Multidimensional PRMA with Prioritized Bayesian Broadcast—A MAC Strategy for Multiservice Traffic over UMTS

Alex E. Brand and A. Hamid Aghvami, *Senior Member, IEEE*

***Abstract*—Multidimensional packet reservation multiple access is proposed as a medium-access control (MAC) strategy for the uplink channel of the UTRA (UMTS terrestrial radio access) time-division/code-division multiple-access (TD/CDMA) mode to benefit from efficient statistical multiplexing on the large common pool of available resources (i.e., slots defined both in time and code domain). A prioritized Bayesian broadcast algorithm is derived to stabilize multidimensional packet reservation multiple access (MD PRMA) and to allow for access delay discrimination of four different access classes. Access delay spread can be selected, and trading voice-dropping ratio against data-access delay is possible. To control multiple-access interference, Bayesian broadcast can be combined with load-based access control. Performance of both frequency-division duplex (FDD) and time-division duplex (TDD) mode is evaluated, the latter particularly relevant for TD/CDMA. For mixed voice, worldwide web (WWW) browsing, and e-mail traffic, the UMTS WWW model is used, while the e-mail traffic model is derived here.**

***Index Terms*—CDMA, MAC, prioritization, PRMA, stabilization, traffic models, UMTS.**

I. Introduction

A. UMTS Terrestrial Radio Access

ETSI is currently in the process of standardizing the European version of third-generation mobile communication systems UMTS. For the basic mode of operation of UTRA (UMTS terrestrial radio access), a difficult choice had to be made between orthogonal frequency-division multiple access (OFDMA), wide-band time-division multiple access (TDMA), wide-band code-division multiple access (WCDMA), and wide-band TDMA/CDMA (TD/CDMA). After a heated debate, a consensus was reached in early 1998 to use WCDMA in the paired band of the UMTS frequency allocation and TD/CDMA in the unpaired frequency band, which appears to imply that WCDMA is to be used only in frequency-division duplex (FDD) mode, while TD/CDMA is to be used in time-division duplex (TDD) mode.

WCDMA [1] is a direct-sequence wide-band CDMA system operating with a basic carrier spacing of 5 MHz. Sixteen time slots of a 0.625-ms length form a frame of 10 ms. They structure the transmitted data, but do not provide a TDMA feature, since in FDD mode a link between a mobile station and base station is continuously maintained at some minimal bit rate.

TD/CDMA [3] is essentially based on a scheme proposed in [4] and was (along with WCDMA) further investigated as mode 1 of the two final multiple-access modes considered in the European ACTS FRAMES research project [2]. It employs both CDMA and TDMA as basic multiple-access methods. Initially, the basic TDMA parameters were taken from the global system for mobile communication (GSM) [5], which would have allowed UMTS to evolve from the GSM air interface. There are now endeavors within ETSI to harmonize TD/CDMA parameters with WCDMA parameters to ease dual mode operation.

This paper is concerned with a medium-access control (MAC) strategy suitable for multimedia packet-type traffic called multidimensional packet reservation multiple access (MD PRMA). The basic strategy was presented at the ETSI UMTS workshop held in December 1996, and its voice multiplexing performance was assessed for all initial proposals submitted to ETSI except WCDMA, i.e., all proposals featuring a TDMA element. A further article (significantly different in form and content, although carrying the same title as the ETSI presentation) briefly introduced Bayesian broadcast for MD PRMA [6]. In the following, the concepts presented in these references will be significantly extended while focusing on the TD/CDMA radio interface parameters listed in [3] to investigate resulting performance with voice, worldwide web (WWW) browsing, and e-mail traffic sources. Before listing in detail the contributions of this article, earlier research on PRMA and its derivatives are briefly reviewed.

B. PRMA and Its Derivatives

From a MAC layer point of view, resources may be allocated using circuit switching or packet switching. Packet switching allows for statistical multiplexing of bursty sources such as multimedia traffic sources and packetized voice when applying voice activity detection (VAD). The amount of bursty multimedia traffic is expected to increase significantly in the near future. Therefore, packet switching appears to be the access strategy of choice for third-generation systems, with circuit switching to be supported optionally for the provision of some constant bit-rate services or very high-quality voice transmission.

Manuscript received January 15, 1998; revised July 14, 1998. This work was supported by DTI and EPSRC.

The authors are with the Centre for Telecommunications Research, King's College, London, WC2R 2LS, U.K. (e-mail: alex.brand@kcl.ac.uk; h.aghvami@kcl.ac.uk).

Publisher Item Identifier S 0018-9545(98)08270-X.

PRMA was suggested as the MAC protocol for the uplink mobile station (MS) to base station (BS) channel for microcellular third-generation systems supporting packet switching in [7]. Considerable research effort has been invested since, for instance, to investigate the performance of the basic protocol [8], to better support multimedia traffic, to increase efficiency by introducing minislots for contention [9], and to allow operation in macrocellular environments, where immediate acknowledgments are no longer possible [10], [11]. We contributed to these efforts by designing joint CDMA/PRMA [12], [13], a MAC protocol suitable for the uplink channel of hybrid CDMA/TDMA air interfaces, where load-based dynamic access control limits multiple-access interference (MAI).

C. MD PRMA with Prioritized Bayesian Broadcast

In conventional PRMA, time slots of fixed length are grouped into frames, and resources are allocated on the basis of packet spurts. In MD PRMA, slots are not only defined in the time domain, but also in an additional dimension, either the frequency domain or the code domain. This allows several *subslots* per time slot to be provided, either by pooling a number of frequency carriers together or by distinguishing these subslots with different spreading codes. Increasing the total number of slots per frame in this way results in adequate trunking or multiplexing efficiency even if the number of time slots per frame is low.

MD PRMA is suitable for any hybrid frequency-division multiple access (FDMA)/TDMA and CDMA/TDMA air interface, but here the focus is restricted to the UTRA TD/CDMA mode. MD PRMA as such is not directly applicable to the WCDMA mode, since discontinuous slotted transmission will not normally be used. Nevertheless, the concepts presented in the following, such as load and backlog-based access control with prioritization for quality of service (QoS) discrimination, can also be applied to optimize the random-access phase on a WCDMA air interface and are of particular benefit if both random access and normal traffic are to be carried in the same frame.

Pseudo-Bayesian broadcast control (henceforth referred to as Bayesian broadcast or BB) was initially introduced in [14] to stabilize slotted Aloha, and its adaptation to MD PRMA was briefly presented in [6] without including a proper derivation. This derivation is delivered in the following. Furthermore, the impact of acknowledgment delay, with and without Bayesian broadcast, and the effect of interleaving on voice-dropping performance is studied. The performance of a TDD mode referred to as MD frame reservation multiple access is investigated (FRMA, in accordance with the terminology used in [15]). The impact of MAI on Bayesian broadcast is assessed, and load-based access control similar as in [12] is combined with BB to reduce the sum of packets lost due to both MAI and packet dropping.

To investigate traffic environments in which several media are involved, voice, WWW browsing, and e-mail traffic are modeled. For voice, a common on–off model is used. The model for WWW browsing is taken from the UMTS selection procedures [16]. As a further contribution of this paper, parameters for an e-mail model are derived from large e-mail log files. Finally, prioritized pseudo-Bayesian broadcast suggested in [17] for the random access in the GSM general packet radio service (GPRS) [18] is tested in this mixed traffic environment. With this algorithm, different permission probabilities for contending terminals of different service classes can be computed to discriminate access delay experienced by these services. It will be shown how a prioritization parameter k allows trade off of voice-dropping performance against data-access delay.

In Section II, conventional PRMA will be explained briefly, and two variations of an MD PRMA scheme for FDD operation will be introduced together with MD FRMA for TDD operation. The system considered with radio interface parameters from TD/CDMA is defined in Section III, and the traffic models for voice and data are presented. In Section IV, the adaptation of the pseudo-Bayesian broadcast algorithm to the various schemes under consideration is derived, and prioritization is introduced. Simulation results on a perfect collision channel for voice-only and mixed voice-data traffic are presented in Section V. In Section VI, which precedes some concluding remarks, MAI is accounted for and BB combined with load-based access control.

II. Protocol Descriptions

A. Conventional PRMA

In conventional PRMA [8], U time slots of fixed length are grouped into frames (or TDMA frames, to distinguish them from voice frames). These slots are either available for contention (C slots) or reserved for the information transfer of a particular terminal (I slots), as indicated by the BS. When a packet spurt arrives at a terminal, it will switch from idle to contention mode, try to obtain permission to send a packet on the next available C slot by carrying out a Bernoulli experiment with some permission probability p_x, and, in the case of a positive outcome, transmit the first packet of the spurt. If this packet is received correctly by the BS, it will send an acknowledgment, which implies a reservation of the same slot (which is now an I slot) in subsequent frames for the remainder of the spurt (henceforth referred to as *implicit resource allocation*). The MS in turn switches to reservation mode and enjoys uncontested access to the channel to complete transmission of its packet spurt. In the case of a negative outcome of the random experiment or a collision on the channel with another contending terminal, the contention procedure is repeated. With delay-sensitive but loss-insensitive services, packets are *dropped* when exceeding a delay threshold value $D_{\max}$, and contention will be repeated with the next packet in the spurt. As packet dropping will cause deterioration of the perceived quality of, for instance, voice or video, some maximum admissible *packet-dropping ratio* P_{drop} will normally have to be specified.

The state diagram for mobile terminals is depicted in Fig. 1. Note that the transition from CONT to IDLE is only possible for a terminal which drops packets and may in exceptional cases have to drop an entire packet spurt. For loss-sensitive and

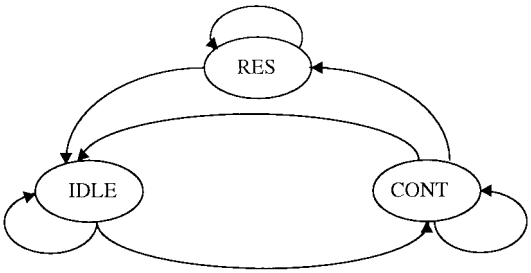

Fig. 1. State diagram of mobile terminals (MAC entity).

delay-insensitive services (e.g., e-mail and WWW browsing), henceforth referred to as nonreal time (NRT) data, packets are never dropped at the MAC, and, therefore, this transition is not possible.

B. Basic MD PRMA Protocol (Scheme A)

In MD PRMA, the time slots known from PRMA are further subdivided into Q *subslots* or simply *slots*, either using an FDMA component, thus introducing a frequency axis and frequency subslots, or in the case considered here for TD/CDMA, a CDMA component introducing a "code" axis and code subslots, as depicted in Fig. 2. As in PRMA, the channel parameters are adapted to the bit rate of the standard service (e.g., the rate of the full rate voice coder) such that during a packet spurt with this service one packet per frame needs to be transmitted on one single subslot. Due to this periodic resource requirement, such a source is termed a *periodic* information source.

The BS will have to indicate on the downlink for each subslot individually, whether it is a C or an I slot. A mobile that holds no reservation, but is admitted to the system may only access C slots in contention mode with some service and time-slot specific *permission probability* p_x (where $x = s$ for speech transmission). The fundamental extension to PRMA required in the basic MD PRMA scheme (called *Scheme A* from now on) is that in the case of multiple C slots on the same time slot an MS in contention mode will select any one of these with equal likelihood.

C. Acknowledgment in Scheme A

Basic PRMA in [7] was designed for microcellular environments to justify the assumption that the base station is able to send and the mobile stations to receive acknowledgment for a particular slot before the following slot starts. However, system constraints will normally not allow for immediate acknowledgment, furthermore, in cells serving larger areas propagation delay will also play a significant factor. Here, the impact of the BS delaying acknowledgment is studied in the following manner. A terminal which has sent a packet in contention mode in a particular time slot will not be allowed to contend again in the next x time slots, regardless of whether there are C slots in this period. The choice of the parameter x is influenced by processing delay, propagation delay, and the structure of the downlink channel. It is assumed that successfully contending mobile terminals will receive their acknowledgment in time to make use of their first reserved slot, therefore $x < U$.

D. Accounting for Interleaving—Scheme B

In conventional PRMA and Scheme A, each packet, whether sent in contention or in reservation mode, carries an addressing header, some further signaling overhead and information data. Once a logical context is established between a mobile terminal and the BS and the latter knows for instance the destination of a mobile originated call, there is no requirement to transmit the full addressing information in every packet over the air interface. The full header is therefore only required in the contention packet. On the other hand, given the adverse propagation conditions in a mobile environment, data needs to be error coded and interleaved over several time slots to provide some protection against deep fades. These considerations lead to the following evolution of the basic proposal, termed Scheme B. When a packet spurt arrives, the MS generates a dedicated request burst for contention which fits into one slot and contains most of the signaling overhead required for the packet spurt, but no user data. Upon successful contention, the MS sends its data in groups of bursts using rectangular interleaving, each burst again fitting into one slot. The group size is determined by the interleaving depth d_{il}. For the basic voice service, d_{il} is chosen such that the transmission time of these bursts corresponds to the voice frame duration D_{vf} and the choice of air interface parameters must ensure that the data in one voice frame fits onto the payload of the bursts in one group. For data services, the data transmitted in d_{il} bursts is referred to as a packet data unit (PDU).

In the case of voice service, once reservation is obtained, the newest voice frame is transmitted. This is equivalent to saying that $D_{\max}$ corresponds to D_{vf}, and dropping occurs framewise rather than packetwise, such that P_{drop} denotes the frame-dropping ratio.

E. MD FRMA for TDD Operation—Scheme C

In [15], a scheme derived from PRMA called frame reservation multiple access (FRMA) was studied in which the BS signals acknowledgment only at the end of a TDMA frame. Contending mobiles are allowed to contend repeatedly on C slots in this frame, hence, before receiving feedback. Should the BS receive several contention packets from the same MS, it will acknowledge only one of them. This scheme is particularly suitable for TDD operation, where the BS can send feedback on slot status of the uplink slots, permission probabilities, and acknowledgment in one of the downlink slots (Fig. 3), placed in a way that provides both MS and BS with suitable processing time (see also [19]).

Scheme C considered here is based on Scheme B, that is, dedicated request bursts are generated for contention and interleaving is carried out over the duration of a voice frame or a PDU. As in FRMA, the BS signals feedback after the last uplink slot of a particular TDMA frame, but before the first uplink slot of the following TDMA frame. Within the uplink slots of a TDMA frame, an MS may send multiple request bursts if it obtains permission, but at most one per time slot. In

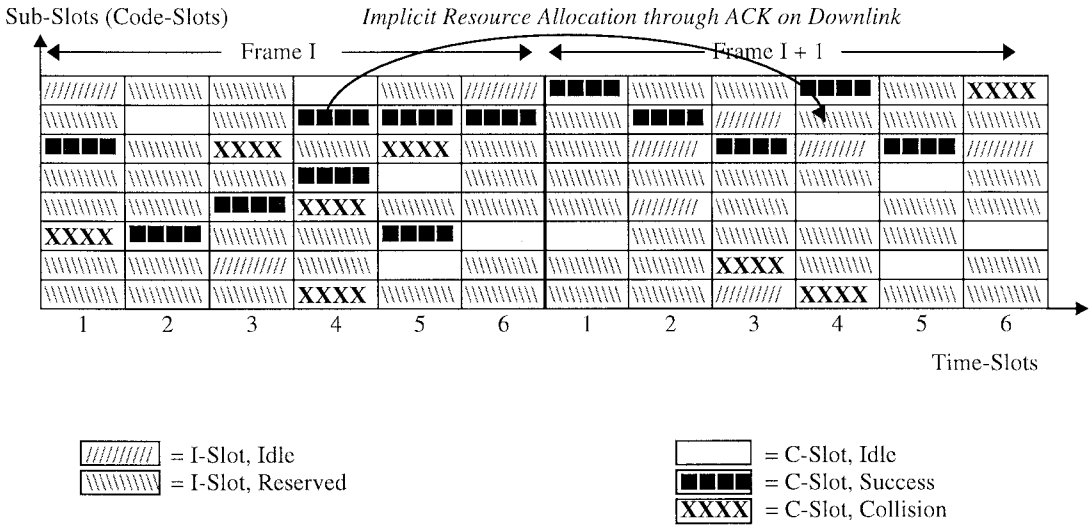

Fig. 2. Slots and frames in MD PRMA.

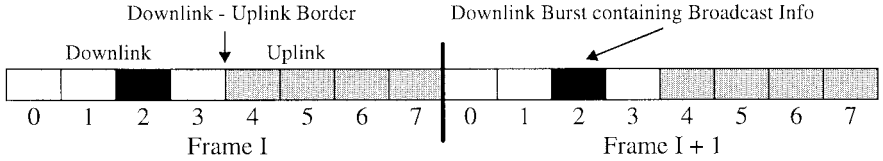

Fig. 3. Organization of uplink and downlink slots in TDD mode.

the implementation chosen, if the BS receives multiple request bursts from a single MS, it will acknowledge the first one.

F. Duration of Reservation Phase

In PRMA as in [8], an MS with periodic traffic may hold a reservation until the end of its packet spurt. An idle I slot indicates the end of a reservation to the BS and will cause it to change the slot-status to C slot. The same approach is considered here, although some protection against loss of reservation during deep fades will be required in practice. Optionally, for NRT data, the reservation phase may be limited to an allocation cycle (i.e., a limited number of PDU's) as suggested in [3]. It is assumed that terminals need to recontend for resources on C slots after expiration of their allocation cycle (another option would be to piggyback extension requests onto data transmitted in I slots). Random data traffic, which is transmitted in contention mode only, is not considered here.

G. Resource Allocation Strategies for Different Services

Some high-bit-rate services will require the allocation of multiple slots in a frame (be it an accumulation of time slots, codes, or a combination thereof) to a single user. If an MS requests several slots, the BS will have to respond with a resource grant which specifies *explicitly* the reserved resources. A simple implicit assignment of resources through acknowledgment is insufficient. Using only explicit resource allocation for all services would allow the BS to keep full control of if and when to allocate what kind of resources to which type of user. On the other hand, implicit resource allocation requires simpler acknowledgment and is particularly well suited for voice services, since their resource requests should always be satisfied to avoid deterioration of voice quality. Therefore, a hybrid approach may be the preferred option to cater for all the different needs while limiting complexity.

The scope of this paper is limited to implicit resource allocation. As a consequence, multislot allocation is not possible. Furthermore, prioritization of particular services in terms of resource allocation can only be achieved by controlling the access to C slots and choosing the length of allocation cycles, as discussed in Sections IV and V. Preemption mechanisms are not considered here.

III. System Definition

A. Radio Interface Parameters

TD/CDMA parameter values specified in [3] are considered, i.e., prior to harmonization with WCDMA parameters. $U = 8$ time slots are grouped in a TDMA frame of duration $D_{\mathrm{tf}} = 4.615$ ms as known from GSM. The time slot duration D_{slot} of 577 μs corresponds to 1250 chip periods. Carrier spacing is 1.6 MHz. Deducting guard period and training sequence and accounting for a spreading factor of 16 for the two burst types defined for data transmission, 56 and 68 symbols payload remain, respectively. For the scenario considered here, where every user may make use of at most one code per time slot, up to $Q = 8$ spreading bursts or codes may be used in the same time slot, subject to interference constraints. The data rate available for the user depends on the choice of

TABLE I
AIR INTERFACE PARAMETERS RELEVANT FOR MD PRMA

Description	Symbol	Parameter Value
TDMA Frame Duration	D_{tf}	4.615 ms
Time-slots per Frame	U	8
Codes per Time-slot	Q	8
Interleaving Depth (in bursts)	d_{il}	4 (voice, short), 96 (long)
Voice Frame Duration	D_{vf}	18.462 ms
Information Bits per Voice Frame		150
Information Bits per PDU		150 (short), 3600 (long)

data modulation, burst type, and the amount of forward error correction coding applied.

With QPSK, burst-type 1, a code rate of 0.33 and interleaving depth $d_{\mathrm{il}} = 4$ suggested in [3] for voice and short UDD 8 data (unconstrained delay data 8 kb/s), 150 b can be transmitted in a voice frame of duration $D_{\mathrm{vf}} = 18.462$ ms or in a short PDU. With $d_{\mathrm{il}} = 96$ (long UDD 8 data), 3600 b fit into a long PDU (see Table I).

B. Radio Channel

1) Perfect Collision Channel: Results reported in Section V were obtained considering a perfect collision channel and parameters listed in Table I. A packet or burst is correctly received if it is the only one sent in a given time-code slot (i.e., slot) and always rejected if it collides with another packet in the same slot, which means that the possibility of capture and the impact of MAI (whether from within the cell or other cochannel cells) and adverse channel conditions are ignored.

2) Accounting for Multiple-Access Interference: While a perfect collision channel is a suitable model for investigations on fundamental MAC features, in a MAC protocol designed for an air interface with a CDMA element MAI should be accounted for somehow. However, it would be beyond the scope of this article to establish an accurate model for performance of receivers employing joint detection (JD, a major feature in TD/CDMA) subject to complex propagation mechanisms. For results reported in Section VI, we therefore resort to simple standard Gaussian approximation to account for burst erasures due to MAI on top of those due to code collisions. Benefits (such as suppression of intracell interference) and potential problems of JD (such as channel estimation when contending users collide on the same code) are ignored. Perfect power control is assumed, and BCH codes on a burst level (no interleaving) are used to assess the burst success rate according to (2), (5), and (6) in [12]. Burst-type 2 and QPSK modulation [requires a factor of two in the denominator of (5)] are required to support BCH codes of length 127. A (127, 43, 14) code [21] is suitable for provision of a 8-kb/s service. Intercell interference is accounted for by assuming a path-loss coefficient $\gamma = 4$, reuse of one, and 80% load in interfering cells, hence, the right-hand side of (15) in [12] is modified to $P_0 \cdot I_{\mathrm{Intercell}} \cdot 0.8 \cdot Q$, with $I_{\mathrm{Intercell}} = 0.37$. This results in a burst error rate P_{ber} for noncollided bursts of no more than 10^{-5} for up to six bursts transmitted in a time slot; $4 \cdot 10^{-3}$ for eight and $1.6 \cdot 10^{-2}$ for nine bursts, respectively.

C. Downlink Signaling

It is assumed that the BS broadcasts and all MS receive information on slot status in a particular TDMA frame before (the uplink part of) this frame starts. Assumptions made on acknowledgment for Schemes A, B, and C are outlined in Section II. In the case of Bayesian broadcast, the BS will also have to broadcast regularly the permission probability values for the different services as discussed in the following section.

D. Traffic Models and Mobile Terminals

1) Voice Model and Terminals: The widely used two-state (on–off) voice activity model is used with exponentially distributed duration of voice spurts and gaps. For voice-only traffic, the mean duration of a spurt is assumed to be $D_{\mathrm{spurt}} = 1.41$ s, of a gap $D_{\mathrm{gap}} = 1.74$ s as in [10], which results in a voice activity factor α_s of 0.448. However, to be fully in line with [16], for mixed traffic, $D_{\mathrm{spurt}} = D_{\mathrm{gap}} = 3$ s and $\alpha_s = 0.5$.

A finite population of voice terminals M is considered which are all involved in a conversation, as is common in investigations on PRMA multiplexing efficiency. This means that the number of ongoing conversations M per simulation run is fixed.

2) Data Traffic and Terminals: For data traffic, an infinite population of mobiles is assumed in the sense that data sessions are generated centrally according to a Poisson process and each session is associated with a new mobile terminal. A random experiment, with parameters according to the chosen fraction of traffic for each service type, determines whether a particular session is to be a WWW session or corresponds to the transmission of a single e-mail message.

3) Model for WWW Browsing: In [16], a model for WWW browsing on both uplink and downlink was chosen in which a session is made up of several packet calls that in turn contain multiple packets or datagrams. The number of packet call requests per session N_{pc} is a geometrically distributed random variable with a mean number of packet calls $\mu_{N\mathrm{pc}} = 5$. The reading time D_{pc} between two consecutive packet call requests in a session, which starts when the last packet of a call is completely received by the receiving side, is distributed according to a geometrical distribution (in terms of simulation time steps, here D_{slot}) with mean $\mu_{D\mathrm{pc}} = 4$ s. Since $D_{\mathrm{slot}} \ll \mu_{D\mathrm{pc}}$, an exponential distribution with mean $\mu_{D\mathrm{pc}}$ can be used instead. The number of packets in a packet call N_d is again geometrically distributed and has a mean μ_{Nd} of 25 packets. For the time interval D_d between the start instances of two consecutive packets inside a packet call, the same considerations apply as for D_{pc}, thus, an exponential distribution is used here with mean $\mu_{Dd} = 0.5$ s.

Finally, the size S_d of packets in bytes is modeled using a Pareto distribution with probability density function

$$f_x(x) = \begin{cases} \frac{\lambda \cdot (e^{\beta})^{\lambda}}{x^{\lambda+1}}, & x \geq e^{\beta} \\ 0, & x < e^{\beta}. \end{cases} \tag{1}$$

Note that these packets need to be segmented into smaller packets at the MAC, hence constitute a packet spurt for MD PRMA. If X is Pareto distributed with parameters e^{β} and λ, then $\ln(X)-\beta$ is distributed according to an exponential distribution with parameter λ. The mean of the Pareto distribution is

$$\mu = \frac{\lambda(e^{\beta})}{\lambda - 1}, \qquad \lambda > 1 \tag{2}$$

and the variance is only finite for $\lambda > 2$.

In [16], $e^{\beta} = 81.5$ and $\lambda = 1.1$, therefore, all outcomes $> c = 67$ Kbytes are set equal to c to ensure finite variance. The mean of this truncated Pareto distribution is

$$\mu = \frac{\lambda \cdot e^{\beta} - c(e^{\beta}/c)^{\lambda}}{\lambda - 1} \tag{3}$$

and the variance (provided that $\lambda \neq 2$) is

$$\sigma^2 = \frac{\lambda \cdot e^{2\beta} - 2c^2(e^{\beta}/c)^{\lambda}}{\lambda - 2}. \tag{4}$$

In the case considered here, the mean before truncation (2) is 896.5 bytes and after truncation (3) 481 bytes, which shows the impact of a few very large packets on the mean of the Pareto distribution. In fact, the main feature of a Pareto distribution compared to a shifted exponential distribution is its elongated tail. With these parameters, the service has a "peak" rate of roughly 8 kb/s. Given the nonnegligible probability that packets of size 67 Kbytes are generated which require 68 s to be transmitted with single slot allocation, queues of significant sizes have to be introduced to avoid dropping of packets in a packet call. A queue which can hold up to 50 packets (regardless of their size) is included here with every data terminal. This queue is conceptually situated above the MAC layer.

4) Model for E-Mail Traffic: Several attempts were made in the LINK ACS project to model e-mail traffic using e-mail log files of the different partners involved. From these files, a minimum e-mail size of around 500 bytes can be deduced, which is mainly due to a minimum amount of header information required in an e-mail. Furthermore, due to the increasing number of attachments in business traffic, a distinctly elongated tail can normally be observed resulting in large mean e-mail sizes, as depicted in a histogram of one of the log files with S_d up to 5.6 Mbytes (Fig. 4). Therefore, a prime candidate distribution to model e-mail size is the Pareto distribution. There are, however, some problems.

First, according to (1) both minimum size and mode (i.e., the value of x yielding maximum f_x) equal e^{β}, whereas from histograms with a bin size of 100 bytes one obtains between 1600–1800 bytes for observed modes (note that the bin size in Fig. 4 is 1000 bytes). Next, using (1), a small minimum size e^{β} does not preclude arbitrary large μ (choosing λ close to one), which allows capture of these two important features of observed e-mail statistics. One would thereby, however, ignore the fact that e-mails are in practice subject to size limitation and have to live with an infinite variance, which is somewhat of an obstacle when trying to get meaningful performance results through simulations. With

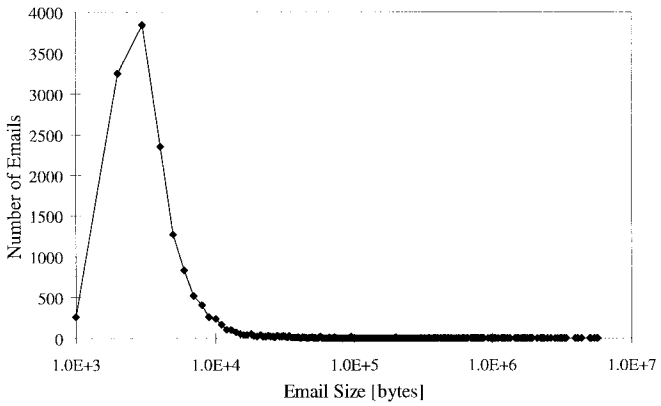

Fig. 4. Example of an e-mail log file histogram with 15 000 samples.

the Pareto distribution truncated at c, on the other hand, the choice of e^{β} imposes an upper limit on μ which will make it impossible in practice to fit both minimum size and mean of observed distributions for any reasonable value of c, say 500 Kbytes. Choosing $e^{\beta} = 500$ bytes to fit the observed minimum size would result in $\mu < 4000$ bytes, which is much lower than for instance the mean size of 26 970 bytes over these 15 000 e-mails considered in Fig. 4. Summarizing, it is not possible to fully fit minimum size, mode, and mean of observed traffic concurrently with a truncated Pareto distribution.

At this point, it is worth mentioning that the data collected on e-mail size is considerably affected by size limitations at the gateways of different companies. Furthermore, traffic statistics are likely to be affected by the change of medium from wire to air, the latter normally associated with rather low bit rates and high costs. With these comments in mind, we might ignore e-mails larger than 500 Kbytes, which reduces the observed mean to 9450 bytes. Doing this, a compromise parameter choice of $e^{\beta} = 1400$ bytes, $\lambda = 1.01$, and $c = 500$ Kbytes which yields $\mu = 9423$ bytes using (3), in combination with using only long PDU's carrying 450 bytes, seems justifiable. While with this model, no e-mails with $S_d < 1400$ bytes can be generated, thanks to the long PDU's, only those e-mails with $500 < S_d < 900$ bytes would require fewer resources (two PDU's instead of three) than can actually be modeled with our model. Compared with the fundamental limitations of a model for UTRA e-mail traffic based on observations from computer networks, this compromise appears acceptable and allows adherence to the model structure for WWW browsing. Only a single e-mail message is generated per session, hence, $\mu_{N_{\mathrm{pc}}} = \mu_{Nd} = 1$, and the values for $\mu_{D_{\mathrm{pc}}}$ and μ_{Dd} are irrelevant.

IV. Prioritized Bayesian Broadcast for MD PRMA

A. Stabilization of Slotted Aloha with Ternary Feedback

In [20], various stabilization methods for slotted Aloha based on ternary (hole or idle, success, collision) channel feedback were compared. It was found that the best delay-throughput performance was provided by methods that use deferred first transmission (DFT) and estimate the number of backlogged terminals (i.e., those having something to

transmit) to calculate the transmission permission probability p, like pseudo-Bayesian broadcast [14]. DFT means that a terminal with a newly arriving packet is considered to be backlogged immediately and is subject to the same permission probability p as a terminal which has to retransmit a packet following a collision.

Let N_t denote the number of backlogged terminals at the start of time slot t. In Bayesian broadcast, Bayes' rule is used to update, after each time slot, the estimated probability that $N_t = n$ stations are backlogged, for each $n \geq 0$. In pseudo-Bayesian broadcast, it is assumed that these probability values can be approximated reasonably well by a Poisson distribution, hence, instead of individual probability values, only the mean v of the Poisson distribution needs to be estimated, and the optimum permission probability value p in terms of minimum delay or minimum backlog is simply the inverse of the mean $p = 1/v$. If the backlog estimation is accurate, this will ensure that the expected offered traffic G assumes a value of one which is the well-known optimum offered traffic for slotted Aloha. In the following, when referring to Bayesian broadcast or BB we actually mean pseudo-Bayesian broadcast.

The algorithm, which only needs binary feedback, reads [14] as follows.

1) At $t = 0$, set v to one.
2) Each station which has a packet to transmit obtains permission to transmit with probability $p = 1/v$.
3) Decrement v by one if the current slot is a hole or success and increment v by $(e-2)^{-1} = 1.3922$ if the current slot is a collision.
4) Set v to $\max(v + \hat{\lambda}_{\text{ar}}, 1)$, where $\hat{\lambda}_{\text{ar}}$ is an estimation of the arrival rate of new packets per slot (see below) and go to step 2) for slot $t+1$.

B. Bayesian Broadcast on Multiple Simultaneous Orthogonal Resources

Suppose for now that $R = Q$ slots in a time slot (i.e., all of them) are C slots available for contention. If they were observed globally to derive the feedback, $1+2+\cdots+(R+1)$ different global observations could be made, for instance six for $R = 2$ (two holes, one hole, and one success, etc.). While it is very tedious, we did calculate all necessary probabilities and followed the approach in [14] for six different observations with $R = 2$ to obtain the relevant update values, the same has, however, not been tried for $R = 3$ and is certainly not advisable for $R = 8$, as there are 45 different observations.

Alternatively, note that the choice of a particular slot out of R in the case of a successful outcome of the Bernoulli experiment with parameter p is not conditioned on this Bernoulli experiment other than that it would normally not be made if the outcome was negative. Thanks to this independence, if the total backlog is Poisson with mean v, then the "backlog per slot" is Poisson with mean v/R. Furthermore, under perfect channel assumption, the slots are orthogonal. As we are dealing with relative increments or decrements in the Bayesian broadcast algorithm, which do not depend on the value of v [except for the "max" operation in step 4)], we can update the backlog estimation for each subslot individually and add the resulting sum of update values in this time slot to the previous estimation of v. The pseudo-Bayesian algorithm for R available subslots per time slot can therefore be written as follows.

1) At $t = 0$, set v to R.
2) Each station which has a packet to transmit obtains permission to transmit with probability $p = \min(1, R/v)$ and selects a slot at random, if $R > 1$.
3) Observe the number of collision slots C in this time slot and set v to $v + C/(e-2) - (R-C)$.
4) Set v to $\max(v + \hat{\lambda}_{\text{ar}}, R)$ and go to step 2) for slot $t+1$.

The "max" operation in step 4) is required for arguments equivalent to those in [14] for $R = 1$.

C. Bayesian Broadcast for MD PRMA (Schemes A and B) with Immediate Acknowledgment

In MD PRMA, only a subset R of the Q slots per time slot may be available for contention, and this number of slots relevant for Bayesian broadcast varies from time slot to time slot, hence, $R[t]$. At desirable operating points of the protocol, packet dropping should be such that the probability of an entire packet spurt being dropped is negligible. This means that packet dropping has no impact on the backlog distribution, since a terminal will remain backlogged after dropping a packet. If immediate acknowledgment is assumed $(x = 0)$, the algorithm for MD PRMA (Schemes A and B) reads as above, but with $R = R[t]$.

Until now it has not been specified whether the feedback should be evaluated by each MS or by the BS. In a mobile communications system, feedback evaluated at the MS is not reliable and should therefore be evaluated at the BS. The BS could broadcast the feedback, which is not convenient, since even an inactive MS would have to listen permanently to this feedback to maintain an accurate estimation of the backlog. Hence, either v or the resulting p value is broadcast instead. We carried out earlier investigations for GPRS to assess the impact of quantization of p and of reduced update intervals. It was found that a 4-b geometric quantization yielded close to maximum performance, while updating of p only every fourth or eighth slot had to be paid for with some performance degradation. Here, for Schemes A and B, it is assumed that p is broadcast at the end of each time slot in such a manner that it is available to all MS with full precision before the next time slot starts.

D. Accounting for Acknowledgment Delays

Next, the impact of acknowledgment delays (i.e., $x > 0$) on BB is studied. Since $x < U$, the introduction of acknowledgment delays only has an impact on terminals with colliding contention packets, but not on an MS which contends successfully. The backlog estimation as such needs no modification, but when calculating p for a particular slot one has to consider that those backlogged terminals which suffered a collision less than x time slots ago will not attempt to obtain permission. If broadcast control is accurate, the offered traffic per C slot will be Poisson with rate $G = 1$, and, consequently,

the average number of terminals involved in a collision is

$$\mu_{\text{coll}} = \frac{1}{1-2/e}\sum_{k=2}^{\infty} k\frac{e^{-G}G^k}{k!}\bigg|_{G=1} = \frac{1-1/e}{1-2/e} = 2.3922. \tag{5}$$

The number of terminals waiting can then be estimated using

$$w = C_x \cdot \mu_{\text{coll}} \tag{6}$$

where C_x is the number of collisions observed in the last x time slots.

Finally, if $w > v - 1$, set $w = v - 1$ and the permission probability to

$$p = \min\left(\frac{R[t]}{v-w}, 1\right). \tag{7}$$

It is again assumed that this value is broadcast at the end of each time slot in such a manner that it is available to all MS's in full precision before the next time slot starts.

E. Bayesian Broadcast for MD FRMA (Scheme C)

Compared with MD PRMA, the Bayesian broadcast algorithm is modified in the following manner.

1) At $t = 0$, set v and v' to one.
2) Each station which has a packet to transmit obtains permission to transmit with probability $p = \min(R[t]/v, 1)$ and selects a slot at random, if $R[t] > 1$.
3) Observe the number of collisions C in this slot and set v' to $v' + C/(e-2) - (R[t] - C)$.
4) If slot $t+1$ is in the same frame as slot t, set v' to $v' + \hat{\lambda}_{\text{ar}}$ and go to step 2) for slot $t+1$, otherwise, go to step 5).
5) Set v and v' to $\max(v' + \hat{\lambda}_{\text{ar}} + s'', 1)$ and go to step 2) for slot $t+1$.

Since each successful MS can reduce the backlog only by one, but is counted in step 3) for every successful contention, in step 5) the number of received contention packets except the first one of each MS s'' is added for compensation.

The backlog estimation v used to calculate p remains constant during the frame. Broadcasting v once per frame before its uplink section starts, together with the slot status information, is therefore sufficient.

F. Estimation of Arrival Rate λ_{ar}

Rivest suggested in [14] either use of $\hat{\lambda}_{\text{ar}} = 1/e$ (which is the maximum arrival rate per slot allowing for stable operation) or estimating the rate of new arrivals λ_{ar} according to

$$\hat{\lambda}_{\text{ar}}[t+1] = 0.995\hat{\lambda}_{\text{ar}}[t] + 0.005I(s[t]) \tag{8}$$

where $I(s[t]) = 1$ if slot t is a success slot and otherwise zero.

In [20], it is reported that for slotted Aloha the effort in estimating λ_{ar} using (8) rather than setting $\hat{\lambda}_{\text{ar}}$ to $1/e$ is not rewarded by improved performance, which agrees with our observations reported in [17]. The analogous approach in MD PRMA would be to set $\hat{\lambda}_{\text{ar}} = \bar{R}/e$ with $\bar{R}$ the average number of contention slots per time slot over some time window. However, since C slots and I slots share the same resources in MD PRMA, such that random access and information transfer cannot be decoupled, this approach yields unsatisfactory performance, and λ_{ar} needs to be estimated. If the statistical parameters of the voice service are known to the BS and since the BS must have control over the admitted voice users M, the voice arrival rate λ_v can be estimated using

$$\hat{\lambda}_v = \frac{M \cdot D_{\text{slot}}}{D_{\text{spurt}} + D_{\text{gap}}}. \tag{9}$$

For data, $\hat{\lambda}_d$ is estimated using (8), obviously counting only successful contentions of data users, and, finally, $\hat{\lambda}_{\text{ar}} = \hat{\lambda}_v + \hat{\lambda}_d$.

G. Introducing Prioritization

A prioritized version of Bayesian broadcast, which allows calculation of multiple permission probabilities per time slot to discriminate the QoS of different service classes was studied for the random-access phase of GPRS [17] and is here applied to MD PRMA. Four different permission probability values p_i are calculated for four different access classes i based on the basic value p in a manner that leaves the offered traffic unaltered compared to the basic approach

$$p_1 = \min(1, m \cdot p) \tag{10}$$

$$p_2 = \min\left(1, \frac{z_1 m + z_2 k}{z} p\right) \tag{11}$$

$$p_3 = \min\left(1, \frac{z_2 m + z_1 k}{z} p\right) \tag{12}$$

$$p_4 = k \cdot p \tag{13}$$

with

$$m = \frac{1 - \alpha \cdot k}{1 - \alpha} \tag{14}$$

$$z = z_1 + z_2 \tag{15}$$

and

$$\alpha = \left(\frac{z_2}{z}S_2 + \frac{z_1}{z_2}S_3 + S_4\right)\frac{1}{S} \tag{16}$$

where S_i is the C slot throughput of class i. Class 1 has highest priority, and parameter k $(0 < k \leq 1)$ determines the global amount of prioritization. With $k = 1$, p_i equals p for $i = 1 \cdots 4$, which is equivalent to Bayesian broadcast without prioritization, while with decreasing values of k the spread of the different p_i values increases and, consequently, also the spread of the access delay values experienced. The algorithm is a generalized nonproportional version of [17], where z_1 was always set to two and z_2 to one, which was referred to as "semiproportional priority distribution." With z_1 and z_2 variables, the relative degree of prioritization of classes 2 and 3 can be controlled, if required.

If α were the backlog proportion rather than the throughput proportion, one could easily show that the application of (10)–(16) results in the same offered traffic as that generated with the single-class scheme [17]. However, the BS knows only the throughput proportion, which is not equivalent to the backlog proportion. To our satisfaction, the algorithm still

controls the offered traffic efficiently to the optimum traffic, as outlined in the same reference, which also explains how the functional relations between p_i and p allow limiting of downlink signaling load. Here, it is assumed that for each access class the p_i value is known with full precision at the start of each time slot.

V. Performance Evaluation for Perfect Collision Channel

A. Simulation Approach for Voice-Only Traffic

In [6], multiplexing efficiency η_{mux} of different UTRA candidates using MD PRMA was compared. Multiplexing efficiency is defined as

$$\eta_{\text{mux}} = \frac{M_{0.01} \cdot \alpha_s}{U \cdot Q}. \tag{17}$$

From Section III, we have $U = Q = 8$ and activity factor $\alpha_s = 0.448$. $M_{0.01}$ is the protocol capacity defined in [7], which is the maximum number of simultaneous conversations M which can be sustained while not exceeding a packet-dropping ratio P_{drop} (averaged over all terminals) of 1%. Here, we are more interested in the dropping behavior as a function of M resulting from the different schemes introduced in Section II to understand how much capacity reduction is required to guarantee lower dropping rates. This dropping behavior is established by extensive simulation studies. During each simulation run of 1000-s simulation time, M is kept constant. Where required, multiple simulation runs with the same M are carried out and the resulting P_{drop} averaged.

B. Results for Voice-Only Traffic, FDD Mode

Fig. 5 shows results for the basic MD PRMA scheme, Scheme A, with a very short packet-dropping delay threshold $D_{\max}$ of 4.615 ms, which corresponds to D_{tf}. Performance when using fixed speech permission probability values p_s is compared with that when Bayesian broadcast is used to calculate p_s. With BB, $M_{0.01} = 131$ and $\eta_{\text{mux}} = 0.92$, with fixed p_s, $M_{0.01}$ lies between 121 (for $p_s = 0.1$) and 131 ($p_s = 0.3$). While it is possible to achieve high capacity with fixed p_s, it is not possible to achieve high capacity and low packet dropping at lower load with the same value of p_s. Furthermore, if p_s is too large, MD PRMA can become unstable, which was experienced with the values considered here for M for $p_s \geq 0.6$ and $M = 140$. Choosing $p_s = 0.5$ offers the best compromise between capacity ($M_{0.01} = 128$) and low dropping at low load, while appearing to allow for stable operation up to $M = 140$. BB on the other hand allows for stable operation at high load while ensuring low packet dropping at low load and performs at least as well as the fixed p_s approach over the entire range of M considered. While it would be possible to run MD PRMA with optimized fixed p_s for each value of M for voice-only traffic and thereby meet performance achieved with BB, this is not straightforward with mixed traffic, particularly with unknown source statistics.

Next, the impact of acknowledgment delays in Scheme A with Bayesian broadcast is studied. Unfortunately, they have a rather negative impact on voice dropping performance (Fig. 6). With increasing x voice dropping increases, particularly at low load. The reason is as follows. At low load, the backlog is normally close to zero and, consequently, $p_s = 1$. A sudden increase in backlog during a single time-slot period at a time when $R[t]$ is low will normally result in a number of collisions, which will cause the algorithm to adapt to the situation and lower p_s. With $x = 0$, this can be achieved quickly and packets are rarely dropped. With $x > 0$, collisions will occur with period $x + 1$ time slots, so spreading them in time, and the algorithm needs longer to adapt. Even worse, it may be deceived by time slots with numerous idle and success C slots lying in between the "collision time slots," which further delays tracking of the real backlog, resulting in packet dropping for those terminals caught in the "collision slots." Interestingly, $x = 3$ produces particularly bad results, which are even worse than those for $x = 4$ and 5. Here, a further factor comes into play. Not only are collisions repeated every four time slots, but also the $R[t]$ patterns will exhibit some repetitive behavior with double the period ($U = 8$ time slots), thus a "bad slot" in terms of backlog may coincide regularly with "bad slots" in terms of low $R[t]$ values. This could probably be described as "resonant behavior" or "local catastrophes." Consequently, increased average P_{drop} in such scenarios is due to a few MS's suffering severe dropping, while those MS's never caught in a "bad slot" experience very moderate dropping.

Fig. 7 compares the impact of acknowledgment delays on the fixed permission probability approach with that on Bayesian broadcast. Dropping performance with $p_s = 0.2$

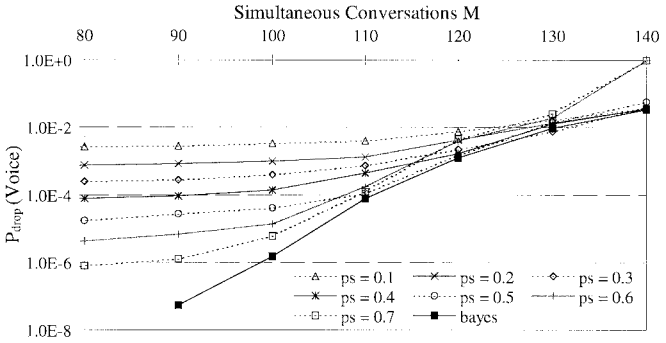

Fig. 5. Bayesian broadcast versus fixed p_s with immediate acknowledgment.

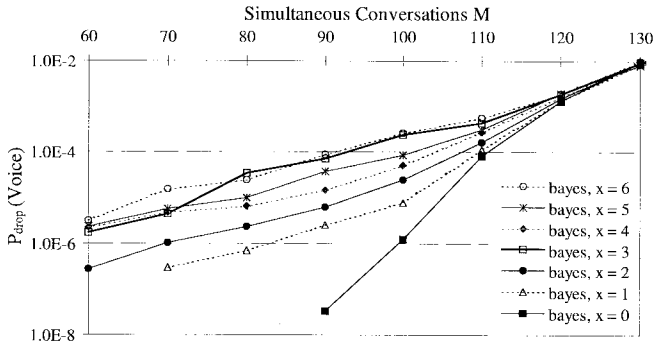

Fig. 6. Impact of acknowledgment delay on Bayesian broadcast (x = number of forbidden time slots after contention).

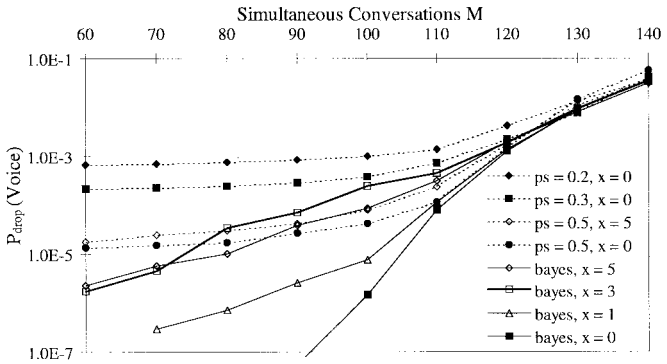

Fig. 7. Bayesian broadcast versus fixed permission probabilities with acknowledgment delays.

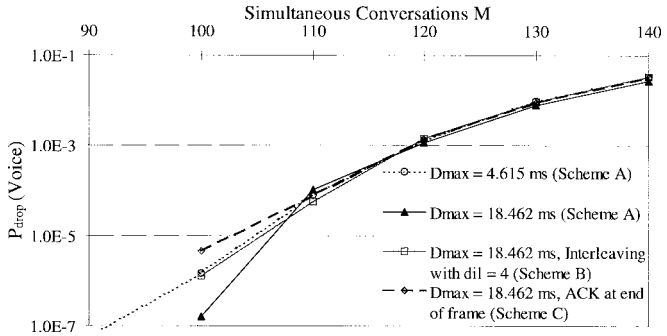

Fig. 8. Performance comparison of MD PRMA with (Scheme B) and without interleaving (Scheme A) and MD FRMA in FDD mode (eight uplink slots, Scheme C).

and 0.3 does not depend on x, and in this case, dropping is almost uniquely due to the MS not getting permission to send contention packets, while C-slot collisions occur rarely. With $p_s = 0.5$, dropping performance depends to some extent on x, although to a far lesser extent than observed with Bayesian broadcast. With $x = 5$, performance advantage of the Bayesian approach dwindles away and $x = 3$ must obviously be avoided for BB. To mitigate the problem with $x = 3$, p_s calculated with BB could be limited to a maximum value $p_{s\max}$ below one, but this would defeat the purpose of broadcast control, which is to stabilize the protocol and ensure efficient operation for all possible traffic scenarios without having to fix traffic dependent parameter values $p_{s\max}$.

Fig. 8 allows comparison of behavior of all three schemes defined in Section II: Scheme A with the short $D_{\max}$ considered before and with a longer $D_{\max}$ of 18.846 ms which is equivalent to that of Schemes B and C. Immediate acknowledgment is assumed for Schemes A and B, while acknowledgment delay is inherent in Scheme C. Note that Scheme C is considered here with eight uplink time slots per frame using FDD mode, which would in practice not leave any time to signal acknowledgment for the entire frame before the start of the next frame.

Judging from Fig. 8, choosing $D_{\max}$ anywhere in between D_{tf} and D_{vf} has a very limited impact on P_{drop}. The same can be said about the (in fact, small) additional load created by the dedicated request bursts and the impact of interleaving and voice framewise dropping in Scheme B. The small gain

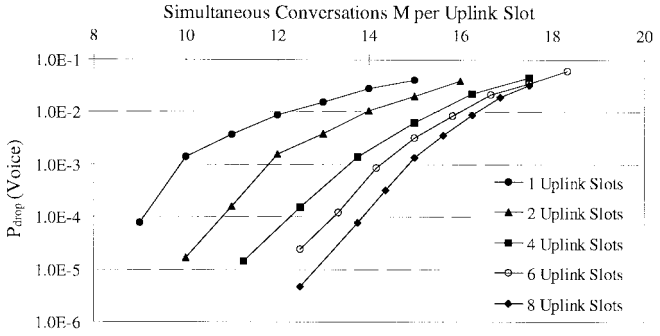

Fig. 9. MD FRMA (Scheme C) with Bayesian broadcast and variable number of uplink slots.

achieved by increasing $D_{\max}$ is almost exactly compensated by the interleaving applied and the request bursts created in Scheme B. Most interestingly, unlike Scheme A, Scheme C does not seem to suffer from acknowledgment delays. With the explanation for the bad performance of Scheme A with acknowledgment delay in mind, it is clear that allowing terminals to contend repeatedly before receiving feedback must mitigate the problems experienced there considerably. As an intermediate conclusion we can state that, even with Scheme C where acknowledgments do not occur immediately, we can exploit statistical multiplexing to a considerable extent while keeping packet dropping at very moderate levels. With $M = 110$ conversations sharing 64 slots, P_{drop} is below 10^{-4}. With $M = 90$, in most schemes not a single packet or frame was dropped. The additional delay incurred by using packet switching instead of circuit switching is moderate and always $< D_{\text{vf}}$.

C. Voice-Only Traffic in TDD Mode Using MD FRMA (Scheme C)

In Fig. 9, the voice-dropping performance of Scheme C using TDD mode with one, two, four, six, and eight time slots in the uplink direction, respectively, is shown. P_{drop} is reported as a function of M per uplink time slot. By doing so, the impact of trunking efficiency on dropping performance becomes immediately apparent. The findings are very similar to those reported in [6] when modifying the number of slots per time slot Q for GSM. Looking for instance at $M_{0.01}$, 12 conversations can be supported with only one time slot, whereas more than 16 per time slot can be supported in the case of having eight uplink time slots per frame.

D. Simulation Scenarios for Mixed Voice-Data Traffic

Three scenarios for mixed voice-data traffic are considered to investigate impact of data traffic on voice-dropping performance and of the prioritized Bayesian broadcast on the access delay performance of data assigned to different access classes: voice and WWW browsing traffic, voice and e-mail traffic, and, finally, voice and both data traffic types together. Access delay is defined as the time between arrival of a request at the MS MAC entity (hence excluding queuing delay while MAC is busy transmitting previous packet spurts) and the end of the successful C slot. Short PDU's are used for

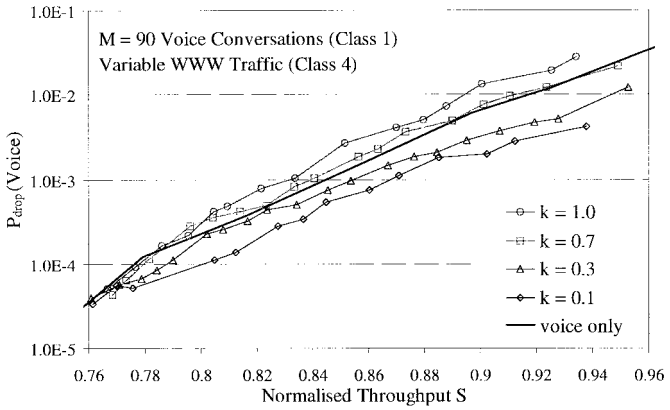

Fig. 10. Voice-dropping performance with mixed voice and WWW data traffic.

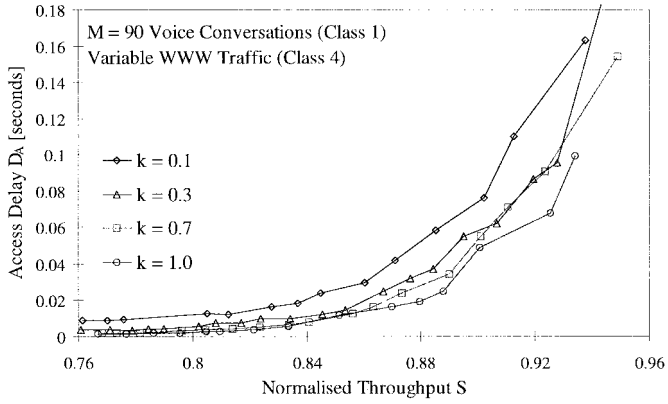

Fig. 11. Delay performance of data with mixed voice and WWW data traffic.

WWW packets, long PDU's for e-mail messages. With voice and a single type of data traffic, voice uses access class 1 and data uses access class 4. In scenario 3, with mixed data sources, voice is again in access class 1, WWW browsing in access classes 2 and 3, and e-mails use access class 4. Performance is investigated using the global mean session interarrival time $\mu_{D\,\mathrm{sess}}$ as the main simulation parameter, while keeping the number of ongoing voice conversations fixed at M. In scenario 3, the fraction of sessions for each access class is chosen such that on average 33% of the data bursts generated carry e-mail traffic, 33% WWW class 2 and 33% WWW class 3 traffic. Each simulation run lasts 1000-s simulation time. While performance is investigated with several values of k, z_1, and z_2 are always set to two and one, respectively. Allocation cycle length is unlimited, unless otherwise mentioned.

Given the relatively few sessions per simulation run ($\mu_{D\,\mathrm{sess}}$ between 0.5–12 s) and the large variance of the packet size S_d (4), the amount of data generated per simulation run even with constant $\mu_{D\,\mathrm{sess}}$ fluctuates considerably and results presented as a function of $\mu_{D\,\mathrm{sess}}$ will not be meaningful. Instead, we report results as a function of the normalized throughput S (normalized to the total user channel rate of 520 kb/s). This has two added benefits: it allows comparison of P_{drop} for voice in the considered scenarios with P_{drop} for voice-only traffic (obviously obtained with the same voice model, that is, $D_{\mathrm{spurt}} = D_{\mathrm{gap}} = 3$ s). Furthermore, dropping of datagrams, due to overflow of the queue of data terminals serving WWW traffic, does not affect results other than altering the traffic model. Note that the respective dropping ratio was always below 1%. Throughput includes all bursts received by the BS including fill bursts in the last PDU and request bursts. If the throughput S generated in two different simulation runs does not differ by more than 1%, results are averaged and reported as one point on the graphs presented.

E. Results for Mixed Voice-Data Traffic with Scheme B

In Fig. 10, voice-dropping performance with $M = 90$ conversations and a variable amount of WWW traffic are depicted for several values of the prioritization parameter k and compared with performance of pure voice traffic. To our satisfaction, heterogeneous traffic does not necessarily result in increased P_{drop} compared to homogeneous voice traffic. Furthermore, the prioritization parameter k not only influences the access delay depicted in Fig. 11 as desired, but also allows trade off of voice dropping against data-access delay. For $k \leq 0.3$, P_{drop} is below that with voice-only traffic at the expense of high data-access delay, and vice versa for $k \geq 0.7$. Note that $S = 0.76$ roughly corresponds to 8% data traffic and $S = 0.95$ to 27%.

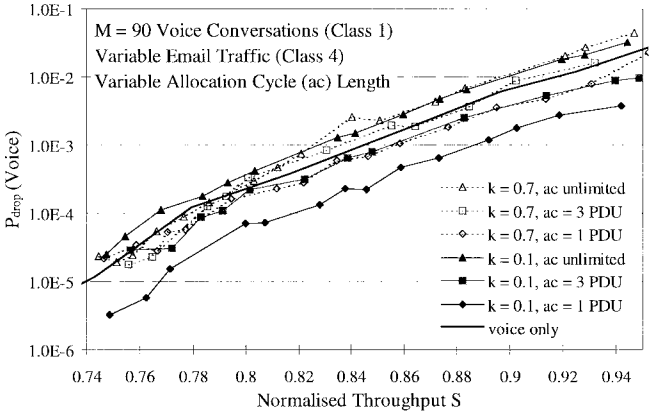

Fig. 12. Voice-dropping performance with mixed voice and e-mail traffic, variable allocation cycle length (in terms of long PDU's per cycle).

Figs. 12 and 13 report results for voice and e-mail traffic. With an unlimited allocation cycle, mainly due to the high variance of the generated e-mail messages (4), both for $k = 0.1$ and $k = 0.7$, voice-dropping performance in the heterogeneous case is slightly worse than that in the homogeneous case. Also, due to the large mean size of e-mail messages compared to the average data contained in a voice spurt and their low arrival rate, varying k has no impact on P_{drop}. This can be corrected by limiting the allocation cycle length. In doing so, both P_{drop} is reduced and the performance spread between $k = 0.1$ and $k = 0.7$ is increased, as shown in Fig. 12 for allocation cycle lengths of three and one PDU, respectively. Obviously, this has to be paid for by increased data transfer delay due to longer time spent in contention state (Fig. 13).

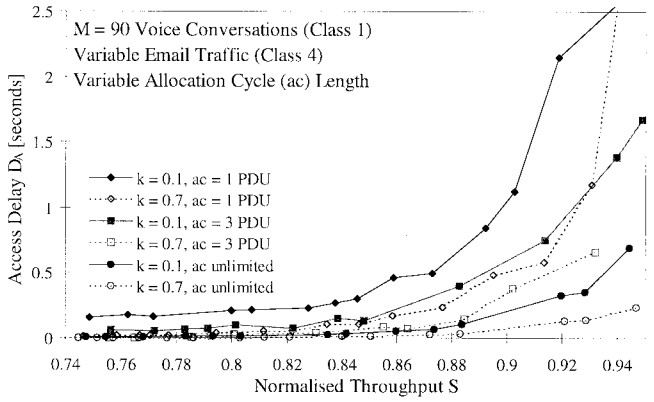

Fig. 13. Mixed voice and e-mail data traffic, access delay of data (cumulative in case of limited allocation cycles, i.e., total time spent in contention state).

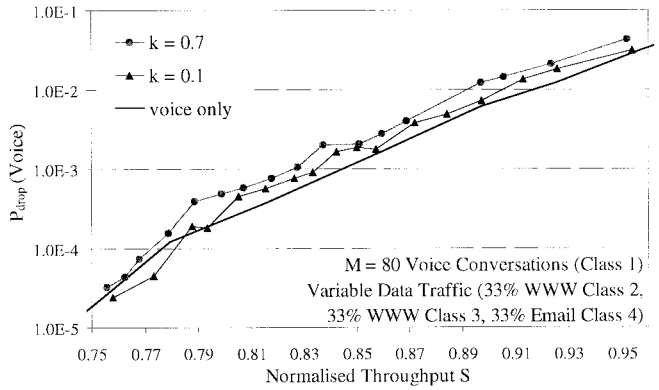

Fig. 14. Voice-dropping performance with mixed voice, WWW, and e-mail data traffic.

Finally, Figs. 14 and 15 report the results for the mixed scenario 3 with $M = 80$. Here, $S = 0.75$ and 0.95 correspond to 17% and 36% of data traffic, respectively. As far as voice dropping is concerned, the performance degradation due to heterogeneous traffic seems rather worse than that shown in both Figs. 10 and 12 for unlimited allocation cycles. This is due to the fact that here data is assigned to access classes 2 to 4 rather than to 4 only, such that the "average access penalty" for data at a given value of k is smaller. Together with the above considerations on e-mails with unlimited allocation cycles, this also explains the small spread between the two curves in Fig. 14. In Fig. 15, the effectiveness of the prioritized Bayesian broadcast in terms of discriminating access delays experienced by the different classes is illustrated. With $k = 0.7$, a significant spread can only be observed for $S > 0.87$, while with $k = 0.1$ class 4 suffers already at low load from significantly larger access delay values than the other classes. Depending on the amount of contention traffic per class, access delay is not always as evenly spread between neighboring classes as observed here for $k = 0.7$, which, if required, can be corrected by appropriate choice of z_1 and z_2. We conclude that by choosing k and depending on the service limiting the length of allocation cycles we can effectively control the amount of prioritization, as desired, and also tradeoff voice-dropping probability against data-access delay.

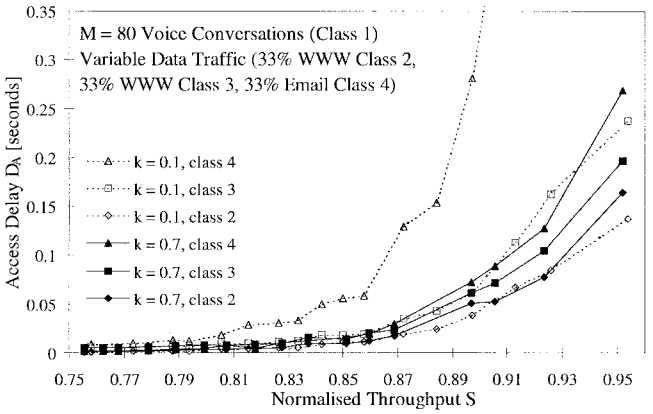

Fig. 15. Access delay performance of data with mixed voice, WWW, and e-mail data traffic.

VI. Controlling MAI with Load-Based Access Control

A. Impact of MAI on Bayesian Broadcast

Bayesian broadcast will in two ways be affected by the introduction of MAI. First, while it was assumed in the derivation of the relevant algorithm that code slots were orthogonal, this is no longer true when considering MAI. It should in theory be possible to account properly for this nonorthogonality in the calculation of update values for the backlog estimation using the first approach mentioned in Section IV-B, but this is likely to be quite cumbersome. Second, as far as feedback is concerned, it is assumed that the BS will consider a code slot carrying an erroneous burst as a collision slot, regardless of whether a code collision has occurred or a single noncollided burst was corrupted due to MAI, although in the latter case v should not be increased. One would therefore expect increased dropping due to both reduced accuracy of the backlog estimation and burst erasure due to MAI. However, if P_{ber} is moderate for up to Q plus a few bursts per time slot, as is the case with parameters and conditions listed in Section III, BB suffers only to a negligible extent from MAI. This is shown in Fig. 16, where P_{drop} is drawn when accounting for MAI for voice-only traffic with Scheme A and compared with the relevant curve without MAI from Fig. 5. Fig. 16 also lists total packet loss ratio P_{loss}, the sum of P_{drop}, and P_{ber} due to MAI (P_{ber} for bursts transmitted in both C and I slots). With the chosen parameters, for $M < 130$, P_{ber} is dominant, hence, the question arises whether P_{ber} can be reduced, even at the expense of increased P_{drop}.

B. Adding Load-Based Access Control

In [12] and [13], we investigated the concept of load-based access control for a PRMA-based protocol on a hybrid TDMA/CDMA air interface to control MAI and thereby reduce P_{loss}. This is here combined with BB according to the following: The basic permission probability p calculated

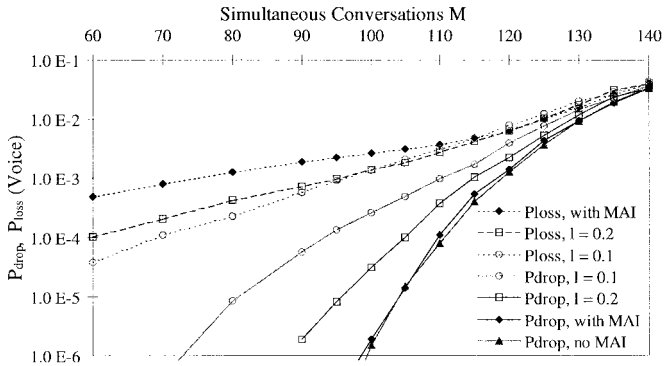

Fig. 16. Voice packet dropping and loss probabilities with and without MAI using Bayesian broadcast, with MAI also combining Bayesian broadcast with load-based access control with $l = 0.1$ and $l = 0.2$.

with BB shall not exceed $p_{\max}$, that is,

$$p = \min\left(p_{\max}, \frac{R[t]}{v}\right) \tag{18}$$

where

$$p_{\max} = \min(1, 2^{R[t]-1} \cdot l) \tag{19}$$

with l subject to optimization. Although (19) was chosen based on experience reported in [12], [13], this is essentially a heuristic approach, but it works well, as shown in Fig. 16. On top of the results previously discussed, P_{drop} and P_{loss} resulting when BB is combined with (18) and (19) are drawn for $l = 0.1$ and 0.2, respectively. While P_{drop} compared to pure BB increases considerably with $l = 0.1$, total P_{loss} is significantly reduced at low load. However, P_{loss} is lower with pure BB for $M \geq 115$. The better compromise seems to be offered by $l = 0.2$, which somewhat increases P_{loss} at low load, but almost meets performance of pure BB at high load. Note that $p = p_{\max}$ in every slot with $l = 0.1$ for $M < 120$, while with $l = 0.2$ for $M < 110$. In other words, access control is essentially load based, and the purpose of BB is to contain occasional backlog excursions at high load.

VII. Conclusion

In a performance assessment for the basic MD PRMA scheme (Scheme A) with TD/CDMA radio interface parameters, the superiority of a system employing Bayesian broadcast control over a system in which access to C slots is controlled by fixed permission probabilities has been shown. With immediate acknowledgment, for voice-only traffic, P_{drop} is as low or lower over the entire range of considered conversations M with broadcast control and stability problems are avoided. However, Bayesian broadcast suffers considerably from delayed acknowledgment and can exhibit "resonant behavior." In such cases, its performance advantage over fixed probabilities dwindles away. These problems are avoided in the MD FRMA scheme where downlink signaling occurs once per frame, but MS's are allowed to contend repeatedly without having to wait for acknowledgment. This scheme has the further advantage of being suitable for TDD operation. With $\alpha_s = 0.448$, when serving $M = 110$ conversations on the 64 time-code slots, P_{drop} is below 10^{-4}. Should a P_{drop} of 10^{-2} be acceptable, more than 130 conversations can be sustained.

Prioritized Bayesian broadcast for MD PRMA with prioritization parameter k was tested with mixed voice, WWW browsing, and e-mail traffic. The ability of this algorithm to discriminate efficiently the access delay of services pertaining to one of four different access classes has been demonstrated. The parameter k not only allows choosing the spread of the access delay values experienced by the different services, but also allows the trade off of voice dropping against data-access delay, in the case of e-mail traffic provided that allocation cycle length is limited. Depending on k and the assignment of services to access classes, at given throughput levels, P_{drop} can even be lower than with homogeneous voice traffic, at the expense obviously of increased access delay for data services. E-mail traffic results in somewhat worse performance than WWW traffic due to the large variance of the message size distribution considered. Using a single prioritization parameter facilitates resource-efficient downlink signaling strategies because individual permission probabilities do not necessarily have to be signalled.

In the system considered here with implicit resource allocation, no specific resource allocation algorithm is required. However, to allow for high-bit-rate data services which require accumulation of multiple codes or time slots and, in particular, to cater for the needs of real-time data services such as video, explicit resource allocation algorithms will have to be looked at which can make use of and enhance the capabilities of prioritized Bayesian broadcast. Unlike the scenario with a large number of low-bit-rate packet services considered here, support of a few high-bit-rate packet data services will likely have an adverse impact on voice traffic.

Finally, we report that if we compare performance of PRMA (on time slots only), MD PRMA (on time-code slots) and a similar scheme with code slots only [22], all with the same frame length and the same number of slots on a perfect collision channel for fair comparison, we observe exactly the same dropping performance, provided that Bayesian broadcast is applied. Obviously, however, the peculiarities of the multiple-access schemes will affect operation and performance of the MAC, as shown in Section VI, where Bayesian broadcast was combined with load-based access control to reduce MAI and thereby also total packet loss experienced.

Acknowledgment

This research was carried out within the UK LINK PCP ACS Project in collaboration with Vodafone, NEC Technologies, and Plextek. The authors thank those members of the companies who were involved in LINK ACS and delivered useful hints for different aspects of this paper. The authors also thank J. Pearson and M. Dell'Anna for reviewing this paper.

References

[1] ETSI SMG2 Tdoc 270/97, "Concept group alpha, wideband direct-sequence CDMA, evaluation document," Pts. I–III, Bad Salzdetfurth, Germany, Oct. 1997.

[2] ACTS AC090, Future Radio Wideband Multiple Access System—FRAMES, "Basic description of multiple access scheme," V. 2.0, Nov. 1996.
[3] ETSI SMG2 UMTS Tdocs 113-117/97, "Concept group delta, wideband CDMA/TDMA, evaluation report," Pts. I–V, Helsinki, Finland, Nov. 1997.
[4] J. Blanz, A. Klein, M. Nasshan, and A. Steil, "Performance of a cellular hybrid C/TDMA mobile radio system applying joint detection and coherent receiver antenna diversity," *IEEE J. Select. Areas Commun.*, vol. 12, no. 4, pp. 568–579, 1994.
[5] ETSI, *ETSI GSM Technical Specifications (01–12 Series)*, European Telecommunications Standards Institute, Sophia-Antipolis, Cedex, France.
[6] A. E. Brand and A. H. Aghvami, "Multidimensional PRMA (MD PRMA)—A versatile medium access strategy for the UMTS mobile to base station channel," in *Proc. PIMRC'97*, Helsinki, Finland, Sept. 1997, pp. 524–528.
[7] D. J. Goodman *et al.*, "Packet reservation multiple access for local wireless communications," *IEEE Trans. Commun.*, vol. 37, no. 8, pp. 885–890, 1989.
[8] D. J. Goodman and S. X. Wei, "Efficiency of packet reservation multiple access," *IEEE Trans. Veh. Technol.*, vol. 40, no. 1, pp. 170–176, 1991.
[9] Y. Li and S. Andresen, "An extended packet reservation multiple access protocol for wireless multimedia communication," in *Proc. PIMRC'94*, Hague, The Netherlands, Sept. 21–23, 1994, pp. 1254–1259.
[10] RACE R2084, "ATDMA system definition, issue 4," Mar. 1996.
[11] J. Dunlop, J. Irvine, D. Robertson, and P. Cosimi, "Performance of a statistically multiplexed access mechanism for a TDMA radio interface," *IEEE Personal Commun. Mag.*, vol. 2, no. 3, pp. 56–64, 1995.
[12] A. E. Brand and A. H. Aghvami, "Performance of a joint CDMA/PRMA protocol for mixed voice/data transmission for third generation mobile communication," *IEEE J. Select. Areas Commun.*, vol. 14, no. 9, pp. 1698–1707, 1996.
[13] ——, "Performance of the joint CDMA/PRMA protocol for voice transmission in a cellular environment," in *Proc. ICC'96*, Dallas, TX, June 1996, pp. 616–620.
[14] R. L. Rivest, "Network control by Bayesian broadcast," *IEEE Trans. Inform. Theory*, vol. IT-33, no. 3, pp. 323–328, 1987.
[15] P. Narasimhan and R. D. Yates, "A new protocol for the integration of voice and data over PRMA," *IEEE J. Select. Areas Commun.*, vol. 14, no. 4, pp. 623–631, 1996.
[16] ETSI, *Selection Procedures for the Choice of Radio Transmission Technologies of the UMTS (UMTS 30.03 Version 3.1.0)*, European Telecommunications Standards Institute, Sophia-Antipolis, Cedex, France, November 1997.
[17] C. Fresco Diez, A. E. Brand, and A. H. Aghvami, "Prioritised random access for GPRS with pseudo Bayesian broadcast control, exponential backoff and stack based schemes," in *Proc. ICT'98*, Chalkidiki, Greece, June 1998, pp. 24–28.
[18] ETSI, *ETSI GSM 03.64 (Phase 2+), Overall Description of the General Packet Radio Service (GPRS) Radio Interface, Stage 2*, Version 5.0.0, July 1997.
[19] F. Delli Priscoli, "Adaptive parameter computation in a PRMA, TDD based medium access control for ATM wireless networks," in *Proc. Globecom'96*, London, U.K., Nov. 1996, pp. 1779–1783.
[20] G. A. Cunningham, "Delay versus throughput comparisons for stabilized slotted Aloha," *IEEE Trans. Commun.*, vol. 38, no. 11, pp. 1932–1934, 1990.
[21] S. Lin, *An Introduction to Error-Correcting Codes*. Englewood Cliffs, NJ, Prentice-Hall, 1970.
[22] L. Tan and Q. T. Zhang, "A reservation random-access protocol for voice/data integrated spread-spectrum multiple access systems," *IEEE J. Select. Areas Commun.*, vol. 14, no. 9, pp. 1717–1727, 1996.

Alex E. Brand was born in 1969 in Zürich, Switzerland. He received the Diploma in electrical engineering from ETH (Swiss Federal Institute of Technology), Zürich, in 1995. Since 1995, he has been working toward the Ph.D. degree at King's College, London, U.K.

He is a Research Associate at King's College, where he worked for three years on the Link ACS Project together with Vodafone, NEC Technologies, and Plextek on enhanced second- and third-generation mobile communication systems. He has attended several ETSI SMG2 GPRS *ad hoc* meetings, held presentations on the ETSI SMG2 UMTS workshops, and also lectured on HSCSD and GPRS in industrial short courses.

A. Hamid Aghvami (M'87–SM'91) received the M.Sc. and Ph.D. degrees from King's College, London, U.K., in 1978 and 1981, respectively.

In April 1981, he joined the Department of Electronic and Electrical Engineering, King's College, as a Post-Doctoral Research Associate. He worked on digital communications and microwave techniques projects sponsored by EPSRC. In 1989, he was promoted to Reader and then in 1992 to Professor in Telecommunications Engineering. He is presently the Director of the Centre for Telecommunications Research at King's College. He carries out consulting work on digital radio communication systems for both British and international companies. He has published over 200 technical papers and lectures on digital radio communications including GSM 900/DCS 1800 worldwide. He leads an active research team working on numerous mobile and personal communication system projects for third-generation systems.

Dr. Aghvami is a Distinguished Lecturer of the IEEE Communications Society. He has been a Member, Chairman, and Vice Chairman of the technical program and organizing committees of a large number of international conferences. He is also the founder of the International Conference on Personal, Indoor, and Mobile Radio Communications. He is a Fellow Member of the IEE.