandwidth requirements cannot be satisfied by the network the corresponding admission request is rejected. NRT sessions will be admitted by concerning the current network state, i.e., queue length as specified later.

In the following, we propose a partitioning of the available bandwidth in one cell to meet the QoS requirements of the three considered service classes: voice calls, RT sessions, and NRT sessions. The bandwidth partitioning constitutes the rationale behind the admission control. Fig. 2 illustrates the partitioning of the available bandwidth into different areas. Let B be the overall bandwidth available in one cell. A portion b_{h} of

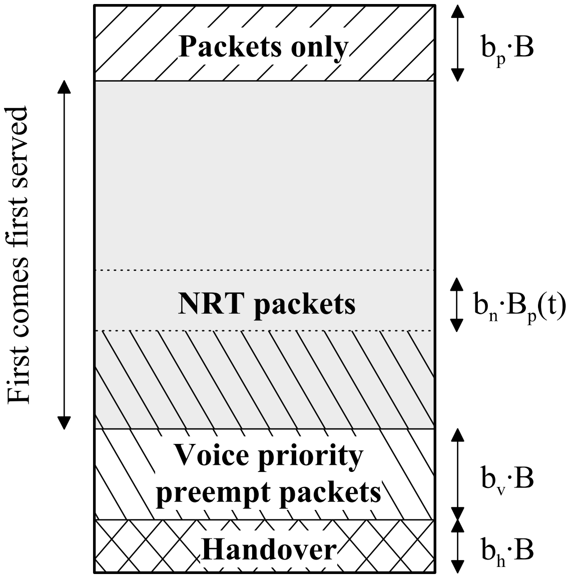


Fig. 2. Dynamic partition of the available bandwidth.

the bandwidth B is exclusively reserved for *handover* calls from neighboring cells in order to reduce handover failures. A portion b_{p} is reserved for RT and NRT data packets, i.e., *packets only*. The remaining bandwidth $(1 - b_{\mathrm{h}} - b_{\mathrm{p}})B$ can be allocated "on demand" by voice calls and data sessions, respectively.

Because in future UMTS networks (around the year 2010) voice calls will still play a major role in bandwidth requirements [25], we introduce a portion b_{v} of the overall bandwidth with priority of voice calls, i.e., *voice priority preempt packets*. This means, if more bandwidth than $(1 - b_{\mathrm{h}} - b_{\mathrm{v}})B$ is allocated for packet data calls and the remaining bandwidth is not sufficient for accommodating a newly arriving voice call, RT sessions have to be degraded below their guaranteed bandwidth. In fact in this case the bitrate guarantee given to RT users will be violated but as can be seen by the experiments presented in Section 4 the probability of RT session degradation is negligible for small values b_{v}. The remaining on demand bandwidth is allocated on a first-comes first-served (FCFS) basis to voice calls or RT sessions. In order to give NRT traffic a certain amount of bandwidth, a portion b_{n} of the bandwidth actually available for packet data is exclusively reserved for non-real-time packets: *NRT packets*. Let $B_{\mathrm{v}}(t)$ be the bandwidth reserved for all voice calls at a certain time point t, then for packet data bandwidth of size $B_{\mathrm{p}}(t) := B - B_{\mathrm{v}}(t)$ is available.

Table 1 shows the different call types arriving in the cell and the corresponding conditions under which these call arrivals will be admitted in a compact notation. Let t be the point in time when an admission request arrives at the admission controller. For RT users, admission is based on the availability of the guaranteed bandwidth specified in the QoS profile. Let $B_{\mathrm{r}}(t)$ be the bandwidth already allocated for RT traffic at time t and let B_{r} be the bandwidth required by the user who requested admission. The user will be admitted according to the bandwidth partitioning illustrated in Fig. 2. That is, if after call admission the handover bandwidth is still available, i.e., $B_{\mathrm{r}} + B_{\mathrm{r}}(t) \leqslant (1 - b_{\mathrm{h}})B - B_{\mathrm{v}}(t)$, and a portion b_{n} of the overall packet bandwidth $B_{\mathrm{p}}(t)$ is also still available for NRT sessions, i.e., $B_{\mathrm{r}} + B_{\mathrm{r}}(t) \leqslant (1 - b_{\mathrm{n}})(B - B_{\mathrm{v}}(t))$. These two cases are combined in the corresponding formula of Table 1 by a minimum operator. New voice calls with bandwidth requirements B_{v} will be admitted, if either less than $b_{\mathrm{v}}B$ bandwidth (voice priority preempt packets area of Fig. 2) is allocated for voice calls or if the voice call can be accommodated in the FCFS area without violating bandwidth requirements of ongoing calls. The corresponding formula presented in Table 1 can be derived in a similar way as for RT sessions.

Data packets arriving at the radio network controller are queued in two distinct queues until they are scheduled to be transmitted over the radio link. We distinguish a real-time queue (RTQ) with capacity K_{RT} and a non-real-time queue (NRTQ) with capacity K_{NRT}. For NRT sessions, the admission is based on the availability of buffer space in the NRTQ. This criteria may be set against certain buffer availability threshold of the capacity K_{NRT}, denoted by η, in order to prevent buffer overflow once the call is admitted. 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 admission controller does not prioritize NRT handovers over new NRT sessions.

Table 1
Call admission conditions for different call types

	Conditions for admission
Voice	$B_{\mathrm{v}} + B_{\mathrm{v}}(t) \leqslant \max\left\{b_{\mathrm{v}}B, \min\left\{(1 - b_{\mathrm{h}})B - \max\left\{b_{\mathrm{p}}B, B_{\mathrm{r}}(t)\right\}, B - \frac{1}{1-b_{\mathrm{n}}}B_{\mathrm{r}}(t)\right\}\right\}$
RT	$B_{\mathrm{r}} + B_{\mathrm{r}}(t) \leqslant \min\{(1 - b_{\mathrm{h}})B - B_{\mathrm{v}}(t), (1 - b_{\mathrm{n}})(B - B_{\mathrm{v}}(t))\}$
NRT	NRT queue length at admission request time $\leqslant \eta K_{\mathrm{NRT}}$
Voice handover	$B_{\mathrm{v}} + B_{\mathrm{v}}(t) \leqslant \max\left\{b_{\mathrm{v}}B, (1 - b_{\mathrm{h}})B - \max\left\{b_{\mathrm{p}}B, \frac{1}{1-b_{\mathrm{n}}}B_{\mathrm{r}}(t)\right\}\right\}$
RT handover	$B_{\mathrm{r}} + B_{\mathrm{r}}(t)(1 - b_{\mathrm{n}})(B - B_{\mathrm{v}}(t))$

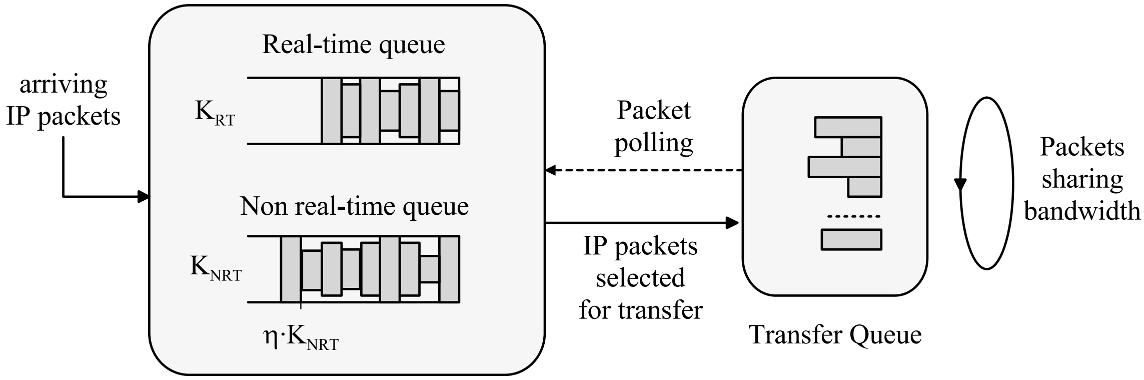


Fig. 3. RTQ, NRTQ, and transfer queue.

2.3. The packet scheduler

At a radio network controller responsible for a cell cluster, data packets from various connections arrive and are queued until bandwidth for transmission is available. The service discipline at this base station controls the order in which packets are served and how packets in transfer share the available bandwidth. Various service disciplines for packet scheduling of guaranteed service and best-effort connections have been summarized in Ref. [27]. Fig. 3 illustrates the proposed queues for data packets that are implemented in the radio network controller. Recall that data packets that arrive at the radio network controller are logically organized in two distinct queues. In each Node B element corresponding to one cell, a transfer queue is implemented that contains the packets actually in transfer. The available bandwidth capacity is shared over the packets in the queue according to their QoS requirements and bandwidth partitioning. Each time one of the following events occurs the available bandwidth is newly assigned to packets in transfer by a packet scheduler:

(i) admission of new voice call or handover,
(ii) termination of a voice call due to call termination or handover,
(iii) transfer of a RT or NRT data packet is finished,
(iv) arrival of a RT packet of user *i* at the radio network controller with no RT packet of user *i* waiting in the RTQ or been in transfer,
(v) arrival of a NRT packet at the radio network controller with no NRT packets waiting in the NRTQ and still bandwidth capacity available for NRT packets.

In order to distinguish different priorities for NRT traffic corresponding to the traffic handling priority defined by 3GPP [2], a weighted Round Robin scheduler or more complex scheduling strategies like weighted fair queueing (WFQ) [13] or WF^{2}Q [7] have to be implemented. An overview of queueing issues in wireless networks can be found in Ref. [8]. Whenever the packet scheduling is initiated due to the occurrence of one of the events (i) to (v) the available bandwidth is newly assigned to packets in transfer. The allocation of bandwidth is performed by the algorithm presented in Fig. 4. Let t be the point in time when the packet scheduling is initiated. Voice calls are assumed to be circuit switched. Therefore, each voice call allocates a fixed amount of bandwidth during its lifetime (see step (1)). In steps (3)–(9) of Fig. 4 the bandwidth requirements $B_{\mathrm{r}}(t)$ needed to satisfy the guaranteed bitrate for RT users is computed. If a portion b_{n} of the remaining packet bandwidth is not anymore available for NRT packets then RT sessions have to be degraded (see steps (10)–(12) in Fig. 4). Recall that this can only happen if bandwidth requirements for RT packets are so exhaustive that they occupy the voice priority preempt packets areas (see Fig. 2) and an additional voice call is admitted in the cell. The de-

```
(1)  calculate new voice bandwidth $B_v(t)$ and share $B_v(t)$ among all active voice calls
(2)  $B_r(t) = 0$
(3)  FOR ALL active real-time users i DO
(4)    IF packet of user i in transfer THEN DO
(5)      increase $B_r(t)$ by guaranteed bitrate of user i
(6)    ELSE IF real-time queue of user i not empty THEN
(7)      poll packet of user i in transfer queue and increase $B_r(t)$ by guar. bitrate of user i
(8)    OD
(9)  OD
(10) WHILE $B_r(t) > (1 - b_n) \cdot B_p(t)$ DO
(11)   degrade real-time sessions stepwise
(12) OD
(13) share remaining bandwidth according to a common scheduling discipline among
     active NRT users
(14) WHILE degraded real-time sessions exist AND still free bandwidth available DO
(15)   increase bandwidth of degraded real-time session
(16) OD
```

Fig. 4. Scheduling discipline assigning bandwidth to different users.

gradation of RT sessions if performed stepwise. That is, in each degradation step all RT sessions are degraded to a specified level before starting the next degradation step if necessary. After assigning bandwidth to RT sessions the remaining bandwidth is allocated to NRT traffic (see step (13)). If still bandwidth available then degraded RT sessions can be increased to their guaranteed bandwidth again (see steps (14)–(16)).

2.4. The performance management information base

This section introduces the framework for adaptively adjusting system parameters at a radio network controller (i.e., a base station responsible for a cluster of cells). This adaptive framework constitutes the main contribution of the paper. To maximize QoS for the mobile users, a performance management entity has to be introduced in a radio network controller that is responsible for corresponding transceiver stations (i.e., Node B elements). Furthermore, a radio network controller has to be extended by an online performance measurement component that derives performance measures in a certain time window (e.g., handover failure probabilities of mobile users). These performance measures form a *system pattern*. The system pattern is submitted in fixed time intervals to the performance management entity which subsequently updates the system parameters (i.e., parameters of traffic controlling components like the admission controller and packet scheduler). The update of system parameters is made as specified in a P-MIB. Fig. 5 shows a detailed view of the proposed adaptive framework for performance management that should be hosted in a radio network controller. System parameters which are adjusted by the performance management entity comprise of

- bandwidth portions b_h, b_p, b_v, b_n,
- NRTQ threshold (portion η of buffer size),
- queueing weights w_i for NRT packets with priority i.

Thus, the proposed framework takes into consideration the different time scales of network control and operation. The parameters b_h, b_p, b_v, and b_n corresponding to bandwidth partitioning and the threshold η are parameters that effect the system at the connection level, i.e., admission control, which operates on a coarse time scale in the order of seconds to minutes. The queueing weights effect the system at packet-level operating on a fine time scale in the order of milliseconds [22,27]. That is they control packet loss probabilities of connections with priority 1 (high), 2 (normal), and 3 (low). Thus, adaptively changing these

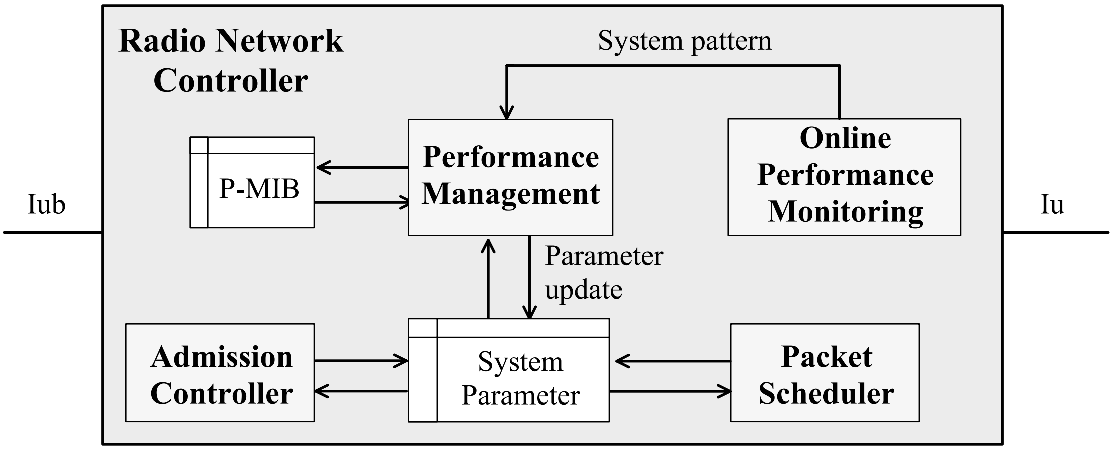


Fig. 5. System architecture for adaptive performance management.

parameters improves the system control at two different time scales.

The UMTS system parameters can effectively be updated by monitoring QoS measures which immediately affect these parameters. For ease of notation, we abbreviate the QoS measures handover failure probability and call/session blocking probability corresponding to voice calls and RT sessions by HFP and CBP, respectively. The packet loss probability of the NRTQ is abbreviated by PLP. The average number of active NRT sessions with priority 1, 2, and 3 is denoted by NRT_1, NRT_2, and NRT_3. Determining the updates for the system parameters, i.e., determining $b_{\mathrm{h}}^{(\mathrm{new})}$ and $\eta^{(\mathrm{new})}$, and the updated queueing weights $w_1^{(\mathrm{new})}$, $w_2^{(\mathrm{new})}$, and $w_3^{(\mathrm{new})}$ can be performed based on the dependencies (1)–(3) to the corresponding old values and the actually observed QoS measures HFP, CBP, PLP, NRT_1, NRT_2, NRT_3. That is:

1. $b_{\mathrm{h}}^{(\mathrm{old})}, \mathrm{HFP}, \mathrm{CBP} \rightarrow b_{\mathrm{h}}^{(\mathrm{new})}$
2. $\eta^{(\mathrm{old})}, \mathrm{PLP} \rightarrow \eta^{(\mathrm{new})}$
3. $\mathrm{NRT}_1, \mathrm{NRT}_2, \mathrm{NRT}_3 \rightarrow w_1^{(\mathrm{new})}, w_2^{(\mathrm{new})}, w_3^{(\mathrm{new})}$

The online monitoring of QoS measures is performed by a sliding window technique as depicted in Fig. 6. The width of the sliding window depends on the number of relevant events that occur according to a performance value (e.g., NRT packet arrivals are relevant events for com-

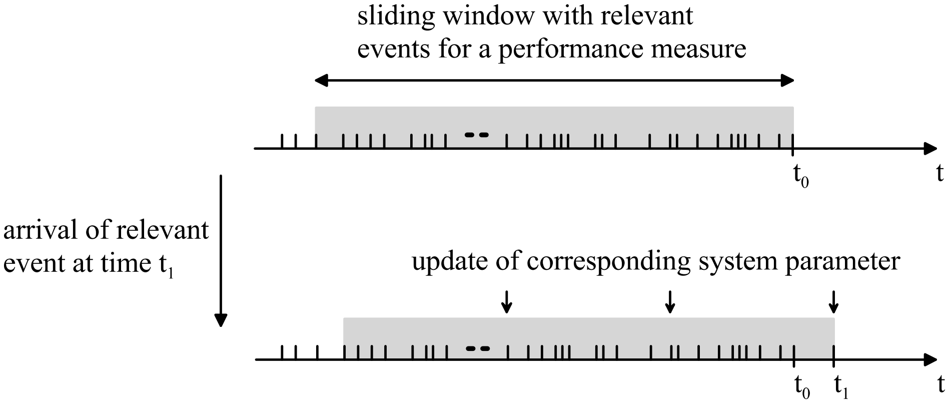

sliding window with relevant events for a performance measure
t_0
t
arrival of relevant event at time t_1
update of corresponding system parameter
t_0
t_1
t

Fig. 6. Online performance monitoring and system parameter update.

puting PLP). The upper part of Fig. 6 shows the sliding window at a certain time point t_0. Assuming that at time t_1 the next relevant event occurs the sliding window moves in time as shown in the lower part of Fig. 6. After a certain number of relevant events are occurred a system parameter update is performed based on the performance measure derived from the sliding window (e.g., update of η according to PLP derived from sliding window).

Note that an accurate online monitoring of performance values requires a sliding window that is not too small. A certain number of events representing the history of the performance value have to be considered to get an expressive performance measure. On the other hand considering a big sliding window prevents the performance management 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 measured performance value. We studied this tradeoff between accurate online monitoring and fast reaction of the P-MIB to changing traffic conditions in several experiments with varying call arrival rate to get the optimal width of the sliding window for each performance measure.

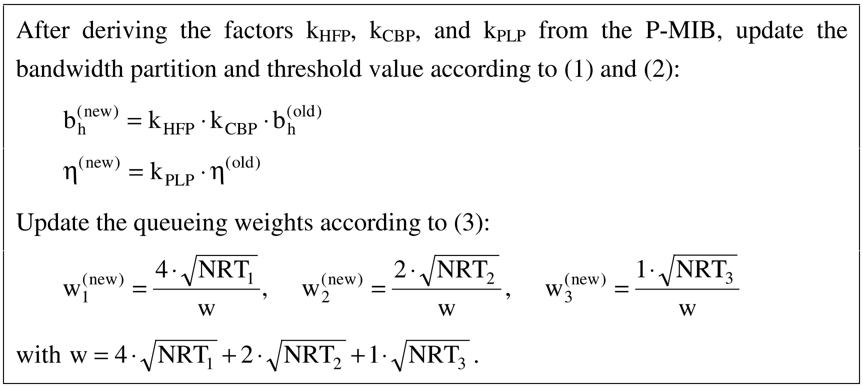
Table 2 shows an extract of the P-MIB controlling the system parameters b_h and η. The entries of the P-MIB are determined off-line using the simulator of the UMTS system. In each simulation experiment, one parameter is varied while all others are kept fixed in order to find the value of this parameter which improves the QoS measures. Thus, the entire parameter space is explored offline using simulation for deriving the entries of the P-MIB. The values derived in this way can be verified by some heuristics. For example, if the handover failure probability is high, the portion of the bandwidth reserved for handover, b_h, should be increased.

Table 2
P-MIB for system parameters b_h and η

	P-MIB for handover bandwidth b_h		P-MIB for NRT threshold η
	k_{HFP}	k_{CBP}	k_{PLP}
(0.5, 1]	1.110	0.936	0.960
(0.4, 0.5]	1.105	0.939	0.965
(0.3, 0.4]	1.100	0.943	0.970
⋮	⋮	⋮	⋮
(0.001, 0.005]	1.014	0.996	1.010
(0, 0.001]	0.980	1.020	1.020

Determining $b_h^{(new)}$ and $\eta^{(new)}$ is performed based on their previous settings as shown in Fig. 7. In fact, for $b_h^{(new)}$ the product of the factors k_{HFP}, and k_{CBP} corresponding to the actually observed HFP and CBP is taken into account. For $\eta^{(new)}$ the factor k_{PLP} representing PLP is taken into account. Note that a high HFP should increase $b_h^{(new)}$ but this obviously also increases the CBP because less bandwidth is available for new voice calls and RT sessions. Therefore, the factors presented in Table 2 are chosen such that if the probabilities HFP and CBP lie in the same interval, e.g., (0.3, 0.4], the product of corresponding factors k_{HFP} and k_{CBP} results in a value greater one. That is, the bandwidth reserved for handovers will be increased. Thus, handover calls are slightly prioritized over new voice calls or RT sessions if their blocking probability is equal. Note that the values $b_h^{(new)}$ and $\eta^{(new)}$ are truncated at a lower bound of 0.1% and an upper bound of $1 - b_v - b_p$ for $b_h^{(new)}$ and 1 for $\eta^{(new)}$, respectively. The truncation at the lower bound guaranties that values of b_h and η do not accumulate near zero for long periods of low traffic load and the truncation at the upper bound guarantees that the computation of $b_h^{(new)}$ and $\eta^{(new)}$ results in valid values.

The update of the queuing weights i.e., determining $w_1^{(new)}$, $w_2^{(new)}$, and $w_3^{(new)}$ is made according to the measured average number of NRT sessions belonging to priorities 1, 2, and 3 in the cell. Fig. 7 specifies how to update the queueing weights for a weighted Round Robin scheduling discipline with three classes corresponding to the three priorities. If the number of users in the cell of priorities 1, 2, and 3 is equal, the weights will be set to 4, 2, and 1. If the number of users is not the same for each priority class, the weights will be adjusted such that the traffic class with the highest user population is prioritized. That is for example, if the majority are low priority users, the weight for low priority should be increased. We use the square root of the measures NRT_1, NRT_2, and NRT_3 in order to smooth the described influence of the



After deriving the factors k_{HFP}, k_{CBP}, and k_{PLP} from the P-MIB, update the bandwidth partition and threshold value according to (1) and (2):

$$b_h^{(new)} = k_{HFP} \cdot k_{CBP} \cdot b_h^{(old)}$$

$$\eta^{(new)} = k_{PLP} \cdot \eta^{(old)}$$

Update the queueing weights according to (3):

$$w_1^{(new)} = \frac{4 \cdot \sqrt{NRT_1}}{w}, \quad w_2^{(new)} = \frac{2 \cdot \sqrt{NRT_2}}{w}, \quad w_3^{(new)} = \frac{1 \cdot \sqrt{NRT_3}}{w}$$

with $w = 4 \cdot \sqrt{NRT_1} + 2 \cdot \sqrt{NRT_2} + 1 \cdot \sqrt{NRT_3}$.

Fig. 7. Procedure for updating the system parameters using the P-MIB.

number of NRT users on the queueing weights. For other scheduling disciplines like WFQ and WF^2Q corresponding schemes can be derived in a similar way.

The overhead introduced by the adaptive performance management is insignificant because the online performance monitoring component and the performance management component both reside in the radio network controller. Therefore, no time consuming signaling is needed. Furthermore, no additional signaling is needed for information exchange among different cells for reserving handover bandwidth capacity. This is due to the implicit allocation of handover bandwidth by the performance management component based on the handover flow and not explicitly by RNC corresponding to neighboring cell clusters.

3. Traffic characterization for third generation wireless networks

3.1. The single user traffic model

We consider UMTS users active in a base station area. Each user generates packet data traffic by utilizing different applications, which can be partitioned in RT and NRT applications. We assume further that RT and NRT applications are mutually exclusive for a single UMTS user.

The traffic model for NRT applications utilizes the notion of the single user traffic model [20], where a user runs NRT applications, i.e., HTTP, Napster, e-mail, UDP, and FTP, according to a characteristic usage pattern. Moreover, a single user can run different applications that may be concurrently active, e.g., WWW browsing while downloading Napster music files. Each application is fully described by its statistical properties, which comprise of an alternating process of ON- and OFF-periods with some application specific length or data volume distribution, respectively. Within each ON-period the packet arrival process is completely captured by the packet interarrival times and the corresponding packet sizes. Thus, the single user traffic model characterizes the NRT traffic that an individual user generates and is employed on three different levels:

1. The session level describes the dial-in behavior of the individual users, characterized by the session interarrival-time distribution and the session data-volume distribution.
2. The connection level describes for each individual application the corresponding distribution of connection interarrival times and connection data volume, respectively.
3. The packet level characterizes the packet interarrival-time distribution and the packet size distribution within the application specific connections.

During an ON-period, i.e., an application specific connection, the user applies the appropriate application in an active fashion. The interarrival time between two successive connection starting points of the same application type and the data volume of each connection are drawn from general distributions, respectively. The packet interarrival times within each connection and the corresponding packet sizes are also drawn according to an application dependent distribution. Thus, the overall traffic stream of a user constitutes of the superposition of the packet arrival process of all application connections within the user's session.

3.2. Real-time and non-real-time traffic characteristics

For RT applications we utilized the approach proposed in Ref. [21], where VBR video traffic is modeled in terms of time-discrete $M/G/\infty$ 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 $M/G/\infty$ input process. Subsequently, the generated traffic is transformed utilizing a hybrid Gamma/Pareto numerical transformation in order to capture the marginal distribution of the measured traffic stream. For further details of the RT traffic model we refer to Ref. [21]. 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 4 for details).

Recent recommendations for NRT traffic models are proposed in Refs. [5,15]. As a drawback, these traffic models are not derived from real measurements, which motivates a characterization of future NRT UMTS traffic based on measurements in network environments comprising of comparable characteristics. In Ref. [20] Kilpi and Norros showed that IP traffic of current ISPs inhibits many characteristics of future UMTS traffic, i.e., different access speeds, influence of the user behavior due to different tariff limits, as well as asymmetric up- and downlink traffic. The main difference between the measured IP traffic at dial-in modem/ISDN link and future UMTS traffic constitutes of the different bandwidth classes of individual users. Thus, based on measurements conducted at the ISP dial-in modem/ISDN link of the University of Dortmund, we present a NRT traffic model for UMTS networks applying the idea of the single user traffic model described above. The key insight of this modeling approach lies in an appropriate scaling procedure of the measured trace data towards UMTS bandwidth requirements.

During the measurement over a four-week period in January 2001, approximately 110,000 user sessions with a total data volume of 120 GB have been logged. All measurements have been conducted at the Ethernet link between the dial-in routers and the router connection to the Internet. We used the *TCPdump* software package running on a Linux client for sniffing all IP packet headers sourced or targeted by dial-in users. For each IP packet the arrival timestamp, the source port, the target port, the packet length, and other TCP header information have been recorded. Moreover, we partitioned the dial-in users in the following bandwidth classes: 9.6, 14.4, 28.8, 33.6, 56, and 64 kbps (ISDN).

In Ref. [19], we observed that each statistical measure of the three traffic levels (session-, connection-, and packet-level) comprise of a *characteristic distribution* which is independent of the dial-in user's bandwidth class, e.g., the HTTP connection interarrival times are distributed according to a lognormal distribution. Therefore, the distribution of a specific statistical measure differs only by the parameter values of the characteristic distribution for different bandwidth classes. In order to find such a characteristic distribution for a specific statistical measure we use a least-squares regression with respect to the bandwidth classes 9.6, 14.4, 28.8, 33.6, 56, and 64 kbps. We utilized the following set of probability density functions (pdf) that can closely match the considered statistical measures: Lognormal, Pareto, Weibull, Gamma, and Exponential.

In the following, the maximum bandwidth capability of future UMTS handheld devices is denoted as bandwidth class. To obtain the traffic characteristics of the UMTS traffic model, we employ the scaling algorithm introduced in Ref. [19] (see Fig. 8). This scaling procedure utilizes the notion of (1) bandwidth-independent characteristic distributions for the statistical measures on the three traffic levels, and (2) bandwidth-dependent

Step 1: Find bandwidth-dependent trends in the mean and the variance of the considered statistical measures. Utilize the least-squares regression method on the mean and variance with respect to increasing bandwidth for each statistical measure. We use the following functions as underlying regression models.

(a) $f_1(x) = a + b \cdot \log^2(c \cdot x)$, a double logarithmic shape.

(b) $f_2(x) = a + b \cdot \log(c \cdot x)$, a logarithmic shape.

(c) $f_3(x) = a + b \cdot \log(c \cdot x) + d \cdot x$, a mixture of a logarithmic and linear shape.

(d) $f_4(x) = a + b \cdot x$, a linear shape.

Step 2: Get the parameterized function of Step 1, which comprises of the least squares residual value. Subsequently, use this parameterized function in order to derive values for mean and the variance corresponding to the UMTS bandwidth classes 64 kbps, 144 kbps, and 384 kbps.

Step 3: For each statistical measure, utilize its characteristic distribution and the mean and variance, calculated in Step 2, to get the parameter values of the characteristic distribution. This task can be performed by solving a non-linear equation system, comprising of the analytical formulas for corresponding mean and variance and the values for mean and variance derived in Step 2.

Fig. 8. Algorithm for bandwidth scaling for UMTS networks [19].

trends in the mean and the variance of each unique statistical measure. In the first step the identification of bandwidth-dependent trends for each statistical measure utilizing a regression method is conducted. The basic idea of the underlying regression models constitutes of the notion of bandwidth-dependent trends and the evolution of mean and/or variance, i.e., bandwidth-dependent changes that are naturally described by one of the functions (a)–(d). We consider this set of functions because they comprise of different asymptotic behavior, e.g., a linear or a logarithmic asymptotic behavior. Subsequently, we utilize the parameterized function, which comprises of the least-squares residual value, in order to get the mean and the variance values corresponding to the UMTS bandwidth classes 64, 144, and 384 kbps. For each statistical measure, we utilize its characteristic distribution and the derived values for mean and variance, in order to get the parameter values of the characteristic distribution (see Step 3 in Fig. 8). The following presents characteristic distributions and the corresponding parameter values for the statistical measures of the session-level, connection-level, and packet-level.

The reason for restricting the NRT traffic model to the bandwidth classes 64, 144, and 384 kbps is twofold. First, we think that after the commercial launching of UMTS most users will run applications on their hand-held devices using the cheaper and, thus, lower bandwidth classes. Second, the notion of trends in the statistical measures and the utilization of a scaling procedure by regression methods are based on measurements that comprises of bandwidth classes from 9.6 to 64 kbps. From this point of view, a trend spotting up to the maximum UMTS bandwidth class (i.e., 1920 kbps) involves too many unknowns.

The statistics at the session level are mainly influenced by user behavior, which is difficult to predict. As the session interarrival time depends to a large extend on the prizing policy of the UMTS

Table 3
Distribution of session volume

Distribution	64 kbps	144 kbps	384 kbps
Lognormal($\mu;\sigma^2$)	(11.1170; 1.9095)	(11.4107; 1.9509)	(11.6795; 1.9781)

provider, we assume an exponential distribution with rate λ for session interarrival times. Note, λ is determined by the call arrival rate (see Section 4). However, in order to take into account the bandwidth-dependent transfer data volumes, we get the characteristic distributions and corresponding parameter sets as shown in Table 3.

At connection level, the measured data indicates that almost all users, who utilize Napster or FTP applications, only run a single session. Therefore, an interarrival-time distribution for connections for those applications is misleading. Thus, we take into account the fractions for using Napster or FTP measured at the ISP. Furthermore, we assume that users run these applications in a single connection. These fractions are derived from measured data of ISDN users, i.e., 1.47% for Napster and 3.05% for FTP. We focus on this bandwidth class because users of lower bandwidth classes hardly utilize these applications. For the remaining applications, the interarrival-time distributions for connections and the data volume distributions per connection are presented in Table 4. Recall that UDP applications are connectionless. Thus, UDP applications are omitted in the statistics for the connection level. UDP applications are assumed to be active during the entire user session and are fully described by its packet interarrival-time distribution.

Table 5 presents the application dependent packet interarrival-time distributions. Note that the packet size distributions for HTTP, Napster, e-mail, and FTP follow to a large extend a discrete distribution, i.e., packets of the sizes 40, 576, and 1500 bytes constitute the largest amount of the overall packet sizes. This phenomenon relies on the maximum transfer units (MTU) of Ethernet and SLIP (serial line IP) networks. Most TCP transfer protocols like HTTP, FTP, and POP3 are used to transfer files as fast as possible. Therefore, within a connection the packets are filled up to the MTU of the underlying network protocol. This is usually 1500 bytes in Ethernet networks and 576 bytes in SLIP networks. Packets with a length of 40 bytes are at most TCP acknowledgments with

Table 4
Statistical properties at connection level

		Distribution	64 kbps	144 kbps	384 kbps
HTTP	Interarrival time	Lognormal($\mu;\sigma^2$)	(0.5967;2.6314)	(0.1580;3.1507)	(−0.4760;3.8787)
	Data volume	Lognormal($\mu;\sigma^2$)	(7.4343;3.4714)	(7.4708;3.7598)	(7.5458;3.9745)
E-mail	Interarrival time	Pareto(k;α)	(14.4360;2.1345)	(15.1334;2.1254)	(16.0229;2.1223)
	Data volume	Lognormal($\mu;\sigma^2$)	(8.1934;3.3852)	(8.2944;3.5288)	(8.4124;3.6439)
Napster	Interarrival time	Not available			
	Data volume	Lognormal($\mu;\sigma^2$)	(12.3025;1.5385)	(12.3677;1.5311)	(12.5410;1.5268)
FTP	Interarrival time	Not available			
	Data volume	Lognormal($\mu;\sigma^2$)	(8.4944;3.6674)	(8.6403;4.1059)	(8.8409;4.3343)

Table 5
Parameters of packet interarrival times

	Distribution	64 kbps	144 kbps	384 kbps
HTTP	Lognormal($\mu;\sigma^2$)	(−3.2441;4.5137)	(−3.9124;5.1794)	(−4.8507;6.1159)
E-mail	Lognormal($\mu;\sigma^2$)	(−4.4052;4.4970)	(−4.8790;4.9687)	(−5.4096;5.4978)
Napster	Lognormal($\mu;\sigma^2$)	(−4.2614;3.7790)	(−4.0340;3.3242)	(−4.4335;3.5226)
FTP	Lognormal($\mu;\sigma^2$)	(−3.6445;4.9564)	(−3.9076;5.2186)	(−4.1089;5.4194)
UDP	Lognormal($\mu;\sigma^2$)	(−3.2770;5.2887)	(−3.7830;5.6710)	(−4.3020;6.0997)

missing data field. Recall, that the TCP/IP header without any options consists of 40 bytes. Table 6 displays the fractions of these discrete packet sizes. We observe further, that the remaining packet sizes are distributed uniformly between 40 and 1500 bytes. In contrast to the TCP packets, the UDP datagram sizes follow a bandwidth-independent lognormal distribution with parameters $\mu = 1.8821$ and $\sigma^2 = 5.4139$.

4. Performance improvement due to the adaptive control framework

4.1. The simulation environment

The simulator considers a cell cluster comprising of seven hexagonal cells. We assume that a mobile user requests a new *session* in a cell according to a Poisson process with call arrival rate λ. 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. Recall that voice calls are assumed to be circuit-switched connections that require a constant amount of radio channels (i.e., a constant amount of bandwidth corresponding to 16 kbps uncoded). 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. Moreover, we assume that 25% of the sessions are voice calls whereas RT and NRT services constitute 15% and 60% of the overall sessions (see Table 7). The frequency spectrum covered by traditional GSM services (i.e., 890–915 and 935–960 MHz) can be also utilized by voice calls but is not considered in our study.

Subsequently, we have to specify the QoS profile for RT and NRT sessions. For RT sessions we define two QoS profiles, i.e., a low bandwidth profile comprising of a guaranteed bitrate of 64 kbps corresponding to streaming audio and a high bandwidth profile comprising of a guaranteed bitrate of 192 kbps corresponding to streaming video. According to the RT traffic model presented in Section 3, 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 presented in Section 3, 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 7), respectively.

Before a mobile user can start a new session, he/she has to pass the admission controller introduced in Section 2.2. The amount of time that a mobile user with an ongoing session remains within the cell is called *dwell time*. If the session is

Table 6
Fractions of different packet sizes in overall traffic

	Packet size: 40 byte (%)	Packet size: 576 byte (%)	Packet size: 1500 byte (%)	Other packet: sizes (%)
HTTP	46.77	27.96	8.10	17.17
Napster	34.98	45.54	4.18	15.30
E-mail	38.25	25.98	9.51	26.26
FTP	40.43	18.08	9.33	32.16

Table 7
Characteristics for different UMTS session types

	Circuit-switched voice service	Streaming RT		Interactive NRT		
		Audio	Video	High priority	Normal priority	Low priority
Portion of arriving requests (%)	25	12	3	6	18	36
Session duration (s)	120	180	180	Determined by session volume distribution		
Session dwell time (s)	60	120	120	120	120	120

still active after the dwell time, a handover toward an adjacent cell takes place. We assume the duration of voice calls and RT sessions to be exponentially distributed. The dwell time can be better modeled by a lognormal distribution as shown in Ref. [6]. Parameters of the distribution are given by $\mu = 1.7770$ and $\sigma^2 = 4.6347$ for voice calls and $\mu = 2.1260$ and $\sigma^2 = 5.3230$ for RT and NRT sessions, respectively. All corresponding mean values are shown in Table 7. As described in Section 3, a NRT session remains active until a specific data volume drawn according to a bandwidth-dependent lognormal distribution (see Table 3) is transferred, assuming it completes without being forced to terminate due to handover failure. Thus, in the simulation environment, a session of a mobile user is completely specified by the following parameters: service class (e.g., voice, RT, NRT), packet arrival process, dwell time, session duration (defined by the amount of time or data), and QoS profile. For voice calls, the QoS profile is empty, since a fixed portion of bandwidth is reserved for a voice call during its dwell time in the cell. The QoS profile for RT sessions consists of a guaranteed bitrate as specified in Section 2.2. To distinguish between NRT traffic, the UMTS simulator implements three packet priorities: high, normal, and low. 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. It is impossible to specify in advance the intensity of the incoming handover flow. This is due to the fact that the handover rate out of the cell depends on the number of active customers within the cell. On the other hand, the handover rate into the cell depends on the number of customers in the neighboring cells. Thus, the iterative procedure introduced in Ref. [4] is employed for balancing the incoming and outgoing handover rates. The iteration is based on the assumption that the incoming handover rate $\lambda_{h,i}^{(n+1)}$ of application i at step $n+1$ is equal to the corresponding outgoing handover rate computed at step n.

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 protocol. As shown in Ref. [23] for the 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. Therefore, a reliable link level can be assumed when considering higher layer protocol actions. To accurately model the UMTS radio access network, the simulator represents the functionality of one radio network controller and seven omni-directional 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 $\rightarrow$ Node B $\rightarrow$ MS) of the radio interface. The amount of uplink traffic e.g., due to acknowledgements is assumed to be negligible. The simulation study focuses on the performance management for the radio interface of UMTS. Management functions for the UMTS core network are not considered. The simulator considers the UTRAN access scheme based on wideband-code division multiple access (W-CDMA) in frequency division duplex (FDD) mode proposed by 3GPP [1]. In FDD downlink, a division of the radio frequencies into four physical code channels with data rates of 1920 kbps each up to 512 physical code channels with 15 kbps data rates each is possible. Therefore, the overall bandwidth B that is available in one cell is 7680 kbps. We assume this bandwidth to be constant over time. Considering an overall bandwidth $B(t)$ depending on the actual interference situation is beyond the scope of this paper and is subject for further research. For the channel coding, we assume a convolution-coding scheme with coding factor 2.

Table 8 summarizes the base parameter settings of the network model underlying the performance experiments. Parameters are divided in fixed parameters and parameters that can be adjusted through the P-MIB. The simulations environment was implemented using the simulation library CSIM [11]. In a pre-simulation run, for each cell at the boundary of the seven cell

Table 8
Base parameter setting of the simulator

Parameter	Base value
Fixed	
Available bandwidth in one cell, B	7680 kbps
RTQ buffer size, K_{RT}	1000 IP packets
NRTQ buffer size, K_{NRT}	1000 IP packets
Adaptive	
Bandwidth partitioning: b_{h}, b_{v}, b_{p}, FCFS	5%, 10%, 20%, 65%
Bandwidth for NRT packets, b_{n}	10%
NRTQ threshold, η	90%
Weights for packet priorities high, normal, low	4/7, 2/7, 1/7

cluster the incoming handover flow of mobile users are derived iteratively from outgoing handover flows. All simulation results are derived with confidence level of 95% using the batch means method. The execution of a single simulation run requires about 30–50 min of CPU time (depending on the choice of the call arrival rate) on a dual processor Sun Enterprise with one GByte main memory.

4.2. Performance results

Using simulation experiments, we illustrate the benefit of the proposed adaptive framework for performance management of UMTS systems introduced in Section 2. The curves presented plot the mean values of the confidence intervals for the considered QoS measures. In Figs. 9–14, the arrival rate λ of new mobile users is varied to study the cell under increasing load conditions.

In a first experiment, we study the effect of the adaptive performance management in a rush hour scenario. That is, the flow of mobile users into the innermost cell is greater/smaller than out of this cell. Therefore, we balance the handover flow with a factor α. That is, the handover flow into the cell is α times the handover flow out of the cell for each service class. Fig. 9(a) plots the voice and RT handover failure probability with and without adaptive performance management for $\alpha = 0.5$, 1.0, and 1.5. Curves with adaptive performance management are denoted with MIB. Note that the handover failure probability with adaptive performance management is (nearly) independent of the value α. This is due to the fact that the handover bandwidth b_{h} is updated according to the online measured handover failure probability which immediately depends on α. As a consequence, different values of α result in different

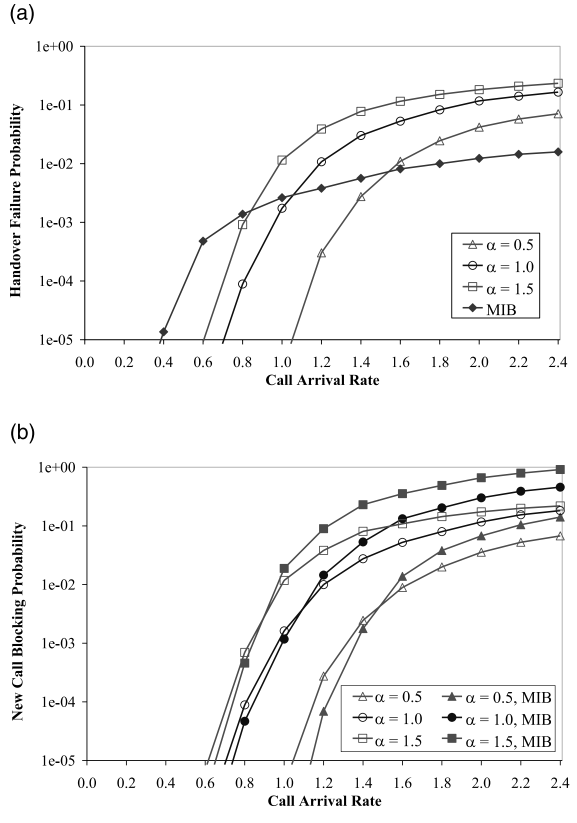


Fig. 9. Effect of adaptive performance management in a rush hour scenario.

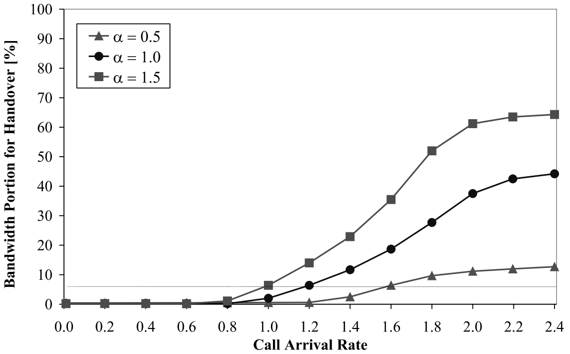


Fig. 10. Adaptive adjustment of handover bandwidth.

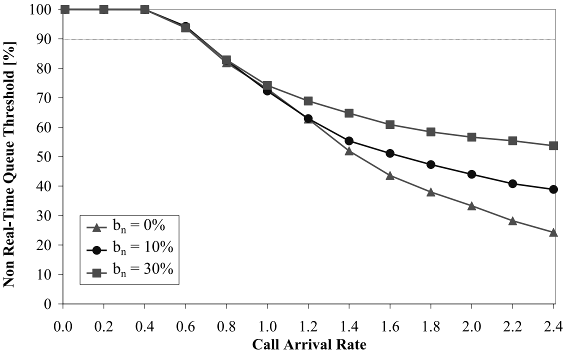


Fig. 11. Adaptive adjustment of NRTQ threshold.

values of b_h with the effect of (nearly) the same resulting handover failure probability for each α. Therefore, Fig. 9 contains only one curve for adaptive control. The adjusted values of b_h are presented in Fig. 10. The solid line corresponds to the case without adaptive control of the handover bandwidth, i.e., b_h equals to 5% for all arrival rates. As can be seen from Fig. 10 the handover bandwidth is adjusted to values below 5% for low arrival rates. Therefore, the handover failure probability is increased compared to the case without adaptive control (see Fig. 9(a)). Studying the blocking probability of new voice calls and RT sessions (see Fig. 9(b)) we surely find a higher blocking probability of new calls in the case with adaptive control and high arrival rate. From these curves we conclude that the adaptive control of b_h is very successful since the handover failure probability can be kept below 10^{-2} for a width spectrum of call arrival rates.

In a next experiment, we investigate the effect of adaptive control of the NRTQ threshold η. Fig. 12 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 bandwidth portions b_n reserved for NRT traffic as introduced in Section 2.2. We observe that the adaptive performance management achieves a significant improvement for the packet loss probability. In fact, the packet loss probability is independent of the bandwidth portion b_n. Therefore, only one curve is depicted in the figure. The adjusted values of η are presented in Fig. 11. The solid line corresponds to the case without adaptive control of the NRTQ threshold, i.e., η equals to 90% for all arrival rates. As can be seen from Fig. 11 the NRTQ threshold is adjusted to values above 90% for arrival rates below 0.7 arrivals per second. Therefore, the packet loss probability is increased compared to the case without adaptive control (see Fig. 12(a)).

Fig. 12(b) shows the average number of NRT users admitted in the cell. For all curves, the number of RT users in the cell first increases up to 53 users for an arrival rate of 0.6 arrivals per second and then it decreases depending on the choice of b_n and whether the P-MIB is active or not. This is due to the fact that for low arrival rates the bandwidth capacity that is assigned to NRT users is more than the guaranteed portion b_n. Thus, packets in the NRTQ will be served faster and more NRT users can be admitted in the cell. Comparing the curves with and without adaptive control we surely find that the number of NRT users is smaller in the case of adaptive control because the threshold parameter η is decreased in order to admit less NRT users in the cell. From these curves we again conclude that our approach successfully controls the packet loss probability of NRT traffic. In fact, the packet loss probability can be kept below 0.05 even for a highly loaded system independent of the NRT bandwidth guarantee.

Next, we study the impact on performance of NRT users by the adaptive control of the queueing weights. Fig. 13 plots the average carried traffic measured in kbps for each priority class of NRT traffic. As shown in Table 7, we

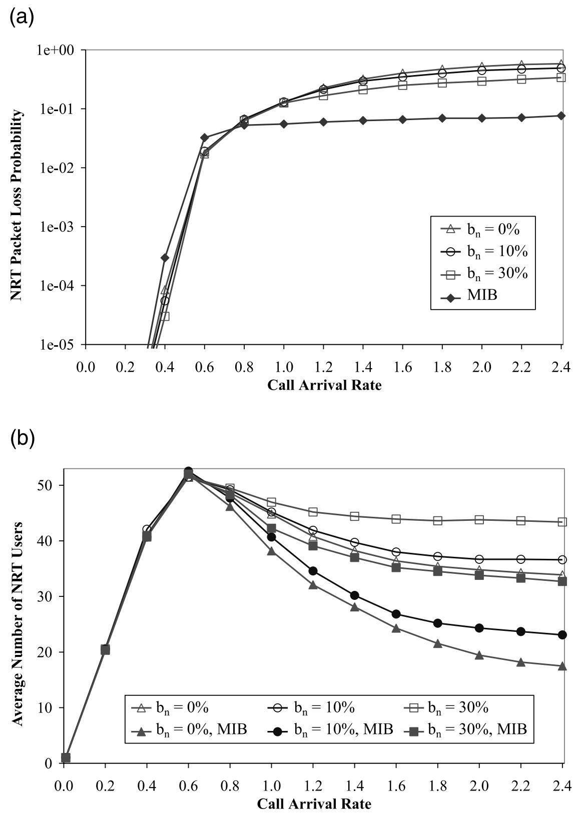


Fig. 12. Effect of adaptive performance management on NRT traffic.

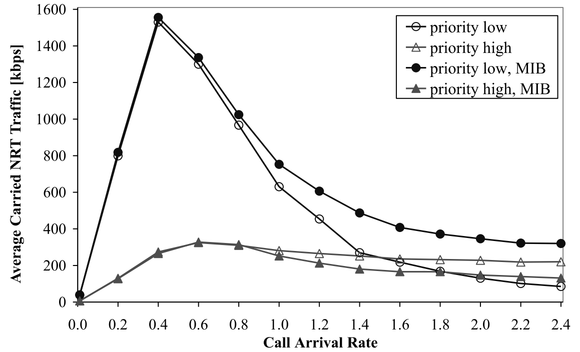


Fig. 13. Effect of adaptive performance management on queueing behavior.

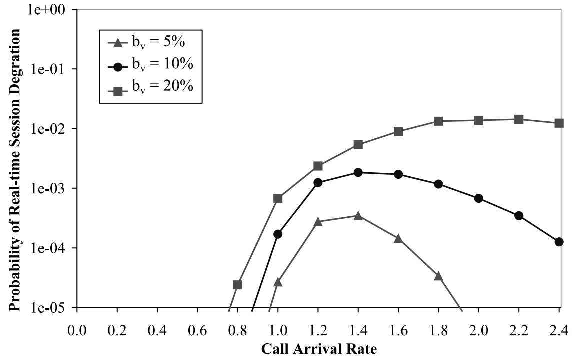


Fig. 14. Probability of RT sessions degradation for different values of b_v.

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 low (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. However, when the cell gets heavily loaded (arrival rate of more than 1.5 arrivals per second), bandwidth for low and normal priority packets will be more utilized by high priority users. The intention of adaptively controlling the queueing weights is to reduce heavy bandwidth degradation of low priority users in this case. The performance increase of low priority users and the decrease of high priority users is shown in Fig. 13.

In a further experiment, we study the effect of QoS provisioning for RT sessions, i.e., the probability of violating the QoS guarantees of corresponding QoS profiles. Fig. 14 plots curves for the probability of RT session degradation below the guaranteed bitrate. Recall that the degradation of RT sessions steams from the introduction of the voice priority preempt packets bandwidth b_v. That is, a newly arriving voice call will be admitted if less than the bandwidth portion b_v is allocated to voice calls even than if bandwidth guarantees for RT sessions will not be fulfilled anymore. Therefore, we studied the probability of RT session degradation for different values of b_v. Note that for $b_v = 0\%$ the probability of RT sessions degradation equals zero. As explained in Section 2.3 degradation of RT sessions is performed stepwise. The probabilities shown in Fig. 14 correspond to the first degradation step to 32 kbps assuming that the RT session continues in lower quality (e.g., only the audio component of a video RT session is transmitted). The probability of degrading a RT session more than one step was in all experiments less than 10^{-5}. An interesting effect than can be observed from the curves of Fig. 14 is that for increasing traffic load the probability of RT session degradation first increases and then decreases. The decrease for high traffic load is due to the fact that RT sessions get no chance to be accommodated in the voice priory preempt packets area because at all time there are sufficient voice calls in the cell. From Fig. 14 we conclude that average bandwidth requirements can almost always be maintained. In fact, the degradation of RT sessions below their guaranteed bandwidth requirements takes place only in one out of 1000 sessions for $b_{\mathrm{v}} = 10\%$.

5. Conclusions

We introduced a P-MIB for dynamically adjusting the packet scheduler and admission controller of the UMTS air interface. The aim of this adaptive control framework lies in improving bandwidth utilization of the UMTS radio channels. The proposed framework distinguishes three different types of services: circuit-switched services as well as packet-switched RT services and NRT services. The P-MIB adaptively adjusts the system parameters of the admission controller at a base station responsible for a cluster of cells. Controlled system parameters constitute the portion of bandwidth reserved for handovers, the buffer threshold of the NRTQ, and the queueing weights for scheduling NRT packets by the packet scheduler.

Using the UMTS traffic model of Ref. [19] 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 P-MIB. Considering a rush hour scenario, we showed that the adaptive performance management makes the handover failure probability insensitive against changes in the ratio between incoming and outgoing handover flows. Furthermore, we observed that the adaptive performance management achieves a significant improvement for the packet loss probability independent of bandwidth guarantees for NRT users.

Throughout the paper, we considered the services and QoS profiles standardized for UMTS. Thus, the proposed adaptive control framework is tailored to UMTS networks. However, by considering other services and QoS profiles, the basic ideas underlying the control framework can also be applied for the adaptive control of wire-line multi-service IP networks (e.g., in terms of a bandwidth broker [18] that performs intra-domain resources allocation through admission control).

References

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

[2] 3GPP, QoS Concept and Architecture, Technical Specification TS 23.107, 2001.

[3] 3GPP, UTRAN Overall description, Technical Specification TS 25.401, 2001.

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

[5] E. Anderlind, J. Zander, A traffic model for non real-time data users in a wireless radio network, IEEE Communications Letters 1 (1997) 37–39.

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

[7] J.C.R. Bennett, H. Zhang, WF^{2}Q: worst-case fair weighted fair queueing, 15th Conference on Computer Communications (IEEE Infocom), San Francisco, CA, 1996, pp. 120–128.

[8] V. Bharghavan, S. Lu, T. Nandagopal, Fair queueing in wireless networks: issues and approaches, IEEE Personal Communications 6 (1999) 44–53.

[9] C. Coutras, S. Gupta, N.B. Shroff, Scheduling of real-time traffic in IEEE 802.11 wireless LANs, Wireless Networks 6 (2000) 457–466.

[10] M. Crovella, C. Lindemann, M. Reiser, Internet performance modeling: the state of the art at the turn of the century, Performance Evaluation 42 (2000) 91–108.

[11] CSIM18-The Simulation Engine, http://www.mesquite.com.

[12] S.K. Das, R. Jayaram, N.K. Kakani, 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.

[13] A. Demers, S. Keshav, S. Shenker, Analysis and simulation of a fair queueing algorithm, Proceedings of the ACM Conference on Applications, Technologies, Architectures,

and Protocols for Computer Communication (SIGCOMM), Austin, TX, 1989, pp. 1–12.

[14] M. Elaoud, P. Ramanathan, Adaptive allocation of CDMA resources for network-level QoS assurances, Proceedings of the 6th International Conference on Mobile Computing and Networking (MobiCom), Boston, MA, 2000, pp. 191–199.

[15] ETSI, Universal Mobile Telecommunication System (UMTS); Selection Procedures for the Choice of Radio Transmission Technologies of the UMTS, Technical Report TR 101 112 v3.2.0, 1998.

[16] J.F. Huber, D. Weiler, H. Brand, UMTS, the mobile multimedia vision for IMT-2000: a focus on standardization, IEEE Communication Magazine 38 (2000) 129–136.

[17] Y. Ishikawa, N. Umeda, Capacity design and performance of call admission control in cellular CDMA systems, IEEE Journal on Selected Areas in Communications 15 (1997) 1627–1635.

[18] V. Jacobson, K. Nichols, L. Zhang, A two-bit differentiated service architecture for the internet, Request for Comments 2638, Internet Engineering Task Force, 1999.

[19] A. Klemm, C. Lindemann, M. Lohmann, Traffic modeling and characterization for UMTS networks, Proceedings of the IEEE Globecom 2001, San Antonio Texas, November 2001.

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

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

[22] T. Liu, J. Silvester, Joint admission/congestion control for wireless CDMA systems supporting integrated services, IEEE Journal on Selected Areas in Communications 16 (1998) 845–857.

[23] M. Meyer, TCP performance over GPRS, Proceedings of the 1st Wireless Communications and Networking Conference (IEEE WCNC), New Orleans, 1999, pp. 1248–1252.

[24] M. Schwartz, Network management and control issues in multimedia wireless networks, IEEE Personal Communications 2 (1995) 8–16.

[25] UMTS-Forum, UMTS/IMT-2000 Spectrum, Report No. 6, 1999.

[26] G. Záruba, I. Chlamtac, S.K. Das, An integrated admission-degradation framework for optimizing real-time call mix in wireless cellular networks, Proceedings of the 3rd International Workshop on Modeling, Analysis and Simulation of Wireless and Mobile Systems (ACM MSWiM), Boston, MA, 2000, pp. 20–26.

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

Christoph Lindemann is an Associate Professor in the Department of Computer Science at the University of Dortmund and leading 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 an author of the monograph Performance Modelling with Deterministic and Stochastic Petri Nets published by John Wiley in 1998. Moreover, he co-authored the survey text Performance Evaluation—Origins and Directions, Springer, 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.

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.

Axel Thmmler 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.