QoS-Aware Distributed Resource Management for a WCDMA Uplink

Pratik Das, *Student Member, IEEE*, and Jamil Y. Khan, *Senior Member, IEEE*

***Abstract*—In this paper, a hybrid resource management system for the uplink in Universal Mobile Telecommunications System wideband code-division multiple access (WCDMA) with components in the Node-B and user equipment (UE) has been proposed. A rate scheduler in the client focuses on average packet delays as a means of abstracting application-specific requirements from the rest of the resource management scheme. It controls uplink transmission through variable spreading gain to optimize resource usage while meeting target delays. Service change requests from the distributed rate schedulers are collectively processed through interservice and intraservice priority queuing in a manner that is shown to exhibit fairness in allocation of resources when cumulative load exceeds system capacity. The performance of the proposed algorithm is explored through discrete-event simulations for three classes of traffic, namely voice, video, and data, over the WCDMA uplink in the presence of short-term Rayleigh fading, automatic repeat request, forward error correction, target transmission delays to meet the respective quality of service, and frame error rate targets in a "multicell" environment. The authors analyze two alternatives for distributed resource management with the UE or Node-B in control of rate scheduling and observe the fairness in resource allocation of both systems. Priority of speech, video, and data traffic is respected and reflected in 95th percentile transmission delays for heavily loaded systems.**

***Index Terms*—Radio resource management (RRM), rate scheduling, Universal Mobile Telecommunications System (UMTS), uplink, wideband code-division multiple access (WCDMA).**

I. Introduction

THE POPULARITY of applications such as email, web browsing, streaming, and nonreal-time multimedia have led to tremendous growth in global communication over the Internet. Recently, "broadband" Internet access through digital subscriber line (DSL), cable, and satellite networks has improved the diversity, quality, and quantity of media available to the Internet user. This greater freedom in accessibility has led to a 155% compound annual growth rate (CAGR) in broadband penetration worldwide between 1999 and 2002, with forecasts predicting further acceleration in penetration as costs reduce [1].

The last decade has seen a push for research and development to bridge the gap between the quality of content accessible over mobile systems and what is available over fixed networks. Third-generation (3G) and fourth-generation (4G) mobile communication technology aims to narrow this gap. The relevance of the mobile communication device and the need to provide services through it similar to those available on fixed networks have never been greater. At the end of 2003, there were over 1.35 billion mobile subscribers worldwide as compared with 1.2 billion fixed-line communications users [2].

It is known that different applications require different qualities of service (QoS) from underlying communication links and that traditionally circuit-switched transmission schemes have been used in wireless communications to guarantee QoS. However, with the bandwidth available being less than that required to meet demand in the licensed civilian wireless communication domain, and the high cost associated with spectrum ownership, bandwidth becomes the most valuable resource in such a scenario. In this context, the overallocation of resources for each circuit or call, i.e., the spectral inefficiency, becomes a significant demerit. Efficient use of bandwidth can be promoted through packet-switched data transmission because of the gain in throughput achieved when statistically multiplexing users according to their transmission loads and channel conditions. However, the QoS experienced by a user and the diversity in the link requirements of different applications must also be considered [3]. By maintaining application-specific service requirements while improving network throughput, QoS schemes stand as an enabling technology for the success of 3G systems.

The Universal Mobile Telecommunications System (UMTS) is a 3G system in development within the Third-Generation Partnership Project (3GPP) with wideband code-division multiple access (WCDMA) as the radio access mechanism. In this paper, we concentrate on resource management for the WCDMA uplink. UMTS provides a common channel and a dedicated channel (DCH) for uplink transmissions. While the former would support true packet-switched communication, the latter can be also used to provide packet-controlled communication with similar statistical multiplexing gains because of the ability to throttle the transmission rate over time. UMTS enables different data rates for users in several ways, i.e., through variable spreading gain (VSG) per transmission channel, through multiple codes (MC) with fixed spreading gain (FSG) transmitted in parallel, or a combination of the two [4]. However, for simplicity of implementation, most propose the use of one or the other. In terms of the signal-to-noise ratios (SNRs) and bit-error rates (BERs) experienced by users, both approaches have an identical effect on the system [5], [6]. The spreading gain or the number of codes allocated to users can also be scheduled differently at periodic intervals in UMTS.

Manuscript received September 4, 2004; revised October 26, 2005 and January 25, 2006. This work was supported by the University of Newcastle, Callaghan, N.S.W., Australia. The review of this paper was coordinated by Dr. A. Chockalingam.

The authors are with the School of Electrical Engineering and Computer Science, University of Newcastle, Callaghan, N.S.W. 2308, Australia (e-mail: pratik@ee.newcastle.edu.au; jkhan@ee.newcastle.edu.au).

Digital Object Identifier 10.1109/TVT.2006.878738

We utilize this feature to extract gains in cumulative uplink throughput while maintaining the QoS provided.

A. Related Work

While the framework for end-to-end QoS is already in place within UMTS [7], alternative strategies for scheduling traffic across the air interface are still being evaluated within the 3GPP standardization process [8], [11]. Independent proposals for scheduling resources over a code-division multiple-access (CDMA) uplink have also been made [3], [5], [12]–[18].

Kim *et al.* address the problem of "optimal" dynamic rate adaptation under constrained signal-to-interference-and-noise ratio (SINR) for uplink packet data transmission in multicell multirate WCDMA systems in [19] and propose "suboptimal" dynamic rate adaptation solutions. While the scope for expanding the proposed framework to serve different classes of traffic is mentioned, the subject has not been fully explored.

Dual-traffic-class direct-sequence CDMA (DS-CDMA) systems are managed with spreading gain scheduling schemes in [20] and [21] to meet real-time and packet data QoS requirements. The effect of multipath fading or shadowing is not considered in [21]. Abrardo *et al.* map the QoS requirements of different applications into SNR requirements to express system capacity in terms of the number of user equipments (UEs) of each type that can be admitted in [20]. The lack of an automatic repeat request (ARQ) scheme is compensated for with an adjustment of the target BER. As a result, discrete packet transfer delays have not been analyzed.

A power control and rate selection scheme for WCDMA uplinks is proposed in [22] with the objective of maximizing the cumulative data rate in the system. QoS requirements of different application types and fairness in distribution of capacity across users given these requirements have not been explored.

This paper explores the performance of the proposed resource management scheme through discrete-event simulations for three classes of traffic, namely voice, video, and data, over the WCDMA uplink in the presence of short-term Rayleigh fading, ARQ, forward error correction (FEC), target transmission delays to meet QoS, and frame error rate (FER) targets in a "multicell" environment. We analyze two alternatives for distributed resource management with the UE or Node-B in control of rate scheduling and observe the fairness in resource allocation of both systems.

II. UMTS Radio Access Network

Here, we introduce the logical network elements and interfaces of the UMTS terrestrial radio access network (UTRAN) to set the basis for the rest of this paper. The equivalent of the mobile terminal and the base station in UMTS terminology is the UE and Node-B, respectively. The WCDMA radio access interface, labeled "Uu," is also known as the "air interface" and allows the mobile terminals to connect with the fixed network. Several Node-B's can connect to a radio network controller (RNC) through "Iub" interfaces. Fig. 1 shows the upper Node-B using an omnidirectional antenna within the cell and the lower Node-B using a sectorized antenna to service three cells. How

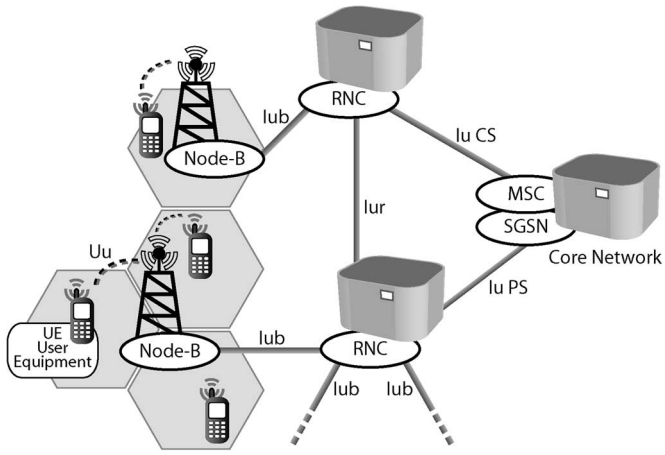

Fig. 1. Components of the UTRAN.

the RNCs connect to the core network depends on the type of services active in the radio access network. For circuit-switched services, the RNC gains access to the core network through the mobile services switching center (MSC), which is also found on second-generation Global System for Mobile Communication (GSM) networks. The "Iu CS" interface is used specifically for circuit-switched data. For packet-switched services, the RNC connects to the serving General Packet Radio Service (GPRS) support node (SGSN) instead through the "Iu PS" interface.

This paper looks at resource management strategies for single-carrier WCDMA in frequency division duplex (FDD) mode. It involves the three outermost network components, i.e., 1) the UE, 2) the Node-B, and 3) the RNC, and we aim to reduce signaling at all levels to improve the rate with which we can redistribute resources among active UEs.

III. Resource Management for the WCDMA Uplink—A Review

Resource management processes operate efficiently when provided with estimates of how much resource is available in total and how much is already in use. In the wireless communication domain, those correspond to the estimates of channel capacity and user load. As we shall see later in this section, CDMA systems pose certain challenges when these quantities are to be estimated.

Channel capacity estimation and resource scheduling differ in the uplink and downlink of a multicelled UMTS network, with some of the factors influencing these differences being

- the number of interferers;
- the degree of orthogonality across the different UE–Node-B transmissions;
- the availability of information on queued data awaiting transmission for optimizing the scheduling process from a QoS perspective;
- the feasibility of employing antenna diversity for greater receiver sensitivity.

Only the last factor makes conditions more favorable in the uplink than in the downlink. While base stations (Node-B) have the space and power required for sensitive receivers with

multiple antennas and amplifiers, mobile stations are typically more constrained in both aspects. Although every UE interferes with the transmission of another in the uplink, the large UE-to-Node-B ratio implies that downlink transmissions have to contend with much fewer interferers. Sectorized transmission in the downlink further improves conditions in this regard. Downlink transmissions to different users are orthogonal at first, but multipath effects can make reflected transmissions appear as multiple-access interference (MAI). However, uplink transmissions are only orthogonal across multiple channels from the same user and not across different users.

Resource management in the uplink through scheduling is typically centralized at the RNC or the Node-B [8]–[11]. It helps to know the queue sizes at all transmitters and the data rate required for the oldest packet in each to meet QoS targets when distributing capacity fairly among multiple users. Naturally, having an estimate of total capacity or remaining capacity relative to current load is essential. However, when scheduling is centralized, having UEs continually relay information about buffer sizes and packet delays to the scheduling authority generates large volumes of zero-revenue signaling data. The remainder of this paper looks at both these issues by reviewing two alternatives for estimating channel capacity as well as alternative configurations for distributed resource scheduling in the uplink.

A. Single Cell Load and Capacity Estimation

CDMA intrinsically has a "soft" capacity, i.e., the cumulative transmission rate in a cell is a function of the channel gain to the active UEs, their output power limits, spreading gains, FEC rates, and target BERs, with more factors coming into play when QoS to admitted users is given importance. Admission control and UE scheduling processes usually rely on some combination of these variables to estimate the system load, the admissibility of a new user, and the priority when allocating resources to UEs.

The QoS requirements of most applications can be resolved into specifications of BER and data rate, which in CDMA systems correspond to the signal-to-noise-and-interference ratio (SNIR) at the receiver and the spreading gain of the channel, respectively, once the FEC scheme and rate are chosen. If only a single cell is to be considered, Leon-Garcia and Arad [5] showed how the SNR and spreading gain requirement of each user combine to provide a representation of the load presented to the system uplink—the "power index."

If each user i transmits with power P_i and a spreading gain G_i, then the SNR (S_i) at the Node-B is

$$S_i = \frac{G_i h_i P_i}{\sum_{j \neq i} h_j P_j + \eta_0 W} \tag{1}$$

where i and j belong to the set of N users connected to the Node-B at any instant of time; h_i is the path gain between user i and the Node-B; η_0 is the power spectral density of the background noise, which is assumed to have a Gaussian profile; and W is the system bandwidth. If the SNR required to meet the BER requirement of any user i is γ_i, then we must have

$$S_i \geq \gamma_i, \quad \forall i. \tag{2}$$

It can be shown that optimum power allocation is achieved and throughput is maximized when the QoS constraint above is met with equality. The resulting optimum power for any user is

$$P_i = \frac{\gamma_i}{\gamma_i + G_i} \frac{\sum_{j=1}^{N} h_j P_j + \eta_0 W}{h_i}. \tag{3}$$

The factor $\gamma_i/(\gamma_i + G_i)$ is referred to as the power index g_i and reflects the signal-to-total-interference ratio for user i. Herein, we have a measure for user load. Intuitively, if a user wanted to double its SNR requirement (γ_i) without changing the load posed to the system, a doubling of the spreading gain would be required as well, which implies that twice as much energy per data bit would be delivered, i.e.,

$$g_i = \frac{\gamma_i}{\gamma_i + G_i}. \tag{4}$$

Using (3) and (4) to form a set of power constraints for all users in the cell and with some manipulation, the optimum power level for each user can be determined as follows:

$$P_i = \frac{\eta_0 W g_i}{h_i \left(1 - \sum_{j=1}^{N} g_j\right)}. \tag{5}$$

Since all the transmission powers should be positive, i.e., $P_i \geq 0, \ \forall i$, a necessary and sufficient condition for the optimum solution to exist is

$$\sum_{j=1}^{N} g_j < 1. \tag{6}$$

The above equation provides not only a measure of system capacity but a connection admission control (CAC) test as well. The maximum transmission power (Tx power) capability $\hat{P}_i$ of UE i, which limits the system capacity, is then incorporated into this constraint in [5] and [15] as follows:

$$\sum_{j=1}^{N} g_j < 1 - \frac{\eta_0 W}{\min_i(\hat{P}_i h_i / g_i)}. \tag{7}$$

1) Dynamic Range of Output Power—Another Constraint: The constraint imposed thus far on UE Tx power is $0 \leq P_i \leq \hat{P}_i$. However, another constraint is the minimum Tx power of the UE $\check{P}$ to reflect the UEs dynamic range of output power $\sigma_{\max}$. Assuming that all UEs have identical Tx power constraints, then

$$\check{P} \leq \bar{P}_i \leq \hat{P} \tag{8}$$

$$\sigma_{\max} = \hat{P}/\check{P} \tag{9}$$

where $\bar{P}_i$ is the scaled Tx power allocated to user i that meets the dynamic range constraint. From (5) and (9), a new constraint emerges that provides insight to the maximum load

g that can be supported per user and its dependence on the path gain to the Node-B, i.e.,

$$\sigma_{m,n} = \frac{P_m}{P_n} = \frac{g_m/h_m}{g_n/h_n}, \quad \{m,n\} \in N \tag{10}$$

$$\max_{\{m,n\}} \{\sigma_{m,n}\} \leq \sigma_{\max}. \tag{11}$$

Feasible values for a UE dynamic range can be taken from [23], wherein $\check{P} \leq -50$ dBm for a minimum dynamic range of 74 and 71 dB in class 3 and class 4 equipment, respectively. If any optimum power P_j is found to be less than $\check{P}$, every user's output power must be scaled by ϕ to meet the minimum output power constraint, i.e.,

$$\phi = \max \left\{ \begin{matrix} 1 \\ \max_j \{\check{P}/P_j\} \end{matrix} \right\} \tag{12}$$

$$\bar{P}_i = \phi P_i \quad \forall i. \tag{13}$$

This scaling operation has the effect of raising the noise floor in the cell and therefore brings the Tx powers of all users nearer to the maximum power limit. Thus, the capacity of the cell is affected as

$$\sum_{j=1}^{N} g_j < 1 - \frac{\phi \eta_0 W}{\min_i (\hat{P}_i h_i / g_i)}. \tag{14}$$

The result in (10) shows that QoS differentiation could have a significant effect on the coverage of a cell. If all users have identical QoS requirements, i.e., $g_m = g_n$, then the maximum diversity in the UE–Node-B path gains for users m, $n \in N$ in the cell is equal to the dynamic range of the UE output power, i.e., $\max\{h_n/h_m\} \leq \sigma_{\max}$. The UMTS specification allows a range of discrete spreading gains from four to 256 in the WCDMA uplink. Using (4), those limits correspond to a maximum diversity in user load (i.e., $10\log_{10}\{g_m/g_n\}$) of 16.3 and 15.6 dB for SNRs 3 and 5 dB, respectively. Using the path loss model for the "vehicular test environment" in [24], alternative capacity distributions that meet both our constraints are shown in Fig. 2. It is assumed that power and interference measurements are perfect and that there is no fading in the channel. Distances between UE and the Node-B increase geometrically to simulate a "hotspot" environment. In all cases, the spreading gain either remains constant or increases with distance from the Node-B. This can be explained by the result in (10). If power index g increases as path gain h decreases, then g/h is divergent, and the available dynamic range gets utilized faster than when g increases with h. Loss in data throughput due to retransmissions has not been considered for the analysis that follows.

In this paper, we assume that a resource manager is "fair" when UEs with similar applications get similar throughput regardless of the quality of their uplink channels. The "fair allocation" curve in Fig. 2 and in subsequent figures exemplifies a scenario with no load diversity in the system, i.e., all users employ $G_i = 16$, $\gamma_i \approx 3.16$ (5 dB). The cell radius does not exceed 0.59 km, and a total of five users can be accommodated with a total bit rate of 1.2 Mb/s. As we increase load diversity

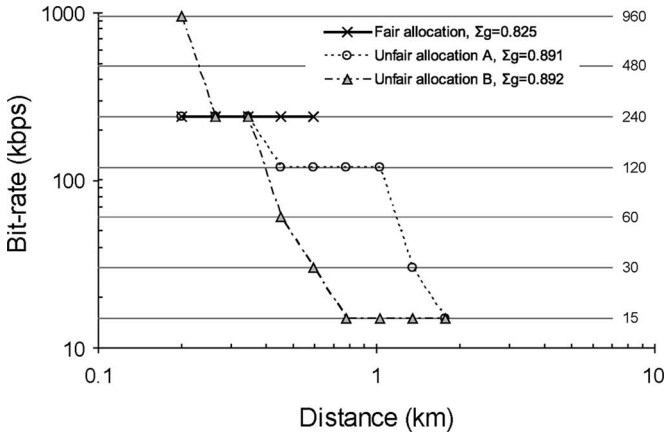

Fig. 2. Example UE uplink bit-rate allocation in a single-cell scenario with SNR $= 5$ dB.

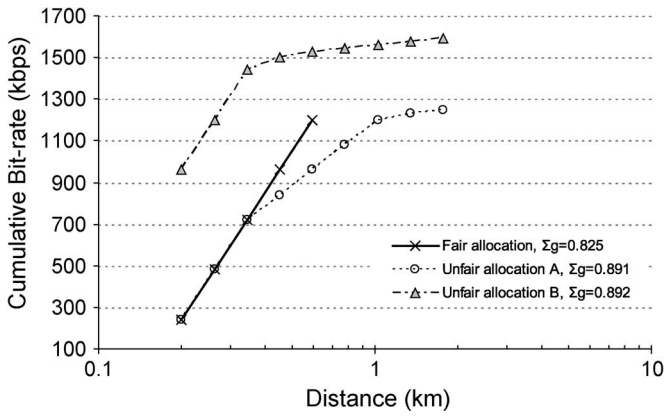

Fig. 3. Example UE uplink cumulative bit-rate allocation in a single-cell scenario with SNR $= 5$ dB.

within the same locations, we find that cell coverage can be extended to include distant users with high spreading gain, as concluded earlier. Interestingly, the cumulative bit rate in the cell also increases when we choose to deviate from fair distribution of available capacity. Although the scenarios in "unfair allocations" A and B have nearly identical cumulative power indices, Fig. 3 shows that they have very different total throughputs over the uplink. Maximizing the cumulative power index (CPI) does not necessarily maximize the cumulative bit rate. That the bit-rate allocations meet the Tx power constraints is confirmed through the individual power allocations in Fig. 4. Fig. 5 summarizes how load diversity is leveraged to increase the path gain diversity among the UEs in the hypothetical deployment. The greater rise-over-thermal (RoT) figure in the unfair allocation scenarios reflects the greater cumulative power in the uplink and, as we shall see later in (18), is due to the greater cumulative power index.

2) Implications for Resource Scheduling Schemes: If the CPI is used in isolation as the CAC criteria, then cell breathing can lead to significant coverage problems, as demonstrated in Fig. 2, where the fair allocation and unfair allocation schemes have similar scores but different coverage areas. Scheduling schemes for sharing uplink capacity must now not only fulfill varying QoS requirements but preserve the coverage area encompassing active UEs as well. However, if maximizing

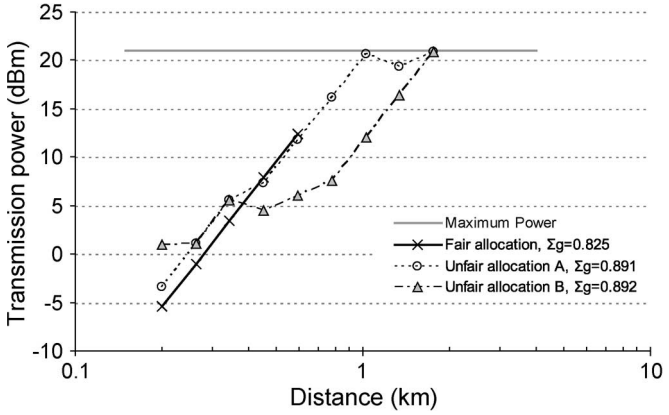

Fig. 4. UE uplink Tx power allocation in a single-cell scenario.

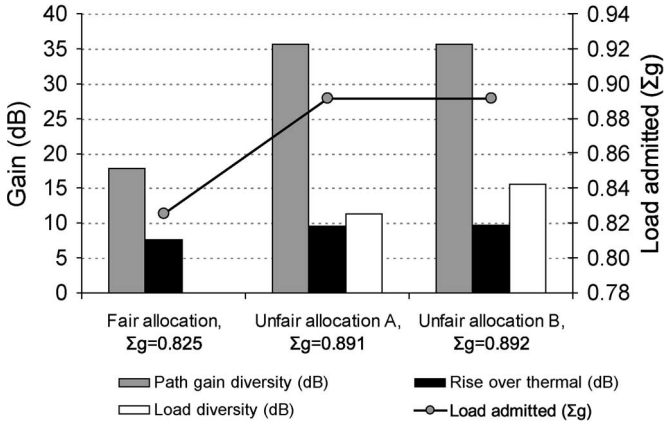

Fig. 5. Effect of load diversity on system throughput and path gain diversity.

cumulative uplink throughput is of the highest priority, the maximum QoS that can be provided to a user becomes severely constrained by the path gain of the communication channel—a constraint that is primarily influenced by the user's distance to the Node-B and mobility.

B. Estimating Optimum Power in a Multicell Scenario

The complexity of determining the optimum Tx power for each UE in a multicell scenario increases with the number of cells in the deployment. A UE compensates for greater path loss with greater Tx power, but in doing so, it also causes greater interference to UEs in neighboring cells. In summary, the optimum Tx power for each UE in the deployment and the uplink capacity in each cell is dependent on the path gain of every UE–Node-B pair. However, approximations can be made to allow for scalable computation of the Tx powers in the same manner as in (3).

The total uplink interference I_k in any cell k due to all the users in the deployment can be assumed to be some factor of the cumulative received powers from users within the particular cell [4], i.e.,

$$I_k = (1 + c_k) \sum_{j=1}^{N_k} h_{j,k} P_{j,k} + \eta_0 W \tag{15}$$

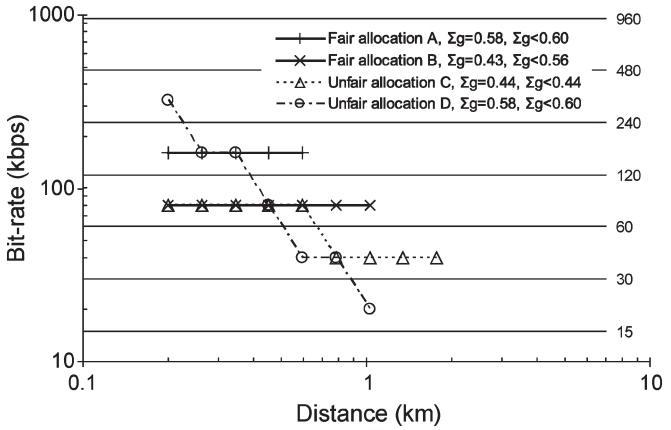

Fig. 6. Example UE uplink bit-rate allocation in a multicell scenario with FEC rate $= 1/3$, $E_b/N_o = 5$ dB, and $c_k = 0.65$. The legend also shows the cumulative load for the configuration (e.g., $\sum g = 0.58$) and the maximum allowable load given that configuration (e.g., $\sum g < 0.60$).

where j is any user in cell k, and c_k is the ratio of other-cell to own-cell interference. The approximated uplink interference I_k can be incorporated into (3) to revise the Tx power and uplink capacity estimations as follows:

$$P_{i,k} = \frac{\eta_0 W g_{i,k}}{h_{i,k}\left\{1 - (1 + c_k)\sum_{j=1}^{N_k} g_{j,k}\right\}} \tag{16}$$

$$\sum_{j=1}^{N_k} g_{j,k} < \frac{1}{(1 + c_k)}\left\{1 - \frac{\phi_k \eta_0 W}{\min_i(\hat{P} h_{i,k}/g_{i,k})}\right\}. \tag{17}$$

Suggested values for c_k are 0.65 and 0.2 in macro cells and micro cells, respectively [4]; the reasoning behind the lower value for micro cells is that street corners provide better isolation. By using different values of c_k for each cell, we can leverage instantaneous soft capacity in the network due to uneven user load distribution. These values can be adjusted periodically or when cumulative received power changes by a certain amount. To visualize the effect of intercell interference on uplink capacity, we look at some hypothetical scenarios for macro cells where, as in the single-cell case, we incorporate path gains and maximum power limits. The effect of FEC through one-third-rate convolutional coding with $E_b/N_o = 5$ dB is also included [25]. We consider four scenarios A, B, C, and D in Figs. 6 and 7, wherein UEs are allocated different bit rates according to their distance from the base station and experience adjacent-cell interference. Comparing Fig. 6 with Fig. 2, as well as Fig. 7 with Fig. 3, demonstrates the reduction in uplink coverage and capacity due to multicell interference for several allocation schemes. While fair allocation B provides greater coverage and accommodates more users than fair allocation A, it does so by halving the data rate to each UE and reducing the total load. Unfair allocation D achieves the highest total throughput by distributing capacity most unevenly.

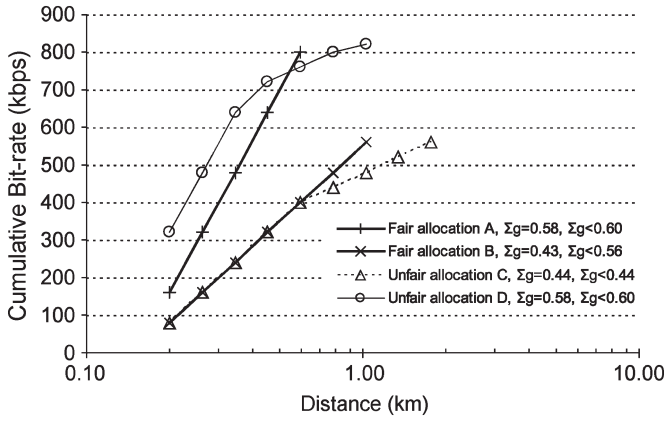

Fig. 7. Example UE uplink cumulative bit-rate allocation in a multicell scenario with FEC rate $= 1/3$, $E_b/N_o = 5$ dB, and $c_k = 0.65$.

RoT is a commonly used indicator of total uplink load and is defined below. The relationship between the CPI and RoT is also shown in the following:

$$\begin{aligned}\text{RoT}_k &= \frac{\text{Total interference and thermal noise in cell } k}{\text{thermal noise}} \\ &= \frac{(1+c_k)\sum_{j=1}^{N_k} h_{j,k}P_{j,k} + \eta_0 W}{\eta_0 W} \\ &= \frac{1}{1-(1+c_k)\sum_{j=1}^{N_k} g_{j,k}}. \end{aligned} \tag{18}$$

An interesting observation from (18) is that when we express other-cell interference as some fraction of the own-cell interference through c_k, and as long as the constraint in (17) is observed, RoT appears to have no relationship with the path gain to the UEs. However, as we shall see in the following section, the position of the UEs has a significant impact on c_k itself, especially when the geographical distribution of users within neighboring cells varies significantly. When we compare fair allocation A with unfair allocation D in Fig. 8, we observe a greater path gain diversity or coverage area in the latter due to greater load diversity. The same applies when we compare fair allocation B with unfair allocation C. However, RoT is similar in both cases. These conclusions match the inferences from (8) and (18).

C. Fair Sharing and Optimization Through Scheduling

In the previous sections, we saw how constraints on CPI and Tx power affect the distribution of capacity among users. When maximizing throughput becomes a priority, it becomes necessary to reduce user bit rates as distances from the Node-B increase. However, when the priority is to fairly distribute capacity within users, the same constraints limit the number of users and the coverage area if all users are to transmit simultaneously. Scheduling allows us to select a subset of active users for simultaneous transmission at any particular instance of time and allows us to vary the spreading gains of each user in a manner that optimizes network throughput while fulfilling the individual QoS requirements.

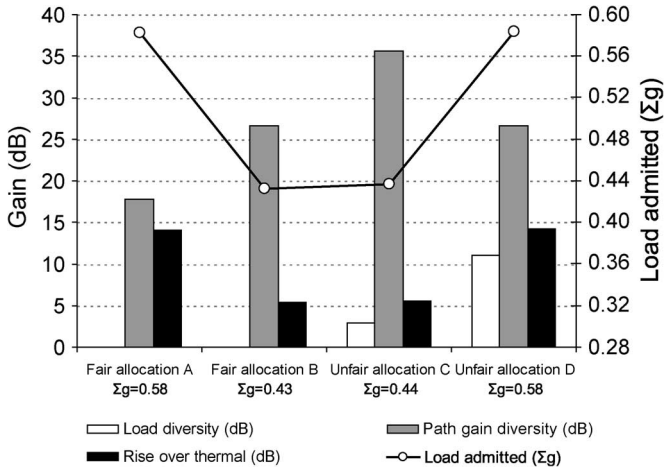

Fig. 8. Effect of load diversity on system throughput and path gain diversity with FEC rate $= 1/3$, $E_b/N_o = 5$ dB, and $c_k = 0.65$.

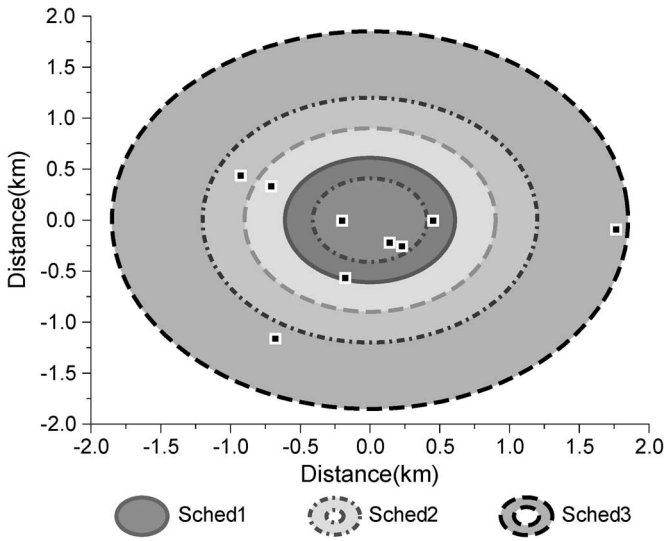

Fig. 9. Separation of users into overlapping ring-shaped subsets according to path gains.

Separating the active users in a cell into subsets and servicing each of them individually reduces the path gain diversity within each subset, which increases cumulative throughput. As before, cumulative throughput is highest when load diversity within users is maximized. Scheduling users in such a manner provides higher bit rates than those permissible if all users in the cell were to transmit simultaneously, albeit only to certain users within each subset. Subsets of users can also be made to overlap such that users get higher throughput on average. Figs. 9 and 10 show how three bit-rate allocation profiles (i.e., Sched1, Sched2, and Sched3) can be defined where a different subset of users gains channel access under each profile, collectively providing greater net coverage. The profiles are activated sequentially in a continuous cycle. When averaged, the three profiles provide an effective spreading gain that does not necessarily match one of the standard UMTS values. The ability to resolve spreading gains more finely through averaging is one of the benefits of scheduling. To illustrate another advantage, let us assume for now that the average spreading gains corresponding to the average bit rates in Fig. 10 were permissible. If all the

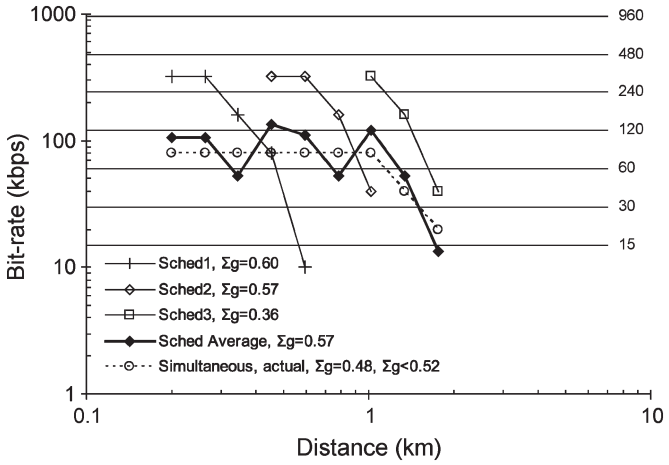

Fig. 10. Sequentially scheduled bit-rate allocation within three subsets of UEs with FEC rate $= 1/3$, $E_b/N_o = 5$ dB, and $c_k = 0.65$.

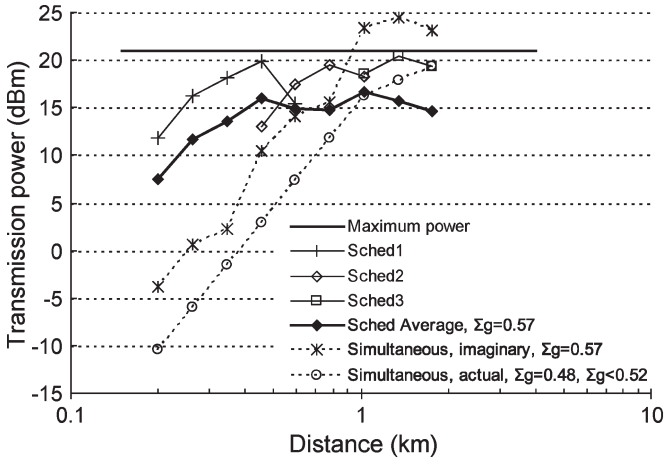

Fig. 11. Comparison of imaginary Tx powers for simultaneous transmission with sequentially scheduled power allocation within three subsets of UEs with FEC rate $= 1/3$, $E_b/N_o = 5$ dB, and $c_k = 0.65$.

UEs in the cell transmitted with these gains simultaneously, Fig. 11 shows with the "simultaneous, imaginary" curve that the Tx power of three UEs would exceed the maximum power limit, whereas the Tx powers during scheduled transmission stay under the limit at all times. The performance of a valid alternative for simultaneous transmission is also shown in Figs. 10 and 11 through the "simultaneous, actual" curves. With a cumulative data rate of 620 kb/s, it provides 130 kb/s less throughput than the scheduled alternative but requires less power. An assumption in the case for scheduled transmission has been that the other-cell to own-cell interference ratio c_k is constant. However, since scheduling periods distinguish UEs by their path gains, and since it is reasonable to assume that with decreasing path gain users are also typically closer to the cell boundary, we should expect c_k to be greater during Sched3 than in Sched1 because of the greater effect on the intercell interference level in neighboring cells. For verification, we simulate a 19-cell deployment wherein each cell has nine users at the distances and with the data rates used in Fig. 10. Path gains are simulated using the vehicular test environment as before and without variation due to fading. Table I gives the interference ratio and RoT figures when the three scheduling

TABLE I
INTERFERENCE RATIOS FOR SCHEDULED SUBSETS

Attribute	Scheduled Subset Sched1	Sched2	Sched3	*Simult. actual*
$\sum g$	0.604	0.565	0.357	0.481
$RoT(dB)$, $c_k = 0.65$	23.8	11.7	3.9	6.8
c_k in simulation	0.001	0.018	0.375	0.108
$RoT(dB)$ in simulation	4.0	3.7	2.9	3.3

subsets were simulated separately, with every cell using the same subset. We observe that the assumption of constant c_k leads to unrealistic estimates of RoT and overly conservative cell capacity estimates. The earlier inference from (18) that c_k does not depend on user position no longer holds because of an artificial deviation in path gain diversity due to scheduling. In real scenarios, similar deviations would also prevail because of slow and fast fading.

D. Alternative Load Estimation and Admission Control Method for a Multicell Scenario

In Section III-A, uplink load due to each UE was estimated independent of its uplink path gain. The role path gains play toward limiting cell capacity is considered during admission control through the term $\min_i(\hat{P}_i h_i/g_i)$ in (7), where h_i is the path gain of user i. In Section III-B, we adapted the load estimation system to take leverage soft capacity in a multicell scenario by approximating the effect on intercell uplink interference. This section looks at an alternative method proposed in [26] for estimating user load and enforcing CAC to include the cost due to path gain and the benefits due to soft capacity in one expression.

The total relative load L_j at base station j due to all the users in that cell is defined in [26] as

$$L_j = \sum_{i=1}^{M} \frac{\beta_i^t h_{ij}}{\sum_{k \in K_i} h_{ik}} \tag{19}$$

where β_i^t is the target carrier-to-total-interference ratio (CTIR) for UE i, M is the set of all UEs with an active connection to j, h_{ij} is the power gain between UE i and Node-B j, and K_i is the set of all Node-B's that have a path gain greater than 0 with UE i. Relative load L is estimated for each Node-B $k \in K_i$ affected by the admission of the new user i, and the user is only admitted if they are all below δ_{sc}—the soft capacity limit parameter. A value of 0.6 for δ_{sc} corresponds to a noise raise of 4 dB.

We thank the reviewer for pointing out that, with some manipulation, power index g_i in (4) is identical to CTIR β_i^t in (19) in a single-cell scenario. Also, unlike in (17), where the coefficient c_k is used to model intercell interference, the relative load concept includes intercell interference dependencies at the cost of greater signaling and processing overheads for several more network elements, i.e., Node-B's or RNCs.

For this system to be effective, the pilot powers received by a UE from all Node-B's in K_i must be communicated to the admission controller periodically for the estimations of path gain ratios $h_{ij}/\sum_{k \in K_i}(h_{ik})$ to hold sufficient accuracy in between refresh intervals. The rate of reporting these measures

associated with event-driven handover procedures in UMTS is found to be sufficient. Since every Node-B in K_i must also find the load of the new user to fit within the load budget (δ_{sc}), Gunnarsson *et al.* propose for scalability that only the six base stations that register the strongest pilot powers with the mobile form set K_i, as also recommended in [27].

E. Power Control

Power is controlled in several ways for a call in WCDMA. The Tx power for the initial admission request is computed according to the received power of the Node-B pilot. Later, the power used to begin data transmission is determined according to the total uplink noise and SNR target. These constitute "open-loop power control." Once data transmission begins, UMTS incorporates "closed-loop fast power control" to counter the effects of slow and fast fading and also to meet the target SNR [4]. Transmitting with lower power would lead to a greater number of reception errors. Tx power that is higher than necessary deteriorates the call quality of several other users in different cells. We use a fast power control step size of 1 dB at 1500 Hz according to the UMTS specification in our simulation model.

It has been argued that consistency in QoS is perhaps more important than the absolute level of quality itself. Since fading and mobility can cause channel conditions to deteriorate significantly, steps must be taken to maintain quality in such circumstances. One of several building blocks toward achieving this goal is outer-loop power control (OLPC), wherein the target SNR is adjusted after every frame received at the Node-B such that the mean FER is maintained at a predetermined value. In this manner, outer-loop control enables diversity in the QoS provided to users. The OLPC logic for each user i is as follows:

$$\mathrm{Snr}_{i,n} = \begin{cases} \mathrm{Snr}_{i,n-1} + \delta_{\mathrm{olpc}}, & \text{if } \mathrm{Nack}_{n-1} \\ \mathrm{Snr}_{i,n-1} - \delta_{\mathrm{olpc}} \cdot \frac{\mathrm{Fer}_i}{1-\mathrm{Fer}_i}, & \text{if } \mathrm{Ack}_{n-1} \end{cases}$$

where $\mathrm{Snr}_{i,n}$ is the target SNR computed for user i in frame n; Fer_i is the target FER to be maintained for the call; δ_{olpc} is the step size for outer-loop SNR changes, e.g., 0.5 dB; and Nack_{n-1} or Ack_{n-1} is true if the previous frame is received with or without error, respectively.

IV. Congestion Management and Quality of Service

In the interest of service differentiation, network congestion is interpreted differently for different service types. Consequently, actions required to alleviate congestion in a heterogeneous service environment differ as well at any point in time. "Best effort" services, by their very definition, give way to quality-controlled services such as voice and video traffic during periods of heavy load. A greater deterioration in the quality of video is acceptable when compared with that of the voice stream associated with the same video-enabled call. The CAC ensures that no more users are admitted into the system if the QoS to active users were to drop below a minimum threshold. However, separate processes are required for controlling congestion due to ongoing traffic and in a manner that considers the value associated with each service type. The purpose is to reduce the load imposed on the network by every user during congestion, and this action is often referred to as "backoff." Not only must the extent of backoff be linked to the service value, as mentioned previously, but it must also consider the difference in pending loads within users of the same service. The need for an interservice and intraservice prioritization scheme is highlighted here and will be discussed later.

Congestion control can be enforced unilaterally by a Node-B or an RNC connected to several Node-B's. In such a scenario, UEs must relay some information about their pending loads to the controller for it to properly distribute the available channel capacity. Packet buffer sizes could be relayed toward this end or even the required SNR if the UE is tasked with QoS management. The ability to control congestion in a distributed manner aids scalability. The resource manager in a UE could make informed decisions on the extent of backoff in its active sessions. A distributed controller would not only reduce the computation required centrally but could also reduce the amount of signaling that is inherent in a centralized control system. Since the signaling medium is itself vulnerable to channel impairments, a reduction in signaling is beneficial in many ways.

Congestion avoidance plays a significant role in WCDMA systems as well since it can be incorporated into multiple layers of the protocol stack. As we shall see later on, the target FER for each service and the SNR associated with it impacts the total number of admissible users significantly. Increasing the FER target allows retransmission schemes to leverage time diversity in the channel and has been shown to dramatically improve cell capacity, with the optimum value ranging from 10% to 30%, depending on the velocity of the mobile [4]. To match the dynamic rate of packet arrival in video or data services, multicode or VSG transmission mechanisms can be used to minimize resources use while meeting QoS. Congestion avoidance can also be incorporated into the process of selecting the number of parallel transmissions or the spreading gain of the transmission. This could be thought of as traffic shaping at multiple-access control (MAC) level. The method of incrementally upgrading the number of codes or the spreading gain to reach the desired value is found to be more resilient in the presence of high interference and provides for greater cell capacity when compared with the method of directly using some value and holding it for the duration of a packet [17].

The proposed resource management framework for a WCDMA uplink incorporates these fundamental principles into a distributed control system, which aims to meet QoS with VSG and with the cooperation of all entities. The system comprises of a transmission rate scheduler, a traffic-class-based admission prioritizer, a frame admission controller, and an outer-loop power controller, as shown in Fig. 12. They are described below. System performance is verified through extensive simulations. The simulation model and results are presented in later sections.

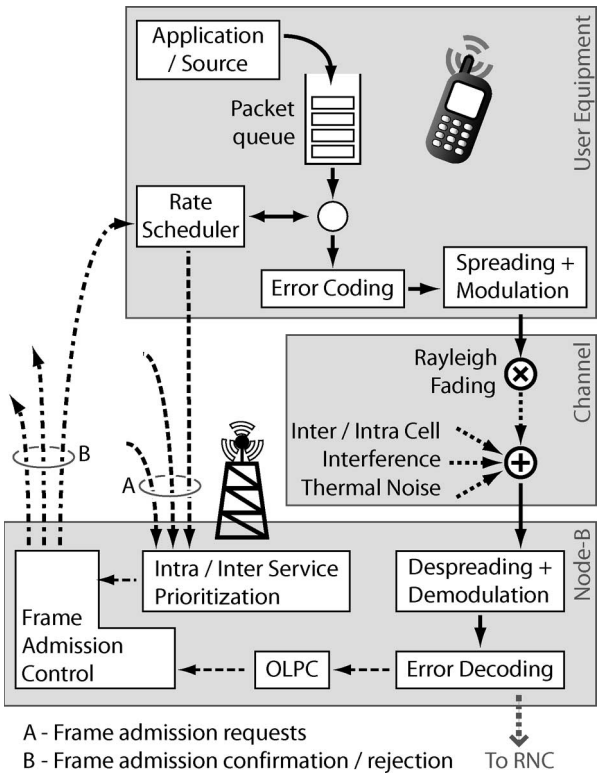

Fig. 12. RRM components for UMTS uplink and simulation design.

TABLE II
ADJUSTING t_{left} AFTER THE ARRIVAL OF A NEW PACKET

if	$q_{n-1} = 0$,	$t_{left,n} = t_{target}$
else if	$\frac{q_{n-1}}{t_{left,n-1}} > \frac{q_n}{t_{target}}$,	$t_{left,n} = t_{left,n-1} \cdot \frac{q_n}{q_{n-1}}$
else		$t_{left,n} = t_{target}$

TABLE III
STATE DEFINITIONS

Rate States		Throughput States	
R_1 :	$Req_n \leq 0.5 \cdot Rate_{n-1}$	T_1 :	Previous frame erroneous, or $Power_{current} \geq Power_{max} - 1dB$
R_2 :	$0.5 \cdot Rate_{n-1} < Req_n$ $Req_n < u \cdot Rate_{n-1}$	T_2 :	Previous frame received successfully, and $Power_{current} < Power_{max} - 1dB$
R_3 :	$u \cdot Rate_{n-1} \leq Req_n$		

A. QoS-Based Rate Scheduler

Spreading gain can be varied in UMTS to accommodate more application data in one transmission frame and increasing the spreading gain also results in energy savings at the UE. However, when the QoS requirements of a particular application must be met, spreading gain primarily depends on the transmission buffer size and the target packet transmission delay. Since our focus is on the uplink, we provide the rate scheduler access to this information through minimal signaling overhead by locating it within the UE.

TABLE IV
OUTPUT OF RATE SCHEDULER

Throughput State	Rate State R_1	R_2	R_3
T_1	Req_n	$Rate_{n-1}/2$	$Rate_{n-1}/2$
T_2	Req_n	$Rate_{n-1}$	if $Power_{current} \cdot 2 < Power_{max} - 1dB$ then $Rate_{n-1} \cdot 2$ else $Rate_{n-1}$

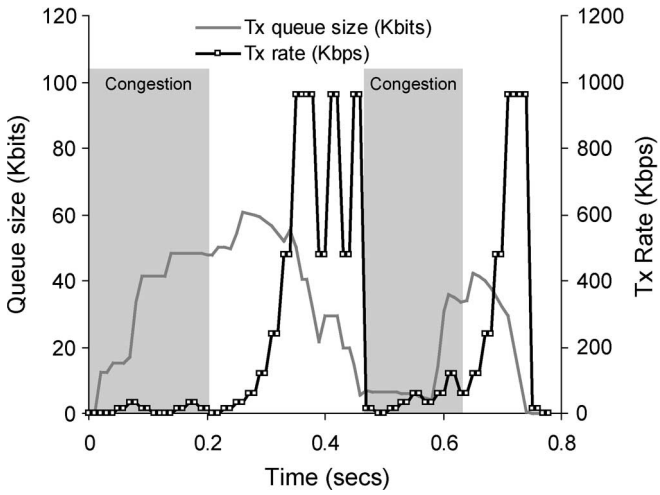

Fig. 13. Rate scheduler behavior for a user in a heavily loaded cell. Scheduling period: one TTI; admission delay: one TTI.

TABLE V
CONTROLLER LOCATION

		Capacity estimation method: CPI	RLE
Admission	*limit-fixed*	Serving Node-B	Multiple Node-Bs or RNC
	limit-free	Controller in UE, Load feedback broadcast by the serving Node-B	

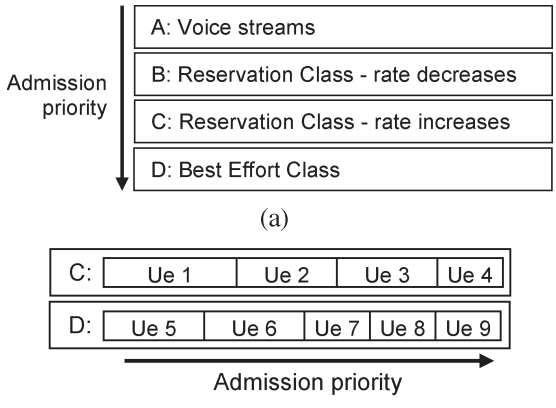

Fig. 14. Prioritized admission. (a) Interservice priority: Since transmission rates allocated to best effort class terminals expire after each scheduling period, there is no need to process rate decreases for such terminals separately. (b) Intraservice priority: For each class, terminals requesting higher transmission rates are processed for admission before those requesting lower transmission rates.

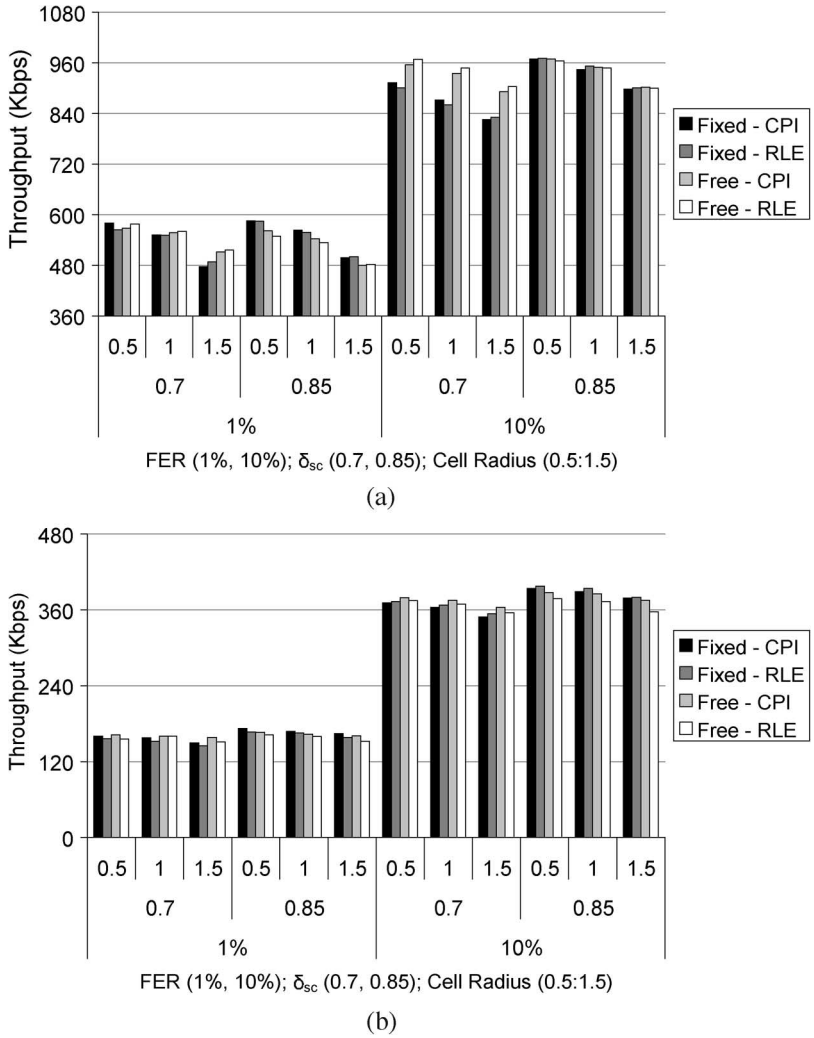

Fig. 15. Maximum cell throughput of application level data packets when the maximum load is 1200 kb/s per cell for different UE velocities. (a) 3 km/h. (b) 50 km/h.

The rate scheduler must adapt with the quality of the radio channel at the MAC level in addition to observing QoS when choosing the best spreading gain or transmission rate for the next scheduling period. This improves the probability of successful reception of future transmissions. Since transmission buffer sizes, channel conditions, and total transmission load in the uplink are dynamic, transmissions must be rescheduled periodically at possibly a new spreading gain to maximize the rate of success while meeting transmission deadlines. In this paper, we set the scheduling period at 10 ms i.e., the minimum interval at which spreading gains can be varied in the UMTS uplink, which is also referred to as the transmission time interval (TTI). Improving the performance of the UMTS uplink by reducing the TTI to an even smaller duration of 2 ms, in addition to other means, has already been proposed and evaluated in [28]. This reduction, in combination with other improvements as part of the enhanced uplink (EUL) or high-speed uplink packet access (HSUPA) effort, has been shown to increase system throughput [29], [30].

The transmission rate required Req_n for the scheduled period n is estimated by dividing the input buffer size q_n by the permissible delay for clearing the buffer while compensating for the mean admission delay (t_{adm}) experienced over past admission requests. It is assumed that packets that arrive before the admission request is approved will not significantly alter the bit-rate requirement for the next frame, i.e.,

$$\mathrm{Req}_n = \frac{q_n - t_{\mathrm{adm}}\mathrm{Rate}_{n-1}}{\max\{t_{\mathrm{left},n} - t_{\mathrm{adm}}, t_{\mathrm{TTI}}\}} \tag{20}$$

where $t_{\mathrm{left},n}$ is the "time left" to transmit all packets in the data buffer and still meet QoS in any frame n. This variable is decremented by t_{TTI} (10 ms) after every TTI. It is also adjusted when a new packet arrives at the input buffer, so that the final value reflects the urgency for transmission of either the packets already queued in the buffer or that of the new packet, choosing whichever is greater, as shown in Table II. The target time period for clearing the buffer is therefore a function of the "target transmission delay" (t_{target}) that matches the QoS requirements of the application being serviced.

Req_n is the ideal required transmission rate and has not been scaled for backoff due to channel conditions or system load. It might not match one of the standard transmission rates in the UMTS uplink either. The final output of the rate scheduling

process is the value Rate_n, which takes all of this into account. The selected transmission rate (Rate_n) is made to depend on the following factors:

- the ratio of the previous bit rate (Rate_{n-1}) and the required data rate (Req_n);
- the throughput in the previous interval (Thput_{n-1});
- the rate adaptation threshold $(u:\ 1 < u \leq 2)$.

The Thput term has two states and is dependent on the successful reception of the previous frame and also the proximity of the output power to the maximum permissible level. It is therefore representative of instantaneous channel quality and cumulative load in the uplink. The transmission rate in the previous frame Rate_{n-1} is compared with Req_n, and the change is classified into three states. It is representative of the urgency in clearing the transmission buffer if QoS is to be maintained. The states are summarized in Table III.

Having characterized the urgency to transmit and the channel condition, the rate scheduler is now able to select the transmission rate. Table IV shows the output of the rate scheduler (Rate_n) for each combination of "rate state" and "throughput state." The transmission rate is only upgraded in state $\{T_2, R_3\}$. A reduction in transmission rate is called for in states $\{T_1, R_2\}$ and $\{T_1, R_3\}$, although the buffer size translates to zero or positive change. This is the implementation of the backoff procedure when the channel is poor or cumulative load is too high. Standard transmission rates in UMTS double from one to the next while spreading gains halve. The congestion avoidance step is in state $\{T_2, R_3\}$, where transmission rates are increased one at a time. Reductions in spreading gain over several set points at once would have to be matched by proportionately large changes in the output power of the UE to maintain SNR. This has the potential of disrupting the transmissions of several other UEs. A stepped approach on the other hand gives other UEs more time to track SNR levels through fast power control. The behavior of the rate scheduler for one user in a heavily loaded cell is shown in Fig. 13. The aim of the scheduler is to clear the transmission queue within the target transmission delay of 0.3 s, but the figure demonstrates how the transmission is delayed due to congestion.

B. Frame Admission Controller

Simply allowing the rate scheduler in each UE to converge to a stable operating point would allow for several undesirable saturation conditions to affect users within a cell and those in neighboring cells as well. We have shown that the rate scheduler encourages users with better channel conditions to upgrade their transmission rates sooner. This implies that coverage can be limited to a small percentage of the cell population during periods of high load if the spread in path gains is large. As the cumulative load varies due to scheduling, cell coverage can vary significantly in a span of a few frames, exacerbating the cell-breathing problem. Allowing users to search for the maximum throughput point also implies that an unknown amount of capacity is left available for future callers in a cell or for handovers from neighboring cells. Frame admission control strategies similar to the call admission techniques presented in Sections III-B and D can rectify this problem. The need for such methods would not arise if all users operated with constant throughput rates and when the cumulative load in the cell is well under capacity. It is when we want to provide QoS for applications with bursty output and QoS constraints, as well as when we want to maximize cumulative throughput in the cell, that we must consider admission control for scheduled changes within a call as well.

TABLE VI
DATA TRAFFIC SIMULATION SET

Constants	
Number of cells	19
Base station height	15 m
Maximum UE power	24 dBm
UE output dynamic range	70 dB
Thermal Noise	-107 dBm
Simulation time	70 s
Cells monitored for statistics	1, centre cell
Warm-up period	10 s
Rate upgrade threshold	1.5
UE priority	Best Effort
Mean UE data rate	60 Kbps
UE packet size distribution	exponential
Mean packet size	2400 bits
UE inter-arrival rate distribution	exponential
Mean inter-arrival time	0.04 s
Error coding	1/3 soft-decision convolutional coding [25]
Frame admission delay	1 TTI
Schedule interval	1 TTI
ARQ	enabled
Variables	
Circumscribing cell radius	0.5, 1.0, 1.5 km
Capacity estimation method	CPI, RLE
Admission controller location	limit-fixed, limit-free
Soft capacity limit δ_{sc}	0.7, 0.85
FER	1%, 10%
Target transmission delay	0.1, 0.3, 0.5 s
Mobile velocity (for Rayleigh fading)	3, 50 kph
Total simulation configuration combinations	*288*
Numbers of UEs per simulation configuration	2 to 20 users per cell (120 to 1200 Kbps)
Cell load step size	2 users

To this effect, we use the two measures of user load presented before in alternative frame admission control strategies. In the first, we limit the CPI in each cell for all scheduled rate changes, where the power index of each user is according to (4), such that the condition in (21) holds. We refer to this as the CPI method in the simulations. In the second, we extend the cumulative relative load estimate (RLE) limit in (19) for within-call frame admission in addition to CAC, such that (22) holds. We refer to this as the RLE method henceforth. In the simulation scenarios presented later on, we evaluate system performance for two different values of δ_{cpi} and δ_{rle}, i.e., 0.7 and 0.85, such as

$$\sum_{j=1}^{N} g_j < \delta_{\text{cpi}} \tag{21}$$

$$L_j < \delta_{\text{rle}}. \tag{22}$$

1) Controller Location and Signaling Load: Depending on the location of the frame admission controller, signaling requirements can vary drastically. A centralized controller at

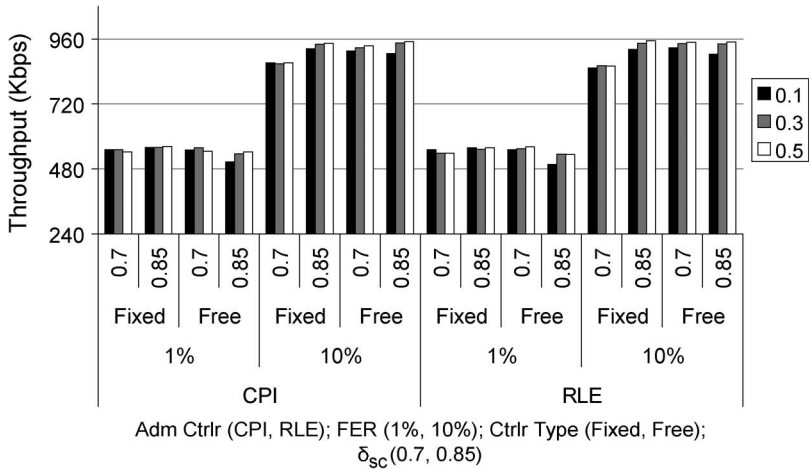

Fig. 16. Effect of target transmission delay on system throughput: Target transmission delay $= \{0.1, 0.3, 0.5$ s$\}$, cell radius $= 1$ km, and mobile velocity $=$ 3 km/h.

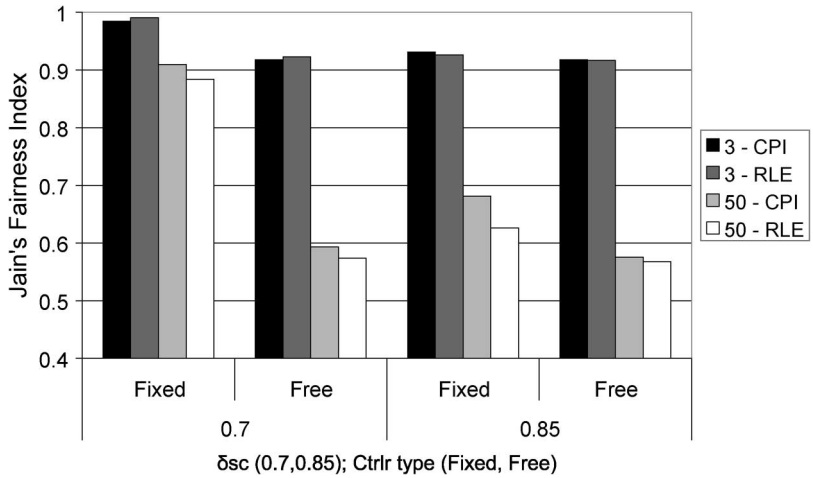

Fig. 17. Jain's fairness index for the different resource management strategies: Cell load $= 1200$ kb/s, mobile velocity $= \{3, 50\}$, cell radius $= 1$ km, target FER $= 10\%$, and target delay $= 0.1$ s.

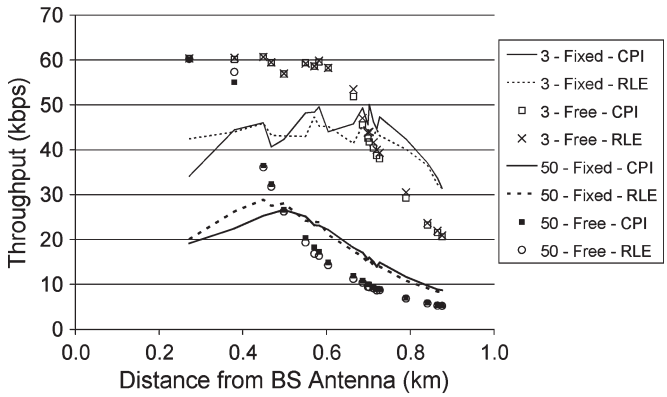

Fig. 18. Variation in mean UE throughput with increasing distance from the base station.

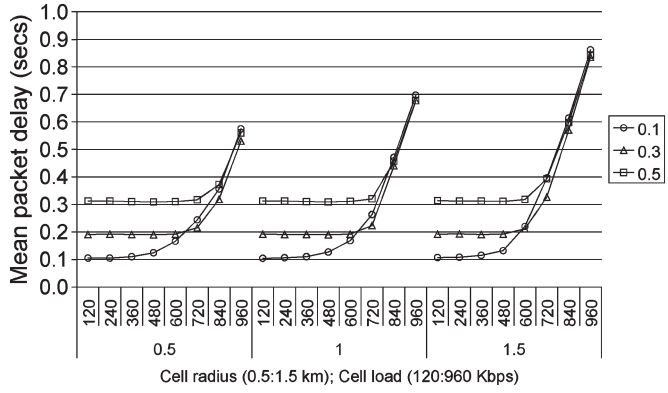

Fig. 19. Control over mean packet transmission delays demonstrated with different delay targets {0.1, 0.3, 0.5 s}, cell radius $= \{0.5, 1.0, 1.5$ km$\}$, and cell load of 120–960 kb/s.

the Node-B would be able to maintain a hard limit on $\delta_{\rm cpi}$ and $\delta_{\rm rle}$ since it can monitor the cumulative intracell and intercell load levels. We refer to this as the "limit-fixed" controller and apply it to both CPI and RLE for comparative evaluation. The performance of the fixed-limit controller for RLE is evaluated only as a benchmark since the signaling load associated with gaining approval from several base stations for all rate changes is prohibitively high. Centralized frame admission control for CPI is, however, feasible since only one base station is required to approve or reject the rate request. An alternative location for the frame admission controller could be within the UE itself (see Table V). The Node-B could transmit cumulative uplink

load levels over a common downlink signaling channel for all UEs within the cell to monitor. For a basic system, the UEs could decide on rate upgrades assuming that the net effect of changes due to other users in the cell would amount to zero. Therefore, we would expect overshoots and undershoots in the cumulative uplink load about the target δ_{cpi} and δ_{rle} values. We refer to this as "limit-free" control. The reduction in signaling from the limit-fixed control method is significant, making this a feasible congestion control scheme for WCDMA systems.

2) Service Prioritization for Frame Admission: The need for interservice and intraservice admission prioritization was highlighted earlier. The frame admission control process provides a point of enforcement of hierarchies in service priority such that high-value services perform better during congestion in uplink traffic. Interservice priority queuing can be enabled by having parallel queues for admission requests from all users sorted according to their service types. Intraservice prioritization is also possible because the rate scheduler escalates the rate requirements of users differently to match the urgency in clearing their transmission queues. The transmission rate itself can therefore be used as a measure of intraservice priority at the time of frame admission. These strategies are summarized in Fig. 14.

V. Simulation Model and Results

We use discrete-event simulations to evaluate the performance of the proposed resource management strategies. Our model simulates a 19-cell deployment with omnidirectional base stations located in the center of each cell. We simulate an ARQ mechanism with FEC using a one-third-rate convolutional code ($K = 7$, soft decision decoding) [25]. Although turbo-coding schemes would improve throughput performance, we use the convolutional coder for ease of implementation and because our focus is not on the absolute values of the performance measures rather the differences between alternative resource management strategies. Similarly, the inclusion of packet interleavers, sectorized antennas, and rake receivers would also improve cell throughput significantly. The received SNR in each slot is mapped to a BER, which leads to a frame error probability computation, where there are 15 slots to a frame (or TTI). Erroneous frames are flagged using a uniform random process [31] referenced to the frame error probability. UEs are made to retransmit those frames in the next TTI. Since our simulation model does not include cell handover, we do not simulate slow fading, but only Rayleigh fast fading using Jakes' model that has been modified to generate multiple uncorrelated fading waveforms for different UEs [32]. Line-of-sight path gains are estimated according to the vehicular test environment model in [24]. Users are distributed uniformly across the cell area, and user positions are identical for the different resource management configurations. Only the center cell performance is measured since it experiences the largest intercell interference. Simulations last 70 s (i.e., 7000 frames or TTIs), but the first 10 s do not contribute toward the results stated to allow the system to stabilize.

TABLE VII
Multiservice Simulation Set

Constants	**Application**		
	Speech	Video	Data
Proportion	4	1	1
Admission Priority	Voice	Reservation	Best Effort
Packet size distribution	constant	VBR MPEG-4 trace	exponential
Mean packet size (bits)	150	2400	
Bitrate (kbps)	15	60	
Mean inter-arrival time (secs)	0.01	0.04	
Error coding	1/2 convolutional	1/3 convolutional	
Mean talk period	1 s	Not applicable	
Mean silence period	1.35 s	Not applicable	
Voice activity distribution	exponential	Not applicable	
Mobile velocity (kph)	3		
Target trans. delay (secs)	0.1		
FER	1%		
Capacity estimation	CPI		
Simulation time (secs)	300		
Cell radius (km)	1		
Variable			
Soft capacity limit δ_{sc}	0.7, 0.8, 0.9		
Total simulation configuration combinations	3		
Number of UEs per simulation configuration	6 to 42 users per cell (145 kbps to 1015 kbps)		
Cell load step size	6		

A. Single Service Simulations—Data Traffic

In this section, we explore the multiplexing efficiency and fairness of our resource management strategies by simulating data traffic only. The following section looks at the performance of multiservice traffic over a system whose parameters have been narrowed down with the help of the findings here. Fig. 15 shows the maximum throughput for different combinations of all the "variable" parameters in Table VI. The legend shows that each column in the graph represents a particular combination of controller location (Fixed, Free) and capacity estimation method (CPI, RLE). The x-axis shows the combination of cell radius (0.5, 1.0, 1.5 km), load limit δ_{sc} (0.7, 0.85), and FER (1%, 10%). The large reduction in cell throughput due to greater mobile velocity is evident in the difference between Fig. 15(a) and (b). As mentioned before, we make no effort to optimize throughput in the presence of Rayleigh fading. However, it is evident from both graphs that in the presence of FEC, a target FER of 10% gives significantly better throughput than a target of 1% because of the lower target SNR, which is in agreement with [4]. It is also clear that there is little change in maximum throughput when the capacity estimation method changes from CPI to RLE. This figure presents the mean of the maximum throughput for the three target transmission delays. This is done to reduce the number of data points for a more concise presentation. Fig. 16 shows that target transmission delay has little effect on the maximum throughput.

Fig. 15(a) also shows that the maximum throughput is lower for the limit-fixed controller when compared with the limit-free controller for $\delta_{\mathrm{sc}} = 0.7$. This difference disappears when we increase the load level to 0.85. However, the gain in throughput

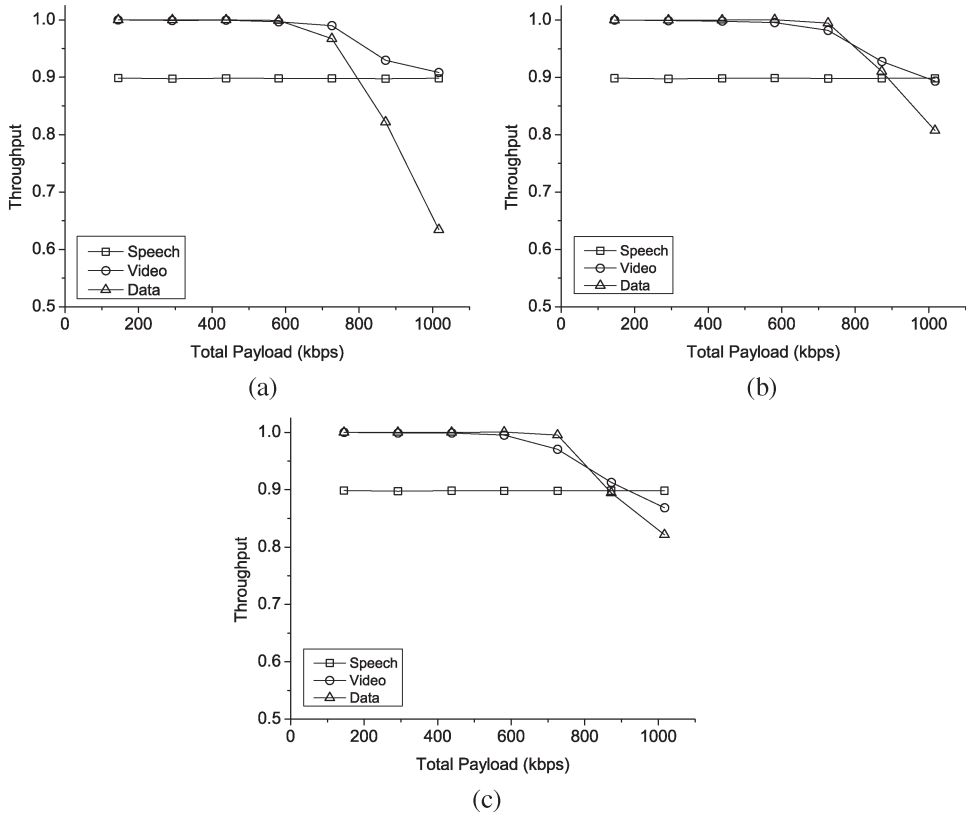

Fig. 20. Average service throughput for different load limits δ_{SC}. (a) 0.7. (b) 0.8. (c) 0.9.

comes with a greater unfairness in allocation of resources among users. This is because increasing the load limit leads to an increase in the "noise rise" in the cell. Since all UEs must transmit within their power constraints, more distant UEs are forced to operate with higher spreading gains to meet their SNR targets when the maximum power constraint restricts them from competing fairly at that noise level. We quantify the fairness in resource allocation for the different simulation configurations under overload (1200 kb/s) using Jain's fairness index [33] to compare the throughput rates of UEs (see Fig. 17). The cell radius is set to 1 km, the target FER to 10%, and the target transmission time delay to 0.1 s. Note that the fairness is greatest when the fixed controller is in operation with a load limit of 0.7 for both mobile velocities. Fig. 18 shows fairness in mean throughput across all users with the same parameters as in Fig. 17 but only for the lower load limit of 0.7.

Since our simulations operate only with a frame admission controller and not a CAC, users are allowed to compete for system resources, although their admission would force the system to operate with a potential load that exceeds capacity. This is done for the dual purpose of realizing the system capacity and for revealing the comparative differences in behavior of the different system configurations when at or beyond capacity.

When operating with the limit-free control mechanism, UEs do not benefit from the interservice or intraservice prioritization of frame admission requests that is available in the centralized fixed control scheme. The rate scheduler has the tendency to probe the channel for better transmission rates by increasing the bit rate one step at a time. When all users attempt to upgrade their transmission rates, nearer users have a greater likelihood of meeting success and moving on to even higher rates. Distant users are more likely to fail and are affected by the backoff procedure in the rate scheduler with greater bias. Therefore, while the goal of meeting target cumulative load limits in a distributed control environment is achieved in the limit-free mode, the system capacity ends up distributed unevenly among all the users in the cell. It is possible to append a measure of average throughput to the average cumulative load measure broadcasted by the Node-B in the limit-free control mode to enable the distributed rate schedulers to prioritize their rate upgrade requests, thereby introducing fairness. However, the centralized limit-fixed mode of control with the CPI capacity estimation method is found to perform at par with the alternative RLE method. This is a positive result because the signaling requirements in CPI are significantly lower than in RLE not only over the fixed infrastructure but across the air interface as well. If the uplink and downlink channel quality is correlated and the FER requirements of UEs change infrequently, then the Node-B only requires a few bits of information from UEs to prioritize and confirm frame admission requests. The success of an admission request could also be conveyed with little overhead across the downlink DCH signaling channel.

The rate scheduler has also been found to provide consistent performance in meeting target transmission delays under a wide range of cell loads in the presence of fast fading. Fig. 19 shows the average application-level packet delays for exponentially distributed packets sizes and interarrival times for the

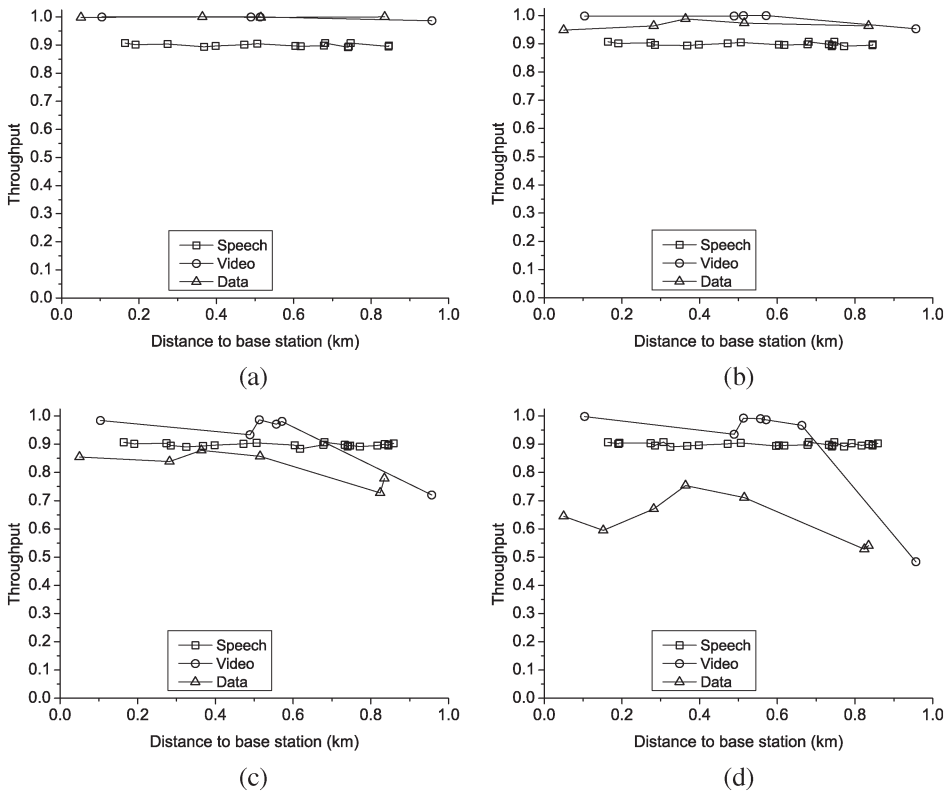

Fig. 21. UE throughput versus distance for different application types and for varying user loads. (a) Twenty-four users per cell. (b) Thirty users per cell. (c) Thirty-six users per cell. (d) Forty-two users per cell ($\delta_{\rm sc} = 0.7$).

fixed-CPI controller with $\delta_{\rm sc} = 0.7$, target FER $= 10\%$, and mobile velocity at 3 km/h.

B. Multiservice Simulations

In this section, we present results indicative of system performance in a multiservice environment by simulating different numbers of mobiles carrying speech, video, or data traffic for 300 s. The limit-fixed frame admission system with CPI capacity estimation is considered since the earlier section showed this combination to provide a good perfomance with low signaling overhead. The proportion of speech to video to data users is kept constant at 4:1:1. Simulation configuration parameters different to those used in the earlier section are presented in Table VII. Fig. 20 shows the average service throughput for mobiles in the center cell of a 19-cell layout and the variation in throughput with the total load over the air interface. In Fig. 20(a), we can clearly see the best effort class data traffic user giving way to other users. We also note that speech users are unaffected at all times despite heavy congestion in the uplink. This is a positive result that supports the radio resource management (RRM) strategies presented here and clearly demonstrates the robustness of the system in the presence of congestion. As the load limit $(\delta_{\rm sc})$ is increased from 0.7 to 0.8 and 0.9 [Fig. 20(b) and (c)], UEs carrying video traffic are affected to a greater extent while data UEs improve their performance. However, speech users remain unaffected. This is because while video UEs have been assigned a higher priority, schedulers in these UEs still expect to empty their buffers with a momentary surge in throughput as depicted in Fig. 13. With the increase in noise rise associated with greater uplink traffic, the fixed-load admission controller denies video UEs of high SNR channels, making it harder for the mobiles to meet their target transmission delays. Therefore, as noise rise increases the net performance of video and data UEs will become increasingly similar. Speech UEs belong to a separate class of traffic of the highest priority because (see Fig. 14) they maintain QoS in all the scenarios presented here. The throughput of 0.9 for speech users corresponds to the FER target of 10%. Since speech traffic is not supported with ARQ, erroneous packets are not retransmitted, and the net throughput is wholly reliant on the target FER and OLPC.

In Fig. 21, we see the scheduler in operation for a load limit $(\delta_{\rm sc})$ of 0.7 and, in particular, the effect it has on UEs of different application types with increasing distance from the base station. As the number of users increases from 24 (580 kb/s) to 42 (1015 kb/s), we observe that data terminals as a group lose throughput to let higher priority services like video and speech reach their QoS targets. However, the most distant video UE suffers a significant reduction in throughput, to the point where it would meet the outage criteria in the 3GPP specification TS 23.107 and the criteria for dropping a call. This highlights the scheduler's inability to resolve fairness among users of the same application when their target transmission delays are identical. Nevertheless, this is a desirable outcome in the presence of congestion because QoS experienced by video users is polarized with throughput being at extremes. Since applications based on video streams typically do not perform well in the presence of significant packet delays or packet error rates, it is better to have fewer users with good

QoS than more users with poor quality. The scheduler's fairness is evident in Fig. 21(d), where data UEs as a group are forced to reduce their throughput due to congestion to let applications of greater priority experience better QoS. Therefore, while the cell breathes with increasing net user load, it does so in a traditional sense only for video users. Data users are affected fairly everywhere in the cell, and speech users are affected last and the least. When the system is made to operate at a higher load limit ($\delta_{\rm sc} = 0.8$ or 0.9), the throughput–distance relation of data UEs is found to approach that of video UEs. The reason for such behavior was explained earlier in this section. Therefore, the need for spare capacity over the air interface is highlighted to fully exploit the service-sensitive multiplexing properties of the proposed resource management system.

We have presented average delay statistics so far because the resource management strategy has been designed to track the average value. 3GPP specification TS 23.107 maps the QoS requirement for streaming traffic to a maximum 95th percentile delay of 250 ms for service data units (SDUs), which are size limited to 12 016 bits per SDU [34]. This does not detract from our contribution because the target transmission delay $(t_{\rm target})$ is simply an input parameter for the scheduler to work with. We leave it to higher protocol layers to match the application-specific 95th percentile delay maximum to an average delay target as much as authors of the 3GPP TS 23.107 specification note that applications that require maximum delay guarantees at the 99.9th percentile must be carefully mapped to a 95th percentile delay maximum. Fig. 22 demonstrates how the proposed scheduler performs at the 95th percentile for packet transmission delays with the same configuration represented by Fig. 21(b). Although we present delay statistics for application data units (ADUs) and not size-limited SDUs, we note that all but the most distant video user conform to the 250-ms delay limit at the 95th percentile when $t_{\rm target} = 0.1$. If SDU delays were to be considered, the perceived delay perfomance would have been even better. ADUs are size limited by the size of the transmission buffer, which can hold 1 s of data at the mean data rate. We also demonstrate that the simulation duration of 300 s is sufficient for the purpose of demonstrating the scheduler's ability to prioritize resource allocation by application by comparing results for 300- and 1800-s simulations with identical configuration parameters. Note that users of all applications were set with $t_{\rm target} = 0.1$. In the presence of congestion and with that delay target, the scheduler outperforms for speech users, right performs for a majority of video users, and underperforms for best effort data users. For the purpose of accurately representing intercell and intracell interference levels in the uplink, the 300-s duration proves to be more than adequate. Fig. 23 demonstrates the mean uplink load estimate for all 19 cells in the simulation model and their 99% confidence limits, given a sample size of 30 000 simulated frames.

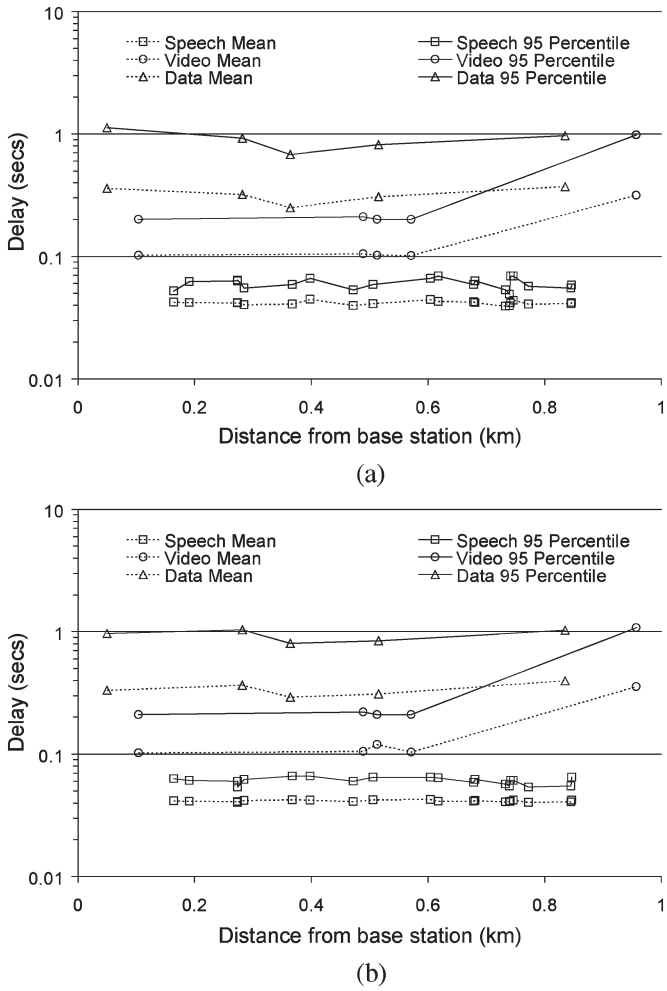

Fig. 22. Ninety-fifth percentile delay trends in a multiservice environment for simulation lengths of (a) 300 s and (b) 1800 s ($\delta_{\rm sc} = 0.7$, 30 users per cell).

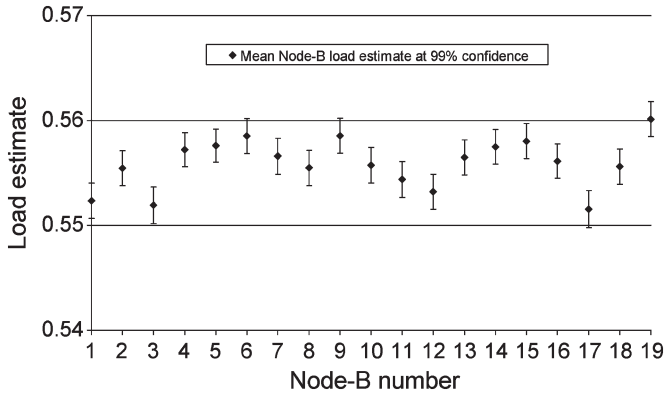

Fig. 23. Average cumulative uplink load by Node-B with 99% confidence limits ($\delta_{\rm sc} = 0.7$, 30 users per cell).

VI. Conclusion

A distributed QoS-aware resource management scheme for the UMTS uplink has been proposed for a triple-class and multicell environment with interservice and intraservice prioritization in place. The scheme has been analyzed through extensive discrete-event simulations for different combinations of system parameters and varying user load while taking ARQ, FEC, and Rayleigh fast fading into consideration. Alternative strategies are considered with control components distributed among the clients (UEs) and the coordinators (Node-B or RNC). The configuration of choice involves a rate scheduler positioned in the UE MAC layer, which regularly requests the Node-B for service changes to optimize uplink resource usage based

on the user's application class. The scheduler meets average delay targets for all applications in the absence of congestion. In congested periods, the scheduler gives preferential access to voice, video, and best effort data users, in that order. The scheduler's fairness also differs by application. All data users in a cell lose throughput as a group to give way to higher priority applications, whereas higher priority applications degrade in a conventional manner with distant users being affected first. This behavior aligns with typical application requirements as well.

ACKNOWLEDGMENT

The authors would like to thank all the reviewers for the constructive comments and ideas that have improved this paper.

REFERENCES

[1] ITU, *ITU Internet Reports 2003: Birth of Broadband*, 5th ed., Geneva, Switzerland: Int. Telecommun. Union, 2003.
[2] L. Srivastava, "Social and human considerations for a more mobile world," in *Proc. ITU/MIC Joint Workshop Shaping Future Mobile Inf. Soc*. International Telecommunication Union, Mar. 2004.
[3] F. Fitzek, A. Kopsel, A. Wolisz, M. Krishnam, and M. Reisslein, "Providing application-level QoS in 3G/4G wireless systems: A comprehensive framework based on multirate CDMA," *IEEE Wireless Commun.*, vol. 9, no. 2, pp. 42–47, Apr. 2002.
[4] H. Holma and A. Toskala, *WCDMA for UMTS: Radio Access for Third Generation Mobile Communications*. Hoboken, NJ: Wiley, 2001.
[5] A. Leon-Garcia and M. A. Arad, "Scheduled CDMA: A hybrid multiple access for wireless ATM networks," in *Proc. IEEE PIMRC*, Oct. 1996, vol. 3, pp. 913–917.
[6] S. Ramakrishna and J. Holtzman, "A comparison between single code and multiple code transmission schemes in a CDMA system," in *Proc. IEEE VTC*, 1998, vol. 2, pp. 791–795.
[7] 3GPP, "End-to-end quality of service (QoS) concept and architecture," 3GPP, Valbonne, France, Tech. Rep. TS 23.207 V5.8.0, Jun. 2003.
[8] 3GPP TSG-RAN1 #34, "Example of Rel-99 TFC control algorithm," 3GPP, Seoul, Korea, Tech. Rep. R1-031004, Oct. 2003.
[9] 3GPP TSG-RAN1 #35, "Rel-99 cell throughput with TFC control, full buffer and various channels," 3GPP, Lisboa, Portugal, Tech. Rep. R1-031243, Nov. 2003.
[10] ——, "Rel-99 cell throughput with TFC control, traffic models, and various channels," 3GPP, Lisboa, Portugal, Tech. Rep. R1-031244, Nov. 2003.
[11] ——, "Reference Node-B scheduler for EUL," 3GPP, Lisboa, Portugal, Tech. Rep. R1-031246, Nov. 2003.
[12] H. Fattah and C. Leung, "An overview of scheduling algorithms in wireless multimedia networks," *IEEE Wireless Commun.*, vol. 9, no. 5, pp. 76–83, Oct. 2002.
[13] IST-1999-10699 WINE GLASS, "Research results on UTRAN (mobility, QoS and interworking with IP core network)—Phase 1," IST, Brussels, Belgium, Tech. Rep. D05, Jan. 2001.
[14] J. Perez-Romero, R. Agusti, and O. Sallent, "An adaptive ISMA-DS/CDMA MAC protocol for third-generation mobile communications systems," *IEEE Trans. Veh. Technol.*, vol. 50, no. 6, pp. 1354–1365, Nov. 2001.
[15] Ö. Gürbüz and H. Owen, "Power control based QoS provisioning for multimedia in W-CDMA," *Wireless Netw.*, vol. 8, no. 1, pp. 37–47, Jan. 2002.
[16] S. Baey, M. Dumas, and M.-C. Dumas, "QoS tuning and resource sharing for UMTS WCDMA multiservice mobile," *IEEE Trans. Mobile Comput.*, vol. 1, no. 3, pp. 221–235, Jul.–Sep. 2002.
[17] F. Fitzek, M. Reisslein, and A. Wolisz, "Uncoordinated real-time video transmission in wireless multicode CDMA systems: An SMPT-based approach," *IEEE Wireless Commun.*, vol. 9, no. 5, pp. 100–110, Oct. 2002.
[18] K. Dimou, C. Rosa, T. Sorensen, J. Wigard, and P. Mogensen, "Performance of uplink packet services in WCDMA," in *Proc. IEEE VTC—Spring*, Apr. 2003, vol. 3, pp. 2071–2075.
[19] D. I. Kim, E. Hossain, and V. Bhargava, "Dynamic rate adaptation and integrated rate and error control in cellular WCDMA networks," *IEEE Trans. Wireless Commun.*, vol. 3, no. 1, pp. 35–49, Jan. 2004.
[20] A. Abrardo, G. Giambene, and D. Sennati, "Capacity evaluation of a mixed-traffic WCDMA system in the presence of load control," *IEEE Trans. Veh. Technol.*, vol. 52, no. 3, pp. 490–501, May 2003.
[21] J. B. Kim and M. Honig, "Resource allocation for multiple classes of DS-CDMA traffic," *IEEE Trans. Veh. Technol.*, vol. 49, no. 2, pp. 506–519, Mar. 2000.
[22] L. Nuaymi, "Association of power control and data rate selection in WCDMA," in *Proc. 14th IEEE PIMRC*, 2003, vol. 1, pp. 301–305.
[23] 3GPP, "User equipment (UE) radio transmission and reception (FDD)," 3GPP, Valbonne, France, Tech. Rep. TS 25.101 V6.4.0, 2004.
[24] ETSI, "Selection procedures for the choice of radio transmission technologies of the UMTS," ETSI, Nice, France, Tech. Rep. UMTS 30.03 V3.2.0, 1998.
[25] J. Omura and B. Levitt, "Coded error probability evaluation for antijam communication systems," *IEEE Trans. Commun.*, vol. COM-30, no. 5, pp. 896–903, May 1982.
[26] F. Gunnarsson, E. Geijer-Lundin, G. Bark, and N. Wiberg, "Uplink admission control in WCDMA based on relative load estimates," in *Proc. IEEE ICC*, 2002, vol. 5, pp. 3091–3095.
[27] 3GPP, "Requirements for support of radio resource management (FDD)," 3GPP, Valbonne, France, Tech. Rep. TS 25.133 V6.6.0, Jun. 2004.
[28] ——, "Feasibility study for enhanced uplink for UTRA FDD (Release 6)," 3GPP, Valbonne, France, Tech. Rep. TR 25.896 V2.0.0, Mar. 2004.
[29] K. Helmersson, E. Englund, M. Edvardsson, C. Edholm, S. Parkvall, M. Samuelsson, Y.-P. Wang, and J.-F. Cheng, "System performance of WCDMA enhanced uplink," in *Proc. Veh. Technol. Conf.—Spring*, May/Jun. 2005, vol. 3, pp. 1427–1431.
[30] H. Boche and M. Wiczanowski, "Optimal scheduling for high speed uplink packet access—A cross-layer approach," in *Proc. Veh. Technol. Conf.—Spring*, May 2004, vol. 5, pp. 2575–2579.
[31] P. L'Ecuyer, R. Simard, E. J. Chen, and W. Kelton, "An objected-oriented random-number package with many long streams and substreams," *Oper. Res.*, vol. 50, no. 6, pp. 1073–1075, 2002.
[32] L. Yingbo and Y. Guan, "Modified Jakes' model for simulating multiple uncorrelated fading waveforms," in *Proc. Veh. Technol. Conf.*, 2000, vol. 3, pp. 1819–1822.
[33] R. Jain, A. Durresi, and G. Babic, "Throughput fairness index: An explanation," in *Proc. ATM Forum/99-0045*, Fremont, CA, MFA Forum, Feb. 1999.
[34] 3GPP, "Quality of service (QoS) concept and architecture," 3GPP, Valbonne, France, Tech. Rep. TS 23.107 V5.9.0, 2005.

Pratik Das (S'97) received the B.E. degree from Massey University, Palmerston, New Zealand, in 2001. He is currently working toward the Ph.D. degree in electrical engineering at the School of Electrical Engineering and Computer Science, University of Newcastle, Callaghan, N.S.W., Australia.

His research interests are in the areas of traffic scheduling, quality of service, and radio resource management for third-generation radio interfaces.

Jamil Y. Khan (S'88–M'90–SM'99) received the B.Sc. (Hons.) and M.Sc. degrees in applied physics and electronics from the University of Dhaka, Dhaka, Bangladesh, in 1982 and 1984, respectively, and the Ph.D. degree in electronic and electrical engineering from the University of Strathclyde, Glasgow, U.K., in 1991.

He was a Lecturer with the University of Dhaka from 1985 to 1987. From 1991 to 1992, he worked for the Research for Advanced Communication in Europe program with the University of Strathclyde. From 1992 to 1999, he was a Lecturer and, later, a Senior Lecturer in information engineering with Massey University, Palmerston, New Zealand. Since 1999, he has been a Senior Lecturer with the School of Electrical Engineering and Computer Science, University of Newcastle, Callaghan, N.S.W., Australia. His research and teaching interests are in wireless networks, multiple-access techniques, Internet Protocol networks, and cross-layer design.