Low-Complexity Medium Access Control Protocols for QoS Support in Third-Generation Radio Access Networks

Timothy O'Farrell, *Member, IEEE,* and Peter Omiyi

***Abstract*—One approach to maximizing the efficiency of medium access control (MAC) on the uplink in a future wideband code-division multiple-access (WCDMA)-based third-generation radio access network, and hence maximize spectral efficiency, is to employ a low-complexity distributed scheduling control approach. The maximization of spectral efficiency in third-generation radio access networks is complicated by the need to provide bandwidth-on-demand to diverse services characterized by diverse quality of service (QoS) requirements in an interference limited environment. However, the ability to exploit the full potential of resource allocation algorithms in third-generation radio access networks has been limited by the absence of a metric that captures the two-dimensional radio resource requirement, in terms of power and bandwidth, in the third-generation radio access network environment, where different users may have different signal-to-interference ratio requirements. This paper presents a novel resource metric as a solution to this fundamental problem. Also, a novel deadline-driven backoff procedure has been presented as the backoff scheme of the proposed distributed scheduling MAC protocols to enable the efficient support of services with QoS imposed delay constraints without the need for centralized scheduling. The main conclusion is that low-complexity distributed scheduling control strategies using overload avoidance/overload detection can be designed using the proposed resource metric to give near optimal performance and thus maintain a high spectral efficiency in third-generation radio access networks and that importantly overload detection is superior to overload avoidance.**

***Index Terms*—Medium access control (MAC), quality of service (QoS), radio resource management, third-generation mobile systems, WCDMA.**

I. Introduction

THE EMERGING wideband code-division multiple-access (WCDMA)-based third generation of radio access networks support multiple services differentiated by their quality of service (QoS) constraints, that is, delay and signal-to-interference ratio (SIR) requirements. Maximizing the spectral efficiency in third-generation radio access networks is commercially important to both the service and network providers but presents a greater challenge than in second-generation systems because of the need to satisfy users' QoS constraints and to provide bandwidth on demand. One approach to maximizing spectral efficiency, on the uplink, is to maximize the efficiency of the medium access control (MAC) mechanism, which is the approach considered in this paper. The more popular MAC approach, when it comes to QoS support, relies on centralized scheduling control [1], [2], where a central MAC entity is responsible for scheduling user-equipment transmissions. This approach offers the potential of delivering near-optimal performance [1] but is complicated by the need to maintain fast/reliable signaling of a significant amount of control information between the user equipment and the central MAC entity. A less complex approach is to use distributed scheduling control, where individual scheduling decisions are made by individual user equipment rather than by a central MAC entity. This implicitly simplifies the nature of the protocol signaling and could potentially reduce signaling overhead.

An issue that is particularly relevant to QoS support in third-generation radio access networks is the efficient allocation of radio resources between services with different SIR requirements. The approaches presented in [3] and [4] are exemplary of the research focus in this area, where the MAC protocol decides whether or not to permit a MAC protocol data unit (MPDU) transmission based on satisfying a *power control feasibility* condition. This approach, though useful, represents a marked departure from conventional resource allocation techniques which require the use of a metric that measures resources. However, such a metric has not previously been available to aid the proper design of resource allocation algorithms in an interference limited environment, as is the case in third-generation radio access networks. This is because there are two dimensions of radio resources, namely power and bandwidth, and resource allocation in either domain alone is insufficient on the interference limited uplink. This paper addresses this fundamental problem by first developing a metric, denoted as *normalized power*, that captures these two dimensions of resource. Subsequently, the resource metric is put to use in order to design, develop, and evaluate new MAC algorithms which maximize the efficiency of resource allocation within their operating constraints. The performance of the new MAC algorithms is evaluated in both single-cell and mobile multicell environments.

The proposed protocols, based on distributed scheduling control strategies, are referred to as single-threshold overload signal spread spectrum (ST-OSSS), multiple-threshold overload signal spread spectrum (MT-OSSS), and overload signal spread spectrum with overload detection (OSSS/OD), where the basic ac-

Manuscript received November 4, 2003; revised December 23, 2003; accepted November 28, 2003. The editor coordinating the review of this paper and approving it for publication is Z. J. Haas.

T. O'Farrell is with the School of Electronic and Electrical Engineering, University of Leeds, Leeds, West Yorkshire LS2 9JT, U.K. (e-mail: T.O'Farrell@leeds.ac.uk).

P. Omiyi is with the International University of Bremen, Bremen 28759, Germany (e-mail: p.omiyi@iu-bremen.de).

Digital Object Identifier 10.1109/TWC.2004.842944

cess procedure of the former two are based on overload avoidance [5], [6], while that of the latter is based on overload detection [7].

Now, it is a well-established fact that a *backoff procedure* is required to achieve protocol stability when using distributed scheduling control (and the proposed protocols are no exception) by introducing controlled random delays. In [8], the backoff procedure is also used to differentiate between services with different delay requirements. In this way, a major problem when using distributed scheduling control, which is the QoS support for services with delay constraints in the absence of a centralized scheduler, is addressed in [8]. However, a potential drawback of this innovative solution is that each of the service-dependent access probabilities needs to be optimized. Therefore, the number of parameters that needs optimization increases as the number of supported services increases. In this paper, a deadline-driven backoff (DDB) procedure is proposed as the backoff strategy for the distributed scheduling MAC protocols. Like the backoff procedure in [8], DDB combines the functions of maintaining protocol stability and providing QoS support for services with delay constraints. However, unlike the backoff procedure in [8], the proposed DDB strategy (implicitly) defines the MPDU access probability, not by the service it belongs to, but by its *delivery deadline*. DDB ensures that an MPDU has a higher access probability than any other MPDU with a later delivery deadline. Hence, unlike in [8], the number of parameters that need optimization does not grow with the number of supported services, which is desirable.

Finally, a new scheduling algorithm, namely the earliest deadline first parallel-link scheduler (EDF-PLS) is presented as a performance benchmark with respect to the objective of maximizing spectral efficiency, because it has been shown in [9] to offer near optimal throughput performance. In addition, the EDF-PLS algorithm represents the ideal case for a centralized scheduling scheme, with the entire buffer information of every user equipment being available to the central scheduling MAC entity at all times and without signaling overhead.

The remainder of this paper is organized as follows. Section II presents the normalized power metric and its application in measuring resource and capacity in a third-generation radio access network. Section III presents the proposed distributed scheduling control MAC protocols, while Section IV presents the benchmark EDF-PLS scheduling policy. Section V presents a system description, and Section VI presents a method of estimating system capacity using the proposed resource metric. Section VII presents a simulation experiment and results that compare the performance of the proposed overload avoidance and overload detection protocols when supporting a mixture of (real-time) voice and (nonreal-time) high-speed data services. Results are presented for both single and mobile multicell environments. Finally, Section VIII presents the conclusions.

II. Quantifying Resource and Capacity

The *normalized power* used or required by a transmission is simply the ratio of received power contributed or to be contributed by the transmission to the total received power. As will become evident in the subsequent discussion, normalized power adequately quantifies resources in an interference limited system, where the resource requirement of a transmission at a receiver is referred to as its normalized power requirement. With reference to the uplink of a radio access network, the *intracell normalized power requirement* of a transmission represents its resource requirement at its serving base station. The *intercell normalized power requirement* of a transmission at a nonserving base station represents its resource requirement at this base station. The normalized power capacity at a base station represents the maximum amount of resources that can be used at that base station. In order to derive mathematical expressions for the intracell normalized power requirement, intercell normalized power requirement, and normalized power capacity, consider a third-generation radio access network with U pieces of user equipment communicating with B base stations on the uplink, where each user equipment is uniquely indexed by a positive integer j in the interval $[1, U]$ and the base station serving user equipment j is indexed by a positive integer b_j in the interval $[1, B]$. Then, it is necessary to establish the condition that guarantees that all transmissions served by all base stations are received with acceptable SIR, beginning with the expression in (1), which is always true. In (1), the symbols $P_{\text{noise}}(b_j), P_{\text{tot}}(b_j), P_i(b_j)$, and $h_i(b_j)$ represent the background noise power at base station b_j, the total received power at base station b_j, the received power at base station b_j from a transmission of user equipment i, and the link gain from user equipment i to base station b_j, respectively

$$\sum_{i=1}^{U} P_i(b_j) = \sum_{i=1}^{U} \frac{h_i(b_j)}{h_i(b_i)} P_i(b_i) + P_{\text{noise}}(b_j) = P_{\text{tot}}(b_j) \quad \text{for } j = 1 \ldots U. \quad (1)$$

Let the symbol $P_{\text{tot}}^{(\max)}$ represent the upper limit on total received power at a base station receiver that guarantees all transmissions are received with acceptable SIR. Then, a necessary condition to ensure that all transmissions served by all base stations are received with acceptable SIR is that $P_{\text{tot}}(b_j) \leq P_{\text{tot}}^{(\max)}$ for $j = 1, \ldots, U$. By substituting the inequality into the left-hand side (LHS) of (1), we obtain

$$\sum_{i=1}^{U} \frac{h_i(b_j)}{h_i(b_i)} P_i(b_i) + P_{\text{noise}}(b_j) \leq p_{\text{tot}}^{(\max)} \quad \text{for } j = 1 \ldots U. \quad (2)$$

The upper limit $P_{\text{tot}}^{(\max)}$ implies that there is an upper limit on the received power necessary for a transmission of user equipment i not to be received in outage at base station b_i and this is represented by the symbol $P_i^{(\max)}(b_i)$. Now, it is clear that when the condition $P_i^{(\max)}(b_i) \leq P_i(b_i)$, for $i = 1, \ldots, U$, is satisfied, then all transmissions are guaranteed to be received with acceptable SIR, provided also that $P_{\text{tot}}(b_i) \leq P_{\text{tot}}^{(\max)}$ for $i = 1, \ldots, U$. Satisfying the inequality in (2) already ensures that the latter condition is met; therefore, substituting the inequality $P_i^{(\max)}(b_i) \leq P_i(b_i)$ for $i = 1, \ldots, U$ in the LHS of (2) and rearranging gives (3), which is a sufficient condition

to guarantee that all transmissions are received with acceptable SIR

$$\sum_{i=1}^{U} \frac{h_i(b_j)}{h_i(b_i)} \times \frac{P_i^{(\max)}(b_i)}{p_{\text{tot}}^{(\max)}} \leq 1 - \frac{P_{\text{noise}}(b_j)}{p_{\text{tot}}^{(\max)}} \quad \text{for } j = 1 \ldots U. \tag{3}$$

Now, from the definition of $P_i^{(\max)}(b_i)$, the relationship in (4) must hold, where γ_i is the SIR target[1] of the transmission of user equipment i at base station b_i. Then, substituting (4) into the LHS of (3) gives

$$\frac{P_i^{(\max)}(b_i)}{P_{\text{tot}}^{(\max)}} = \frac{1}{1 + 1/\gamma_i} \tag{4}$$

$$\sum_{i=1}^{U} \frac{h_i(b_j)}{h_i(b_i)} \times \frac{1}{1 + 1/\gamma_i} \leq 1 - \frac{P_{\text{noise}}(b_j)}{p_{\text{tot}}^{(\max)}} \quad \text{for } j = 1 \ldots U. \tag{5}$$

Let the symbol $z_i(b_j)$ represent the summand on the LHS of (5), then $z_i(b_j)$ is the normalized power requirement of a transmission of user equipment i at base station b_j. If i is equal to j for any summand on the LHS of (5), then the summand $z_j(b_j)$ represents the intracell normalized power requirement of a transmission of user equipment j at base station b_j. A close observation of (5) shows that $z_j(b_j)$ can be expressed as in (6). If i is not equal to j for any summand on the LHS of (5), then the summand represents the intercell normalized power requirement of a transmission of user equipment i at base station b_j, which is represented by $z_i(b_j)|_{b_j \neq b_i}$. Substituting the LHS of (6) into (5) shows that $z_i(b_j)|_{b_j \neq b_i}$ is given by

$$z_j(b_j) = \frac{1}{1 + 1/\gamma_j} \tag{6}$$

$$z_i(b_j)|_{b_j \neq b_i} = \frac{h_i(b_j)}{h_i(b_i)} z_i(b_i). \tag{7}$$

Let the symbol $Z_{(\max)}(b_j)$ represent the term on the RHS of the inequality in (5). Now, the inequality in (5) states that the sum of all transmission resource requirements at any instant, at every base station that is serving at least one piece of user equipment, must not exceed $Z_{(\max)}(b_j)$, in order to guarantee that all transmissions are received with acceptable SIR. Therefore, $Z_{(\max)}(b_j)$ represents the normalized power capacity at base station b_j. Note from the right-hand side (RHS) of (5), the larger the value of $P_{\text{tot}}^{(\max)}$, the higher the normalized power capacity, and as $P_{\text{tot}}^{(\max)}$ tends to infinity the normalized power capacity tends to unity. However, $P_{\text{tot}}^{(\max)}$ is limited by transmit power constraints, where $P_{\text{tot}}^{(\max)}$ is limited by the most vulnerable type of transmission, that is, one with the worst combination of a low peak transmit power, a high SIR requirement, and being far from its intended base station receiver. This suggests that the relationship in (8) is true for $P_{\text{tot}}^{(\max)}$, where the symbol $\min_s [x_s]_{s=1}^{N_S}$ represents the smallest value of the set $\{x_s : s = 1, \ldots, N_S\}$. In (8), $Z_s^{(\max)}$ and $P_{T,s}^{(\max)}$ are the maximum intracell normalized power requirement and the peak transmit power for the transmission of service s, respectively. Also, N_S, H_{edge}, and H_{fade} are the total number of services, the link gain to the edge of the coverage area, and shadow fading margin, respectively. In other words, the inequality in (8) states that $P_{\text{tot}}^{(\max)}$ must be chosen to ensure that the service, with the smallest achievable received power at the edge of the coverage area, achieves a normalized power equal to its normalized power requirement. Therefore, $P_{\text{tot}}^{(\max)}$ can be set equal to the RHS of (8), as long as this value of total received power is not so high as to fall outside the dynamic range of the base station receiver

$$p_{\text{tot}}^{(\max)} \leq H_{\text{edge}} H_{\text{fade}} \min_s \left[P_{T,s}^{(\max)} \Big/ z_s^{(\max)} \right]_{s=1}^{N_S}. \tag{8}$$

To conclude this section, we note from the RHS of (5) and from (8) that the smaller H_{edge}, the smaller H_{fade}, and the smaller $P_{T,s}^{(\max)}$, then the smaller the normalized power capacity. Also, (6) shows that the normalized power requirement increases with increases in the SIR requirement, where the SIR requirement is given by the expression $(R_B)/(W) \times (E_B)/(I_o)$, where R_B, W, and E_b/I_o are the transmission bit rate, channel chip rate in WCDMA, and the average bit energy to interference spectral density or link quality requirement, respectively. The first point implies that the greater the cell size, the greater the shadow fading variation and the smaller the peak transmitter power levels, then the smaller the capacity, while the second point implies that the higher the link quality requirement (the lower the bit-error rate) and/or the faster the data rate the more the resource consumed by the user. These patterns of behavior of normalized power requirement and normalized power capacity are as expected of resource requirement and capacity in an interference limited system.

III. Novel Distributed Scheduling MAC Protocols

The three new distributed scheduling MAC protocols presented in this paper, each comprised of a basic access procedure and a backoff procedure, are distinguished by their basic access procedures. These basic access procedures are described next, followed by a description of DDB, which is the backoff procedure that is proposed for all three MAC protocols.

A. Basic Access Procedure of ST-OSSS

Each slot begins with a short access window of duration τ_{aw} which is used to resolve contention for the remainder of the slot. Now, at the start of a slot, if user equipment has an MPDU that is "ready" for transmission, the user equipment begins listening on the downlink and schedules to begin transmission at a random instant, according to a uniform distribution, within the access window of that slot. If at some time t within the access window and timed from its start, base station b_j detects that its channel load $w_{\text{load}}(t, b_j)$ exceeds the *overload-avoidance threshold* Z_{OL}, then it begins broadcasting an overload signal at this instant t to all "applicable" user equipment. The "applicable" user equipment includes those that significantly interfere with the base station, namely all user equipment that communicates with it, including those that are in its soft-handover region. All "applicable" user equipment that can hear the

[1]Note that the target SIR takes into account the power-control imperfections, propagation-conditions, link-quality (bit-error rate or E_b/I_o) requirements, channel-coding gain, processing gain, and the other diversity techniques employed at the receiver.

overload signal before they are scheduled to begin their MPDU transmissions within the access window are barred from transmitting in the current slot. Therefore, $Z_{\rm OL}$ must be chosen to minimize the frequency of channel overloads without compromising throughput and its value cannot exceed the normalized power capacity $Z_{\max}(b_j)$ of the base station receiver. Note that there is a *protocol response time* or "latency" τ, which is the shortest time necessary for all "applicable" user equipment to be aware of any changes in channel load.

The *channel load* $w_{\rm load}(t,b_j)$ is defined as the sum of the normalized power requirements of MPDU transmissions that have been received by base station b_j at time t and is given by (9). The term $N_{\rm AW}(t)$ in (9) is a random variable representing the number of transmissions that have begun by time t within the access window

$$w_{\rm load}(t,b_j)=\sum_{i=1}^{N_{\rm AW}(t)}\frac{h_i(b_j)}{h_i(b_i)}z_i(b_i). \tag{9}$$

It is worth noting that the values of the parameters necessary for base station b_j to compute are $w_{\rm load}(t,b_j)$ available from the network. The outer loop power control algorithm running on the network determines the target SIR values of transmissions belonging to every user equipment, while user equipment periodically measures and reports to the network the link gains to neighboring base stations. Therefore, a base station receiver monitors its channel load during the access window by first "triggering" to individual transmissions and then obtaining from the network the necessary parameters to compute the normalized power requirements for these transmissions. The base station computes the intracell normalized power requirements for all transmissions belonging to user equipment that are under its power control, which includes some of the user equipment in its soft-handover region. It computes intercell normalized power requirements for all transmissions belonging to user equipment outside its power control but within its soft-handover region. However, the base station ignores the intercell normalized power requirements of all other transmissions, which belong to user equipment that is neither under its power control nor within its soft-handover region. These transmissions typically contribute less interference to the base station than the users under its power control or within its soft-handover region. Finally, the base station sums up the computed intracell and intercell normalized power requirements to give the channel load.

B. Basic Access Procedure of MT-OSSS

There is a distinct *characteristic overload-avoidance threshold* value $Z_{\rm OL}(z)$ that best serves as an overload-avoidance threshold for transmissions with normalized power requirement equal to z. This is because it can be shown that without any loss of generality, $Z_{\rm OL}(z)$ must satisfy the relation in (10) for base station b_j

$$Z_{\rm OL}(z)\le Z_{\max}(b_j)-z. \tag{10}$$

If the relation in (10) is satisfied and the channel load is less than $Z_{\rm OL}(z)$ at any instant in the access window, then it is guaranteed that the subsequent access of a transmission, with a normalized power requirement equal to z, will not cause the channel load to exceed the normalized power capacity. But, the channel load may exceed $Z_{\rm OL}(z)$. Therefore, using an overload-avoidance threshold that meets (10) ensures that when transmissions with normalized power requirement equal to z are supported, a warning overload signal would always be broadcast before the channel load exceeds the normalized power capacity, in order to deter further transmissions. However, if the condition in (10) is not satisfied, then there is no such guarantee since the possibility exists that the channel load may exceed both the normalized power capacity and $Z_{\rm OL}(z)$, simultaneously.

The existence of characteristic overload-avoidance threshold values suggests that the use of a single threshold for overload avoidance, as in ST-OSSS, is not the best policy when there is a plurality of potential normalized power requirement values for transmissions. This is because in order to minimize the frequency of outage, the characteristic overload-avoidance threshold with the lowest value has to be selected and used as the single ST-OSSS threshold. This threshold corresponds to the characteristic overload-avoidance threshold for the highest normalized power requirement value and is the most restrictive to access for transmissions with a smaller normalized power requirement, resulting in higher delays than necessary for such transmissions. Therefore, the multiple threshold approach of MT-OSSS has been proposed to improve on the performance of ST-OSSS. The basic access procedure for MT-OSSS is similar to that for ST-OSSS, with user-equipment scheduling transmissions in the access window in the same way. However, MT-OSSS uses a set of overload-avoidance thresholds that is equivalent to the set of characteristic overload-avoidance thresholds $\{Z_{\rm OL}(z_j)\}$. Whenever the channel load $w_{\rm load}(t,b_j)$ exceeds an overload-avoidance threshold $Z_{\rm OL}(z_j)$, an overload signal is sent only to "applicable" user equipment that currently uses a transmission normalized power requirement of z_j. More processing is inevitably required for MT-OSSS than ST-OSSS, since the former compares channel load against more thresholds than the latter. Also, while the downlink signaling information broadcast by ST-OSSS requires only a single information field for all service types, a similar broadcast by MT-OSSS must include a separate field for each service type with a distinct transmission normalized power requirement, where each user equipment responds only to the field corresponding to its transmission normalized power requirement. This implies that the signaling overhead of MT-OSSS exceeds that of ST-OSSS.

C. Basic Access Procedure of OSSS/OD

The basic access procedure for OSSS/OD is similar to that for ST-OSSS, with user-equipment scheduling transmissions in the access window in the same way, except in this case the access window comprises of a number of minislots. The duration of each minislot must be no less than the protocol response time or latency. If in any minislot m within the access window and counted from its start base station b_j detects that its channel load $w_{\rm load}(m,b_j)$ exceeds the *overload-detection threshold*, then it begins broadcasting an overload signal in this minislot m to all "applicable" user equipment. Then, all "applicable" user equipment, which can hear the overload signal and begin MPDU transmission in the minislot m, aborts transmissions in the cur-

1. **If** T_{boff} equals zero, then DDB attempts transmission as follows.
    1.1. **If** the load in the access window of the preceding slot exceeds a protocol threshold, then the current DDB transmission attempt is considered to have 'failed' and DDB behave as follows.
        1.1.1. Determine the back-off window size T_{win} using the backoff *window size selection rule* $T_{win} = \min\left(2^n \times T_{win}^{(\min)}, T_{win}^{(\max)}\right)$, where n is the number of DDB transmission attempts to date, $T_{win}^{(\min)}$ is the minimum backoff window size and $T_{win}^{(\max)}$ is the maximum.

            Note: $T_{win}^{(\max)}$ equals the delivery deadline T_D of the MPDU less the slot duration T_s (MPDU transmission time) and less the current time t, that is $T_{win}^{(\max)} = T_D - T_s - t$.
        1.1.2. Select a random back-off time T_{boff} in the interval [0, T_{win}].
        1.1.3. The number of DDB transmission attempts to date n is incremented by one.
        1.1.4. **If** T_{boff} equals zero, then DDB delivers the MPDU to the basic access procedure of the MAC.
            1.1.4.1. **If** the MPDU is delayed by the basic access procedure of the MAC, then it remains queued by the user-equipment and, hence, reattempts transmission using the DDB process, starting from step 1 above. (refer to sections III.A, III.B and III.C on the behaviour of the proposed MAC basic access procedures)
    1.2. **Else**, DDB delivers the MPDU to the basic access procedure of the MAC for transmission.
            1.2.1.1. **If** the MPDU is delayed by the basic access procedure of the MAC, then it remains queued by the user-equipment and, hence, reattempts transmission using the above DDB process, starting from step 1 above.
2. **Else**, decrement T_{boff} by one slot.

Fig. 1. DDB algorithm.

rent slot. Also barred from transmitting in the current slot are all "applicable" user equipment that can hear the overload signal and are scheduled to begin MPDU transmission in any minislot with index greater than m. Note that overloads only occur in the access window, since they are detected and mitigated before the end of the access window. So, for a short access window, the impact of the overload is small.

OSSS/OD requires a single overload-detection threshold. Since overloads only occur when the channel load exceeds normalized power capacity, the overload-detection threshold by definition must equal normalized power capacity. The use of a single threshold represents an advantage of OSSS/OD over MT-OSSS. Also, OSSS/OD has the advantage over both overload-avoidance proposals in that the overload-detection threshold is a constant and does not need to be optimized, unlike the thresholds of the latter.

D. DDB

In the context of this paper, "backoff" refers broadly to the mechanism by which user equipment selects the delay before an MPDU attempts transmission with the basic access procedure of the protocol. With the proposed DDB procedure, the later the delivery deadline of an MPDU and/or the more the congestion, the longer the delay on average selected by DDB. In this way, DDB adaptively shapes the offered load statistics with the objective of minimizing both the congestion and the likelihood of an MPDU being late, thus maximizing the useful throughput.

Now, associated with every MPDU is a backoff time T_{boff} measured in slots. When user equipment has an MPDU to transmit, in each slot it behaves according to the algorithm in Fig. 1. Note that DDB continues its attempts to transmit an MPDU via the basic access procedure of the MAC until the MPDU is transmitted or late. Finally, the *window size selection rule* used is the *exponential rule*, made popular by its simplicity and its application in the binary exponential backoff algorithm [11]. However, other rules could be employed instead, such as the linear or μ-law rules, for instance.

The description of DDB in Fig. 1 shows that backoff is triggered by overload, which in turn depends on the amount of normalized power resource being used. Now, as the amount of normalized power resource being used temporarily increases from some equilibrium value, the overload frequency also tends to increase, which forces DDB to increase the MPDU access delay and thus reduce the amount of normalized power resource being used (and the overload frequency) back down toward the equilibrium point. Also, if the amount of resource being used temporarily falls from some equilibrium value, the overload frequency also tends to fall, which forces DDB to reduce the MPDU access delay and thus increase the amount of normalized power resource being used (and the overload frequency) toward the equilibrium point. Therefore, DDB effectively represents a nonlinear control system using negative feedback with respect to the amount of normalized power resource being used.

IV. EDF-PLS

In [9], it is shown that the EDF scheduling policy is near optimal in minimizing the frequency of MPDU lateness when

1. Let $w_L(0,b_j)=0$
2. Start with n=1
3. **While** $n \le N_L$
   - 3.1.Compute
     $$w_L(n,b_j)=w_L(n-1,b_j)+z_n(b_n), \text{ where } b_j=b_n$$
   - 3.2.**If**
     $$w_L(n,b_j) \le Z_{max}(b_j), \text{ for } j=1\ldots U$$
     - 3.2.1. Copy n-th entry from the request list to the transmit list
     - 3.2.2. Then n=n+1
   - 3.3.**Else** (skip n-th entry for another entry for which step 3.2 may be true)
     - 3.3.1. Compute
       $$w_L(n,b_j)=w_L(n-1,b_j), \text{ where } b_j=b_n$$
     - 3.3.2. Then set n=n+1

Fig. 2. EDF-PLS algorithm.

scheduling MPDU transmissions with heterogeneous normalized power requirements. In each slot, and at each base station, the EDF-PLS first organizes a *request list* of all MPDUs requesting access. The request list is organized in a nondecreasing order of the delivery deadlines; that is, the request with the earliest deadline is at the top of the list. Requests that would exceed their delivery deadlines, if transmitted in the current slot, are not included in the request list and are dropped.

Let the positive integer j index the jth entry in the request list from the top and also index the user equipment associated with this entry. Also, let the integer b_j index the base station serving the user equipment associated with the jth entry in the request list, while $w_L(n,b_j)$ represents the sum of normalized power requirements from a subset of the first n entries of the request list at base station b_j, where N_L is the total number of entries in the list at the base station. Then, the EDF-PLS algorithm generates a *transmit list* at base station b_j according to the iterative algorithm in Fig. 2.

All MPDUs with entries in the transmit list are allowed to transmit in the current slot. The EDF-PLS algorithm adheres to the basic EDF scheduling principle [14] except when it is necessary to schedule an MPDU transmission with a smaller normalized power requirement than another MPDU with an earlier delivery deadline. This occurs when the normalized power requirement of the latter is too high to be supported at the base stations. In the limit, when all MPDUs have equal normalized power requirements, the algorithm reduces to the basic EDF scheduling policy.

V. System Description

In order to evaluate MAC protocol efficiency, in the first instance an isolated third-generation radio access network picocell in an indoor office operating scenario, supporting a mixture of voice telephony and nonreal-time high-speed data (NRT-HSD) on the uplink, is modeled. A WCDMA system is considered [13] with services supported on the uplink using dedicated channels between the user equipment and the base station, where each user equipment is allocated at least one dedicated physical data channel and one dedicated physical control channel. Associated with each user equipment is an unique scrambling code. User equipment is considered stationary in the indoor environment. The remainder of this section describes the traffic models and QoS constraints applied to voice and NRT-HSD in this paper.

A. Traffic Models

1) Voice Traffic Model: This model consists of exponentially distributed ON and OFF periods. During ON periods source data is generated at a constant bit rate, while during OFF periods no data is generated. The mean duration for both ON and OFF periods, denoted as $T_{\rm ON}$ and $T_{\rm OFF}$, respectively, equals 3 s, as proposed in [12] for voice telephony. The minimum acceptable duty cycle $\delta_{\rm voice}^{(\min)}$ for a voice service, when using this model, is defined in (11), where T and $T_{\rm DC}^{(\rm voice)}$ denote the slot duration and maximum acceptable MPDU delivery delay for a voice service, respectively

$$\delta_{\rm voice}^{(\min)}=\frac{T_{\rm ON}}{T_{\rm ON}+T_{\rm OFF}+T_{\rm DC}^{(\rm voice)}-T}. \tag{11}$$

2) NRT-HSD Traffic Model: The NRT-HSD traffic model is the WWW browsing/file transfer model presented in [12]. This model considers the generation of layer 3 packets, and each must then be segmented into one or more MPDUs depending on packet size and transmission bit rate (MPDU size). Each session or connection comprises a number of *packet calls* $N_{\rm pc}$ that is geometrically distributed with mean $\mu_{\rm npc}$. The number of packets in a packet call N_d is also geometrically distributed with mean $\mu_{\rm nd}$. The time interval between two consecutive packet arrival epochs within a packet call is D_d (seconds), where D_d+1 is geometrically distributed with mean $\mu_{\rm dd}+1$ (seconds), where time is measured in discrete units of slot duration. The reading time $D_{\rm pc}$ is the time difference between the instant the last MPDU of the last packet of a packet call completes transmission and the instant of arrival of the first MPDU of the first packet of the next packet call. The distribution of $D_{\rm pc}$ is a geometric distribution, in slots, with mean $\mu_{\rm dpc}$. Each layer 3 packet is of length L_p in bytes, with mean $\mu_{\rm lp}$, which is assumed to be distributed according to a truncated Pareto distribution [12], where the probability density function (pdf) of a nontruncated Pareto distribution is given by (12). If X is a nontruncated Pareto distributed variable and c is the maximum layer 3 packet size, then $L_p=\min(X,c)$ where $e^\beta=81.5$ bytes, $\lambda=1.1$, and $c=67$ Kbytes [12]

$$f(x)=\begin{cases}\dfrac{\lambda e^{\lambda\beta}}{x^{\lambda+1}}, & \text{for } x\ge e^\beta\\ 0, & \text{for } x<e^\beta.\end{cases} \tag{12}$$

Given these parameters, the minimum acceptable duty cycle $\delta_{\rm hsd}^{(\min)}$ for an NRT-HSD service, when using this model, is defined in (13), where $L_{\rm pad}$ denotes the mean number of padding bits which are added, when necessary, to complete the MPDU at the end of each packet call. The symbols T and $T_{\rm DC}^{(\rm hsd)}$ in (13) represent the slot duration and maximum acceptable MPDU delivery delay for a NRT-HSD service, respectively

$$\delta_{\rm hsd}^{(\min)}=\frac{\dfrac{\mu_{\rm nd}\times(\mu_{\rm lp}+L_{\rm pad})\times 8}{R_B}}{\dfrac{\mu_{\rm nd}\times(\mu_{\rm lp}+L_{\rm pad})\times 8}{R_B}+\mu_{\rm dpc}+T_{\rm DC}^{(\rm hsd)}-T}. \tag{13}$$

For this experiment, $\mu_{\rm dd} = 0, \mu_{\rm nd} = 15, \mu_{\rm dpc} = 12$ seconds, and so $\mu_{\rm npc}$ is infinite since a very long session is considered in order to obtain steady-state statistics.

B. Performance Measures

The four main performance measures that are used to evaluate MAC protocol performance, in this paper, are the MPDU success probability $P_{\rm succ}$, mean throughput in bits per second, mean delay per bit, and the capacity operating point. The former is the probability that an MPDU is neither late (dropped) nor received in outage. The mean throughput S at the base station is defined as the rate at which information is received without outage (in bits per second) at the base station, and the mean delay per bit is the mean time interval from when a bit is ready for transmission by the MAC to the time when it has been received without outage. We define the QoS requirement of a service as the set of performance measure constraints which, if met, implies that the perceived QoS of a user of a service is satisfactory. In this paper, a data loss constraint, defined as $P_{\rm succ} \geq P_{\rm limit}^{\rm (voice)}$, is specified for the voice telephony service, while both a loss constraint ($P_{\rm succ} \geq P_{\rm limit}^{\rm (hsd)}$) and a mean delay constraint, defined as $D \leq D_{\rm limit}^{\rm (hsd)}$, are specified for the NRT-HSD service. $P_{\rm limit}^{\rm (voice)}$, $P_{\rm limit}^{\rm (hsd)}$, and $D_{\rm limit}^{\rm (hsd)}$ are limits on the achieved values of MPDU success probability and mean NRT-HSD delay, respectively, that result in acceptable perceived QoS. The capacity operating point of a MAC protocol is defined here as the maximum number of connections using service r that can be supported by the protocol in a mix containing N_S service classes, given that M_s connections use each of one of the other services s, where s and r are integers in the interval $[1, N_S]$. The capacity operating point of any scheme is dependent on the exact nature of the service mixture.

VI. Estimating the Capacity Operating Point

One application of the normalized power metric is in estimating the capacity operating point by calculating a "fluid capacity" estimate. The fluid capacity serves as an upper bound on the maximum number of connections $M_r^{\rm (max)}$ using service r that can be supported, on the uplink of a third-generation radio access network, in a mix containing N_S service classes, given that M_s connections use each of one of the other services s, where s and r are integers in the interval $[1, N_S]$. Also given are the minimum acceptable duty cycle, minimum acceptable MPDU success probability $P_{\rm limit}^{\rm (s)}$, and maximum normalized power requirement when transmitting $Z_s^{\rm (max)}$, respectively, of a connection using service s, for all values of s. Note that $\delta_s^{\rm (min)}$ accounts for the combined effect of the duty cycle of the source and the maximum acceptable MPDU access delay of the service s, while $1 - P_{\rm limit}^{\rm (s)}$ accounts for the MPDU loss constraint of the sth service from dropping and/or outage. In [9], $M_r^{\rm (max)}$ is shown to be given by (14), where L_s is the *mean load* of a service s in the cell and is given by (15) in [9]. The mean load of a service states how much (normalized power) resource is required on average by all the users of a particular service in order to satisfy their QoS requirements. Equation (14) simply states that $M_r^{\rm (max)}$ is given by subtracting the average resource used by all services except service r from the total available resource and dividing this unused resource by the average resource used by a typical user of service r

$$M_r^{\rm (max)} = \left\lfloor \frac{1}{\delta_r^{\rm (min)} \times P_{\rm limit}^{(r)} \times z_r^{\rm (max)}} \times \left(Z_{\max}(b_j) - \sum_{\substack{s=1 \\ s \neq r}}^{N_S} L_s \right) \right\rfloor \tag{14}$$

$$L_s = M_s \times \delta_s^{\rm (min)} \times P_{\rm limit}^{(s)} \times z_s^{\rm (max)}. \tag{15}$$

VII. Simulation Experiment

The objective of this experiment is to determine for a fixed number of voice users the upper limit on the number of NRT-HSD users that can be supported by each protocol on the uplink of a third-generation radio access network while satisfying the QoS requirements of all users. This upper limit on the number of NRT-HSD users defines the capacity operating point of the protocol for the investigated scenario. ST-OSSS without DDB (ST-OSSS), ST-OSSS with DDB (ST-OSSS DDB), MT-OSSS with DDB (MT-OSSS DDB), and OSSS/OD with DDB (OSSS/OD DDB) are compared, with EDF-PLS as the benchmark.

A. Simulation Model and Parameters

A discrete-event simulator has been implemented for this investigation and is written entirely in proprietary C code, with the system and MAC modeled using behavioral models. The simulation model comprises of a base station receiver, a radio channel, user equipment, and the considered MAC algorithms. The radio channel is associated with the base station receiver and represents the radio channel between the user equipment and the base station. The radio channel is modeled as a *resource bank* with the resource measured in normalized power units and contains a maximum resource equal to the normalized power capacity at the base station receiver as defined by the RHS of (5). Whenever user equipment transmits an MPDU to a base station receiver, the MPDU "borrows" an amount of normalized power resource, equal to its intracell normalized power requirement from the radio channel for the duration of its transmission (equal to one 10-ms slot). If at any time, the amount of normalized power resource remaining in the radio channel is negative, then an outage is declared by the base station receiver and all involved MPDU transmissions are received with unacceptable SIR. Each piece of user equipment is modeled as a traffic generator with a queue, and depending on the type of service it supports, the MPDU arrival statistics are represented by either the voice or NRT-HSD traffic models presented in Section V-A. The output of user equipment is binary, either indicating an MPDU is ready for transmission, thus invoking the MAC routine, or indicating an empty user-equipment buffer. The MAC routines are all modeled as request/permission algorithms, with user equipment sending requests for transmission and the MAC responding with permissions to transmit in a way that mimics

TABLE I
SUMMARY OF TRANSMISSION BEARER PARAMETERS

Bearer name	E_b/I_o	Link quality	Bit rate R_B	Intracell normalised power requirement
Voice8	3.02 (4.8 dB)	Bit error rate=10^{-3}	8kb/s	1/170.5364
UDD2048	1.1482 (0.6 dB)	Block error rate =10^{-6}	486.4kb/s	1/8.3341

the behavior of the respective MAC strategies as presented in Sections III and IV.

In order to calculate the intracell normalized power requirements for the voice and NRT-HSD MPDUs, the transmission E_b/I_o targets presented in Table I are used, which are obtained for the "Indoor A" model [13]. The normalized power requirement is calculated using (6), where the SIR target in (6) equals the E_b/I_o target weighted by the factor $(R_B)/(W)$, where R_B and W are the transmission bit rate and channel chip rate in WCDMA, respectively. In this study, it is assumed that an MPDU is transmitted using a constant bit rate transmission bearer, with a fixed E_b/I_o requirement and the fluctuation in link gains due to shadow fading have been ignored. These assumptions do not impact on MAC protocol operation, since the MAC is always aware of the normalized power requirement values in the system, whether they are fixed or varying. The voice telephony and NRT-HSD services, considered in the experiment, use the voice8 and UDD2048 transmission bearers, respectively, and Table I summarizes the characteristics of these two transmission bearers [13].

In order to calculate the maximum resource contained on the radio channel (the normalized power capacity at the base station receiver), the values of the following parameters are required: namely, the link gain to the edge of the coverage area H_{edge}, the shadow fading margin H_{fade}, the maximum intracell normalized power requirements for voice $z_{\text{voice}}^{(\max)}$ and NRT-HSD $z_{\text{hsd}}^{(\max)}$, the peak transmit powers for voice $P_{T,\text{voice}}^{(\max)}$ and NRT-HSD $P_{T,\text{hsd}}^{(\max)}$, and the background noise $P_{\text{noise}}(b_j)$. The picocell coverage radius is set at 100 m from the base station, and user equipment is considered to be located on the same floor of the building. Then, H_{edge} is calculated as 2×10^{-10} using the model for the mean distance dependent link gain in [12] for indoor communication, which is given by $H(r) = \alpha r^{-3}$, where $\alpha = 2 \times 10^{-4}$ and r is the distance in meters (on the same floor of the building) from user equipment to base station. Also, $P_{T,\text{hsd}}^{(\max)} = P_{T,\text{voice}}^{(\max)} = P_{T,s}^{(\max)} = 2.5$ mW [12], $P_{\text{noise}}(b_j)$ is set equal to the thermal noise power of 1.99×10^{-11} mW [12], H_{fade} is set equal to unity (no shadow fading), $z_{\text{voice}}^{(\max)} = 1/170.5364$ and $z_{\text{hsd}}^{(\max)} = 1/8.3341$ (from Table I), and $\min_s[P_{T,s}^{(\max)}/z_s^{(\max)}]_{s=1}^{N_S}$ equals $P_{T,s}^{(\max)} \times 8.3341$ (set by the UDD2048 bearer from Table I). Given these parameter values, $P_{\text{tot}}^{(\max)}$ is calculated as 4.17×10^{-9} mW using (8). Substituting this into the RHS of (5) gives a normalized power capacity and, hence, the maximum radio channel resource of almost unity.

For the proposed distributed scheduling protocols, a small access window duration is assumed for efficiency (for example 5% of the slot duration) and a high access window to latency ratio $\tau_{\text{aw}}/\tau = 50$. Both uplink and downlink signaling errors are assumed to be negligible; however, in practice this will depend on whether the processing time available is sufficient for accurate detection of the signals. In [9], it is shown that the performance of the proposed protocols improves the higher the value of the access widow to latency ratio; however, the higher the access widow to latency ratio, the larger the size of the access window required to maintain a high protocol signaling reliability and a large cell coverage. An access window to latency ratio of 50 represents a sufficiently high value to achieve high protocol performance, while not resulting in too large an access window. Achieving this tradeoff, however, may complicate the practical implementation of the protocols. Finally, the DDB strategy uses an exponential window size selection rule, with $T_{\text{win}}^{(\max)}$ equal to 1 slot duration to minimize MPDU delays. The characteristic overload-avoidance thresholds to be employed in ST-OSSS DDB (and ST-OSSS) and MT-OSSS DDB are optimized as in [9] for the voice and NRT-HSD transmission bearer types with normalized power requirements given by $z_1 = 1/170.5364$ and $z_2 = 1/8.3341$, respectively. These characteristic overload-avoidance thresholds are $Z_{\text{OL}}(z_1) = 0.9382$ and $Z_{\text{OL}}(z_2) = 0.8399$ with MT-OSSS DDB employing both thresholds. However, ST-OSSS DDB (and ST-OSSS) employs the lower of the two characteristic overload-avoidance thresholds in order to satisfy the condition in (10).

In [9], the characteristic overload-avoidance threshold $Z_{\text{OL}}(z)$ for a normalized power requirement equal to z is determined numerically, using a mathematical model of the behavior of the ST-OSSS basic access procedure when driven by Poisson MPDU arrival statistics, where the characteristic overload-avoidance threshold is that threshold that maximizes the throughput (minimizes delay and outage). The value of the characteristic overload-avoidance threshold selected by this method tends to be somewhat restrictive (higher delays than necessary and nearly zero outage) when applied to non-Poisson MPDU arrival statistics as depicted in [9]. In practice, characteristic overload-avoidance thresholds for a finite set of normalized power requirement values between zero and unity can be calculated offline and stored in a network database from which base stations can access them as required. In this case, the set of normalized power requirement values with characteristic overload-avoidance thresholds is limited by memory constraints. Therefore, in practice the characteristic overload-avoidance threshold, for a normalized power requirement value not included in this set, is that of the nearest normalized power requirement in that set. Results in [9] show that an overload-avoidance threshold that is 14% too low results in a drop in achievable capacity of about 6%, while an overload-avoidance threshold that is 6% too high results in a drop in achievable capacity of about 47%. Therefore, it

TABLE II
MAC PROTOCOL CAPACITY OPERATING POINTS FOR SINGLE-CELL OPERATION

Constants: Average number of voice users = 86, Fluid Capacity for HSD = 359 users $P_{limit}^{(hsd)} = 0.99$, $D_{limit}^{(hsd)} = 4.88\times10^{-3}$ ms, $P_{limit}^{(voice)} = 0.95$					
	Capacity Operating Points				
	EDF-PLS	ST-OSSS	ST-OSSS DDB	MT-OSSS DDB	OSSS/OD DDB
NRT-HSD Support	[201,301]	[101,201]	[201,301]	[201,301]	[201,301]
Voice Support	[701, 801]	[101, 201]	[201, 301]	[201, 301]	[201, 301]
Joint NRT-HSD and Voice Support	[201,301]	[101,201]	[201,301]	[201,301]	[201,301]

is essential to use optimization strategies that tend to select overload-avoidance thresholds that are too low, rather than too high.

In order to determine the capacity operating points of the various schemes from the simulation results, values for the parameters $P_{\text{limit}}^{(\text{voice})}$, $P_{\text{limit}}^{(\text{hsd})}$, and $D_{\text{limit}}^{(\text{hsd})}$, defined in Section V-B, must be specified. Here, $P_{\text{limit}}^{(\text{voice})}$ is set equal to 95%, which is a maximum data loss rate of 5% and is the same value as the constraint on data loss due to outage that is suggested in [12] for voice. For the NRT-HSD services $P_{\text{limit}}^{(\text{hsd})}$ is set high at 99% and assumes a 1% loss rate is easily compensated for by higher layers through the resubmission of "lost" MPDUs. The minimum acceptable value for the *active session throughput* for the data service is set equal to 10% of the nominal peak source bit rate [12], and therefore for the NRT-HSD service with its UDD2048 bearer it equals 204.8 kb/s, which is equivalent to $D_{\text{limit}}^{(\text{hsd})}$ equaling 4.88 μs, since $D_{\text{limit}}^{(\text{hsd})}$ is the inverse of the active session throughput constraint [12].

The delivery delay constraint $T_{\text{DC}}^{(\text{voice})}$ for a voice MPDU (before it is considered late) is assumed to be one slot or 10 ms (i.e., zero access delay). The choice of delivery delay constraint for the NRT-HSD service should be infinite but is limited by buffer length restrictions and is chosen to be finite but much larger than the mean delay requirement. Therefore, $T_{\text{DC}}^{(\text{hsd})}$ is chosen to be 10 s, which is substantially greater than ten times the mean delay requirement $D_{\text{limit}}^{(\text{hsd})}$. The minimum acceptable duty cycle for voice $\delta_{\text{voice}}^{(\text{min})}$, given the one-slot delivery delay constraint, equals the duty cycle of the source, which is 0.5. From (13), the minimum acceptable duty cycle $\delta_{\text{hsd}}^{(\text{min})}$ for the NRT-HSD service is calculated as 0.0178, where L_{pad} equals 418 bytes. The parameters $\delta_{\text{voice}}^{(\text{min})}$ and $\delta_{\text{hsd}}^{(\text{min})}$ are used in estimating the capacity operating point using (14) and (15).

In this experiment, each scheme is simulated and performance statistics collected for the cases when the number of supported NRT-HSD users equal 1, 101, 201,..., 1001, with the number of supported voice users kept at 86 in each case, which is equivalent to supporting a low voice load of approximately 0.24. The simulation is run for 1000 s of simulation time, in each case, and performance statistics are not collected for the first 100 s (simulation time) of each simulation run in order to allow the system to reach a "steady state," after which no significant fluctuations in the performance statistics are observed. The number of NRT-HSD users is varied in this way, in order to determine the capacity operating points of the proposed protocols and the EDF-PLS algorithm. A total of 11 points of performance statistics are collected all together for each scheme, and these are plotted graphically to give the simulation results presented in the next section.

B. Simulation Results

Table II shows the capacity operating points for the protocols. The capacity operating points when using OSSS/OD DDB, ST-OSSS DDB, and MT-OSSS DDB are very close to that of the benchmark EDF-PLS algorithm and lie in the same range [201, 301], which is an indication of high efficiency (refer to Figs. 3–5). A further indication that these three protocols offer near-optimal performance is that their capacity operating points are slightly less than the fluid capacity estimate for NRT-HSD, where the fluid capacity for NRT-HSD, given 86 voice users, is obtained from (14) as $[(8.3341)/(0.0178 \times 0.99) \times (1 - (86 \times 0.5 \times 0.95)/(170.5364))] = 359$ users. When using ST-OSSS without DDB, the capacity operating point lies in the range [101, 201] (see Table II and Figs. 3–5). This represents approximately a 50% drop in the number of NRT-HSD users that can be supported without compromising QoS when compared to protocol performance with the use of DDB. The use of DDB in ST-OSSS DDB enables a fairer allocation of resources between real-time voice and NRT-HSD, than ST-OSSS alone, resulting in a higher voice P_{succ} (Fig. 3) and a lower NRT-HSD delay (Fig. 4), a higher NRT-HSD P_{succ} (Fig. 5), and a higher NRT-HSD throughput (Fig. 6) than those achieved by ST-OSSS alone.

MT-OSSS DDB is seen to improve on the voice P_{succ} performance (see Fig. 3) of ST-OSSS DDB, without a noticeable decline in the NRT-HSD delay, throughput, and P_{succ} performance (see Figs. 4–6). But this improvement results in no significant capacity enhancement. This improvement is expected since MT-OSSS DDB, by using a more appropriate higher threshold for voice while retaining the lower more appropriate threshold for NRT-HSD, simply permits voice MPDUs to utilize resources which otherwise would have been

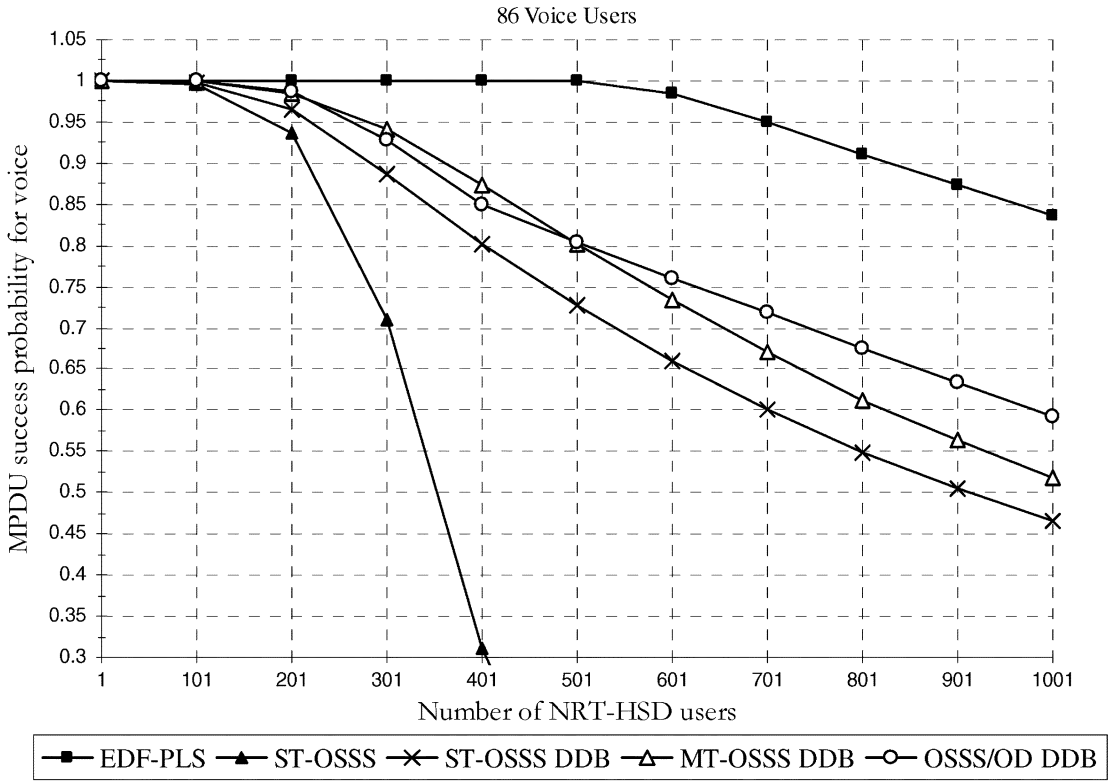

Fig. 3. MPDU success probability P_{succ} for voice versus the number of NRT-HSD users with 86 voice users.

wasted when using ST-OSSS DDB. The "wasted" resource when using ST-OSSS DDB, which is roughly equal to the difference between the two characteristic thresholds, is too small to support an additional NRT-HSD MPDU transmission but can support a number of voice MPDU transmissions with their much smaller normalized power requirement.

OSSS/OD DDB does not improve significantly on the capacity achieved by MT-OSSS DDB. However, OSSS/OD DDB does offer a fairer allocation of resources between voice and NRT-HSD than MT-OSSS DDB, delivering a lower NRT-HSD delay (Fig. 4), a higher NRT-HSD P_{succ} (Fig. 5), and a higher NRT-HSD throughput (Fig. 6) than those achieved by MT-OSSS DDB as well as a higher voice P_{succ} (Fig. 3) over most of the observed range. When the number of NRT-HSD users is large (greater than 501), the voice P_{succ} and NRT-HSD throughput of OSSS/OD DDB improves significantly on that of MT-OSSS DDB, due to an increased frequency of outage in the operation of the latter as the MT-OSSS basic access procedure is overwhelmed by the heavy load. This region is well beyond the capacity operating points of both protocols; hence, the performance difference between the protocols is not as relevant in this region. However, the higher NRT-HSD throughput and the higher voice P_{succ} performance of OSSS/OD DDB at such high loads is an indication that OSSS/OD DDB is a more stable protocol than MT-OSSS DDB.

Summarizing the results, it is observed that the overload detection-based OSSS/OD DDB offers the best and most stable performance compared to the overload-avoidance-based schemes. Also, the use of multiple characteristic overload-avoidance thresholds for overload avoidance results in better QoS support compared to using a single overload-avoidance threshold in a mixed service interference limited environment such as in a third-generation radio access network. The most significant enhancement to protocol efficiency in this environment, however, results from the use of DDB in the proposed protocols.

C. Mobile Multicellular Operation

Results have been presented for the case of a single isolated cell with stationary users in order to evaluate MAC protocol efficiency rather than the overall system spectral efficiency. However, the proposed MAC protocols have been defined in such a manner that their behavior is not altered by user-equipment mobility and/or multicell operation. Multicell operation is accommodated in the protocols in two ways. First, every base station measures intercell normalized power requirements of those users in its soft-handover region but under the power control of other base stations, which are typically the most significant intercell interferers. Second, each base station broadcasts overload signals to both the intracell interferers and the intercell interferers within its soft-handover region. The effect of mobility, with respect to the protocols, is on the value of the intercell normalized power requirement (but not on the intracell normalized power requirement), which results from variations in the link gains between user equipment and the base station. However, each base station tracks the variations in link gain and, hence, intercell normalized power requirement of the intercell interferers within its soft-handover region (these are the relevant intercell interferers). The remainder of this section examines the

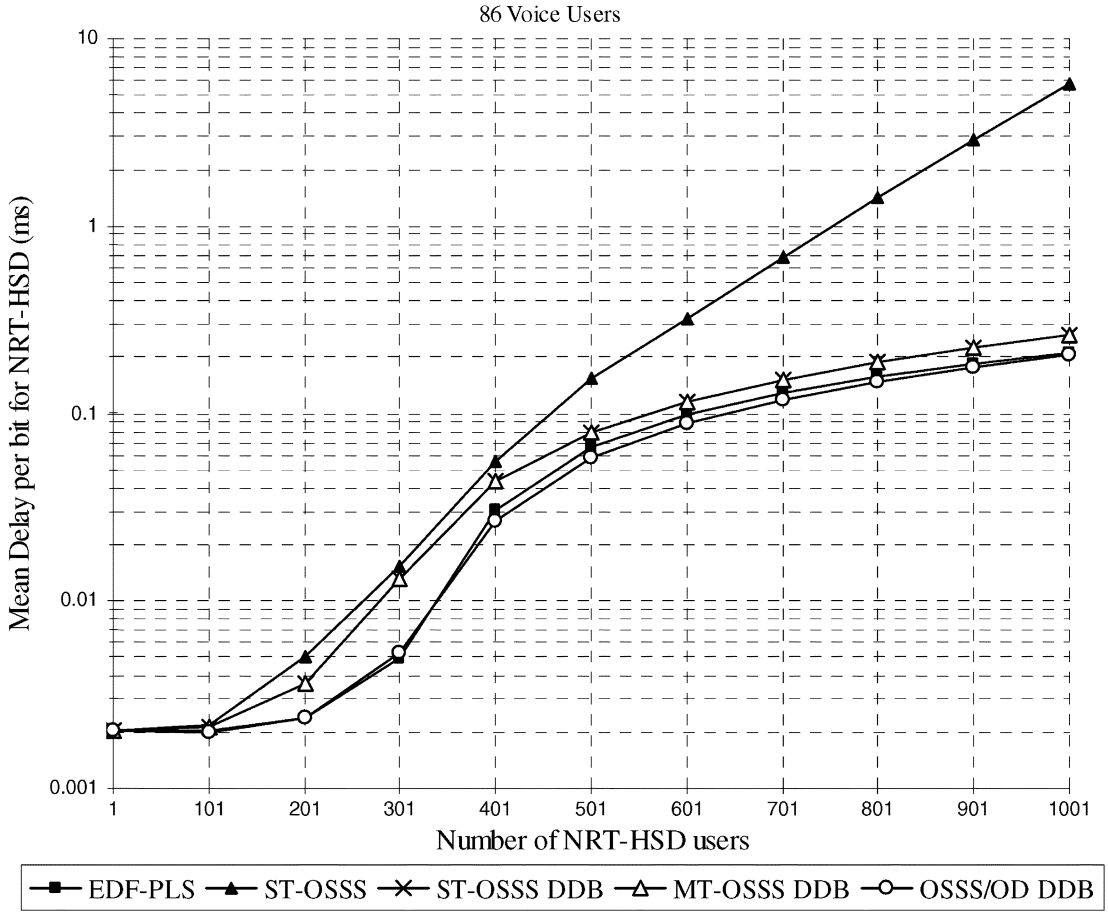

Fig. 4. Mean delay per bit for NRT-HSD versus number of NRT-HSD users with 86 voice users.

impact a mobile multicellular environment will have on the performance of the proposed distributed scheduling protocols when compared to the performance in a single-cell scenario with stationary users.

From the perspective of the "tagged" cell in a mobile multicellular environment, there are three main subsets of users at any time, namely the set $U_{\rm pc}$ of users that are under the power control of the "tagged" cell, the set $U_{\rm SH,out}$ of users that are outside its power control but lie within its soft-handover region, and the set $U_{\rm out}$ of users that lie outside both its power control and soft-handover region. Let $M_{\rm pc}, M_{\rm SH,out}$, and $M_{\rm out}$ denote the average number of users in sets $U_{\rm pc}, U_{\rm SH,out}$, and $U_{\rm out}$, respectively. From the perspective of the proposed distributed scheduling MAC protocols, a "tagged" cell in the multicellular environment is identical to an isolated cell with users in sets $U_{\rm pc}$ and $U_{\rm SH,out}$ under its protocol control and all users in $U_{\rm out}$ outside its protocol control. The protocol threshold(s) of the distributed scheduling MAC protocols must be reduced from their isolated cell values, in order to accommodate the users in $U_{\rm out}$, which results in a proportionate reduction in the average normalized power capacity available to support users in set $U_{\rm pc}$. This reduction equals the average normalized power requirement of the users in $U_{\rm out}$, denoted by $z_{\rm out}$, plus a margin $c_{\rm guard}$ to allow for short term variations of the intercell normalized power requirement of the "uncontrolled" users in $U_{\rm out}$ and for errors in its measurement, respectively. This capacity reduction has the effect of reducing the maximum supportable value of $M_{\rm pc}$ and hence the capacity operating point per cell from its isolated cell value. A further reduction in capacity operating point per cell results from the additional normalized power load on the "tagged" cell from users in set $U_{\rm SH,out}$, denoted by $z_{\rm SH,out}$. Therefore, the average normalized power capacity available to support users in set $U_{\rm pc}$ is reduced in total from its isolated cell value by the sum of $z_{\rm out}, z_{\rm SH,out}$, and $c_{\rm guard}$.

From the perspective of a centralized scheduling scheme, a "tagged" cell in the multicellular environment is identical to an isolated cell with users in set $U_{\rm pc}$ under its protocol control and all users in sets $U_{\rm out}$ and $U_{\rm SH,out}$ outside its protocol control. This is because a centralized scheduler, in a cell, can only schedule users that are local to its cell, in order to avoid conflicting schedules in different cells. Therefore, EDF-PLS in the "tagged" cell has a larger set of uncontrolled users to contend with than the proposed distributed scheduling protocols. By accommodating the additional interference from users in $U_{\rm out}$ and $U_{\rm SH,out}$, the EDF-PLS protocol suffers a proportionate reduction in the average normalized power capacity available to support users in set $U_{\rm pc}$. As for the distributed protocols, the reduction is equal to the sum of $z_{\rm out}, z_{\rm SH,out}$ and the margin $c_{\rm guard}$, where the latter accounts for the short term variations of the intercell normalized power load of the "uncontrolled" users. However, the inability of centralized scheduling to control $z_{\rm SH,out}$ results in large fluctuations of the relevant intercell in-

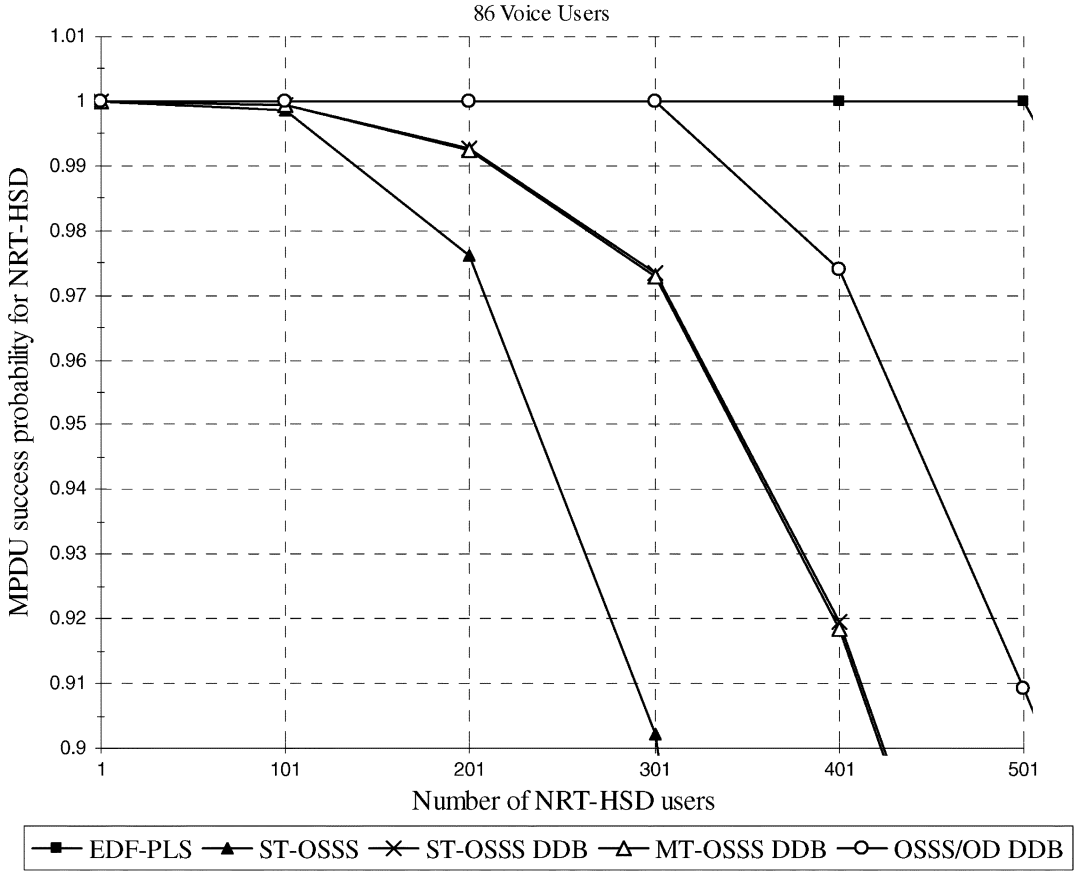

Fig. 5. MPDU success probability P_{succ} for NRT-HSD versus the number of NRT-HSD users with 86 voice users.

terference. For this reason, the centralized scheduling protocol requires a larger value of c_{guard} than the proposed distributed scheduling protocols. In simulation, values of c_{guard} equal to 0.4 and 0.25 of the normalized power requirement were used for the centralized and distributed protocols, respectively. The values are chosen by systematically increasing c_{guard} in simulation until the throughput metric was maximized for the respective approaches. The increased values of intercell interference fluctuation c_{guard} and $z_{\mathrm{SH,out}}$ experienced by centralized scheduling means that this approach suffers a greater loss of system capacity than distributed scheduling.

Simulation results are presented for a small indoor multicellular system consisting of a "tagged" base station at the center of six uniformly spaced base stations forming a circle around it. Therefore, assuming polar coordinates, the coordinates for the "tagged" base station are (0 m, 0°), while that of the other bth base station surrounding it are (200 m, $b \times 60°$), for $b = 0, 1, \ldots, 5$. Users are assumed to move with Brownian-like motion at a speed of 3 km/h but remain within the boundaries of the system. The simulator models the variation of the normalized power requirement at the "tagged" base station of users as they move within the system. For simplicity, shadow fading is ignored and the soft-handover region is approximated by an annular (ring) shape. Therefore, when a user is within a 100-m radius of the "tagged" base station, it is under the power control of the latter and its (intracell) normalized power requirement at the "tagged" base station is independent of location within this region. However, when a user is outside the 100-m radius of the "tagged" base station, it is under the power control of a neighboring base station and its (intercell) normalized power requirement at the "tagged" base station varies with changes in location. An experiment was performed to measure the capacity operating points of the EDF-PLS and OSSS/OD DDB protocols. EDF-PLS upper bounds the performance of state-of-the-art centralized MAC protocols while OSSS/OD DDB represents the best of the proposed distributed MAC protocols. The results obtained are presented in Table III. The capacity operating point per cell, as determined at the "tagged" base station, are measured as the maximum number of NRT-HSD users per cell (under its power control) that can be supported in a mix containing 86 voice users on average per cell (under its power control). The performance of OSSS/OD DDB is presented for different sizes of the soft-handover region outside the "tagged" base station's power control which is assumed to be in the shape of a ring and is measured in terms of its radial width as denoted by $R_{\mathrm{SH,out}}$.

The results show that OSSS/OD DDB outperforms centralized scheduling for all sizes of the soft-handover region considered. The percentage drop in the number of supported NRT-HSD users is about 64% for OSSS/OD DDB with $R_{\mathrm{SH,out}}$ equal to 25 m and even lower for higher values of $R_{\mathrm{SH,out}}$, while centralized scheduling experiences a drop of 86%. The larger drop, in the case of the latter, is because the centralized scheduler, unlike for distributed scheduling, cannot schedule

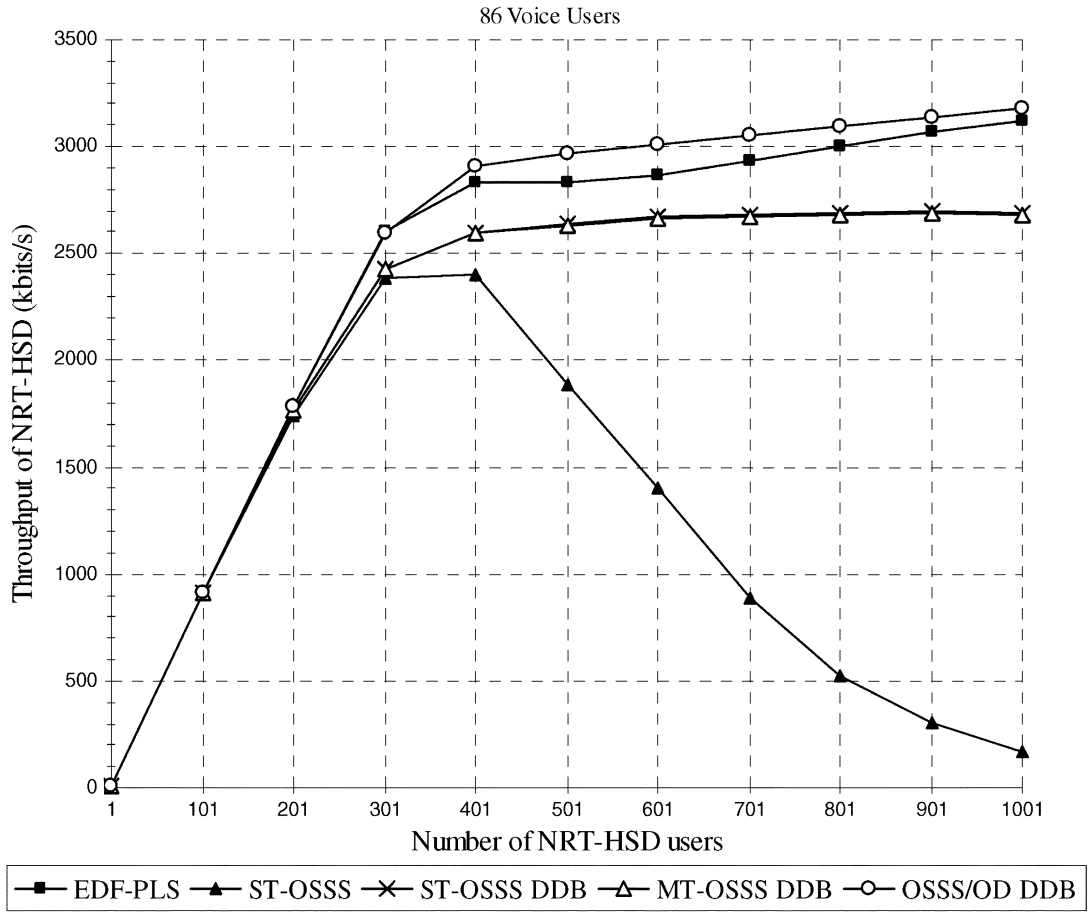

Fig. 6. Throughput for NRT-HSD against number of NRT-HSD users with 86 voice users.

users within the coverage of neighboring cells, even if these users are in the soft-handover region of its own cell.

VIII. Conclusion

Low-complexity distributed scheduling control strategies using overload-avoidance/detection can be designed using the proposed normalized power resource metric to give near-optimal performance and thus maintain a high spectral efficiency on the interference limited uplink in third-generation radio access networks. The overload-detection approach achieves a superior performance compared to overload avoidance and offers a lower implementation complexity. This outcome is fundamental since it suggests that an overload-detection approach, which attempts to minimize the effects of overloads when they do occur, is inherently superior to an overload-avoidance approach, which attempts to minimize the frequency of overloads. The overall low complexity of the proposed distributed scheduling control techniques is attributable to their overload-avoidance and overload-detection features. In addition, the proposed protocols are expected to be straightforward to integrate with the existing standards because of the use of the traditional distributed scheduling control principles of overload-avoidance/detection and DDB in the design of the proposed protocols, where DDB is a QoS-enabled enhancement to standard backoff techniques. A significant enhancement on distributed scheduling control protocol efficiency is achieved by the use of DDB in these protocols, with the results indicating that it is possible to support as many as 50% more NRT-HSD users by using DDB. Finally, the design of the protocols offers flexibility in the support of different service types with diverse QoS requirements in an interference limited environment.

A key advantage of using the proposed distributed scheduling schemes is revealed from the performance of OSSS/OD DDB in the multicellular environment. The results suggest that the distributed nature of these protocols readily allows for intercell cooperation within the soft-handover regions. By influencing the scheduling of mobile traffic within the soft-handover regions, the proposed distributed protocols reduce intercell interference effects, resulting in an enhanced management of the radio resource when compared with centralized scheduling. The size of the soft-handover region affects the performance of the proposed distributed scheduling schemes as indicated in Table III.

Future work should focus on overload detection-based distributed scheduling control schemes and consider a theoretical study and further empirical evaluation of DDB, with an emphasis on optimizing the DDB process through the choice of the window size selection rule and its parameters. Signaling implementation issues and the impact of signaling errors on protocol performance with respect to various signaling implementations need to be addressed. Also, even though the design of the protocols offers the flexibility to support different service types with diverse QoS requirements in an interference limited environment, the results presented in this paper only consider the sup-

TABLE III
MAC PROTOCOL CAPACITY OPERATING POINTS FOR MOBILE MULTICELL OPERATION

Constants: Average number of voice users per cell = 86 $P_{limit}^{(hsd)} = 0.99$, $D_{limit}^{(hsd)} = 4.88\times10^{-3}$ ms, $P_{limit}^{(voice)} = 0.95$				
	Capacity Operating Points			
	EDF-PLS	OSSS/OD DDB		
		$R_{SH,out}$ = 100m	$R_{SH,out}$ = 50m	$R_{SH,out}$ = 25m
Overload-detection threshold	Not applicable	1.0	0.80	0.60
Maximum number of supported NRT-HSD users	[29,43]	[129,143]	[129,143]	[72,100]

port of two extremes (voice and nonreal-time data). Evaluation of the proposed distributed scheduling approach, in supporting more diverse QoS requirements, is important and is currently under investigation by the authors, with promising results. Finally, the effect of improved MAC protocol efficiency on the overall system spectral efficiency of a third-generation radio access network should be investigated, taking into account the impact of connection admission control, macro-diversity, mobility management, and handover algorithms in a multicell environment.

REFERENCES

[1] D. Petras, "Medium-access-control protocol for wireless, transparent ATM access," in *Proc. IEEE WCSS'95*, pp. 79–85.

[2] D. Raychaudhuri and N. D. Wilson, "ATM-based transport architecture for multiservices wireless personal communication networks," *IEEE J. Select. Areas Commun.*, vol. 12, no. 8, pp. 1401–1414, Oct. 1994.

[3] L. Carrasco and G. Femenias, "W-CDMA MAC protocol for multimedia traffic support," in *Proc. IEEE VTC'00 Spring*, pp. 2193–2197.

[4] A. Sampath and J. M. Holtzman, "Access control of data in integrated voice/data CDMA systems: Benefits and tradeoffs," *IEEE J. Select. Areas Commun.*, vol. 15, no. 8, pp. 1511–1526, Oct. 1997.

[5] K. Toshimitsu, T. Yamazato, M. Katayama, and A. Ogawa, "A novel spread slotted aloha system with channel load sensing protocol," *IEEE J. Select. Areas Commun.*, vol. 12, no. 4, pp. 665–672, May 1994.

[6] P. E. Omiyi and T. O'Farrell, "Throughput analysis of novel CDMA-based MAC protocol for wireless LANs," *Electron. Lett.*, vol. 34, no. 12, pp. 1201–1203, Dec. 1998.

[7] D. M. Lim and H. S. Lee, "Throughput analysis of channel load sense multiple-access with overload detection protocol in spread-spectrum packet radio networks," *Electron. Lett.*, vol. 27, no. 7, pp. 1809–1810, Jul. 1991.

[8] A. 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, Dec. 1996.

[9] E. O. P. Omiyi, "Medium-Access-Control for third generation cellular mobile systems," Ph.D. dissertation, Univ. Leeds, Leeds, U.K., Dec. 2000.

[10] A. J. Viterbi, *CDMA: Principles of Spread Spectrum Communications*. New York: Addison-Wesley, 1995.

[11] K. K. Ramakrishnan and H. Yang, "The ethernet capture effect: Analysis and solution," in *Proc. LCN'94*, pp. 228–240.

[12] *Universal Mobile Telecommunications System (UMTS); Selection Procedures for the Choice of Radio Transmission Technologies of the UMTS*, 1998.

[13] *Updated Annex C of ETSI's RTT Proposal*, Sep. 14–15, 1998.

[14] H. Saito, "Optimal queueing discipline for real-time traffic at ATM switching nodes," *IEEE Trans. Commun.*, vol. 38, no. 12, pp. 2131–2136, Dec. 1990.

Timothy O'Farrell (M'91) received the B.Sc. (Hons.) degree in electrical and electronic engineering from the University of Birmingham, Birmingham, U.K., in 1980 and the M.Sc. and Ph.D. degrees in electrical and electronic engineering from the University of Manchester, Manchester, U.K., in 1986 and 1989, respectively.

Since 1996, he has been with the School of Electronic and Electrical Engineering, University of Leeds, Leeds, U.K., where he is currently a Senior Lecturer in communications. His research interests include broadband wireless communications, in particular multiple-access techniques, sequences/coding/modulation, and radio resource management. He has published more than 100 refereed publications in journals and conferences and holds five patents in these areas. He presented the MBCK proposal for the higher rate extension to 802.11b in the IEEE802.11g standardization activity.

Peter Omiyi received the B.Eng. (Hons.) degree in electrical and electronic engineering from the University of Manchester, Manchester, U.K., in 1996, and the Ph.D. degree in electrical and electronic engineering from the University of Leeds, Leeds, U.K., in 2001. His doctoral thesis is titled "Medium access control for third generation cellular mobile systems" in which quality of service sensitive MAC protocols for third-generation radio access networks are researched.

Since July 2003, he has been with the International University of Bremen, Bremen, Germany. His current research interests include radio resource management for third– and fourth-generation RANs.