Channel-Dependent Load Balancing in Wireless Packet Networks

Giuseppe Bianchi, Ilenia Tinnirello

Università di Palermo, Dipartimento di Ingegneria Elettrica

Viale Delle Scienze, Parco D'Orleans, 90128 Palermo, Italy

Phone: +390916615269, Fax:+39091488452

bianchi@elet.polimi.it, ilenia.tinnirello@tti.unipa.it

ABSTRACT

This paper refers to a wireless cellular packet network scenario where fast retransmission of corrupted packets is used to improve the packet error ratio. Since the "gross" packet transmission rate (including retransmission) depends on the channel quality perceived, admitted calls weight unevenly in terms of effective resource consumption. In this paper, we suggest using channel quality information to drive load balancing mechanisms. We propose two novel metrics to determine the best cell to attach to, during handover or new call origination. Extensive simulation results prove the superiority of our proposed schemes with respect to traditional load balancing, which base their operation on the number of admitted calls per cell.

Key words: load-balancing, handover, packet recovery, cell occupancy evaluation.

1 INTRODUCTION

Traditional frequency and time division cellular systems are based on fixed channel assignment. In such systems, a fixed channel resource, e.g. a TDMA slot per frame, is allocated to an admitted call for its whole duration time. However, the small propagation delay characterizing micro-cellular environments allows an efficient command/response communication mode. Several multiple access control (MAC) schemes [1], [2], [3] have been designed to provide dynamic resource assignment via a command/response protocol, either on a slot-by-slot [2] or a frame-by-frame [3] basis. Dynamic packet scheduling features have been also included in the channel access control operation of emerging commercial systems, such as Bluetooth [6], the Point Coordination Function of the IEEE 802.11 [7], and third generation cellular systems (e.g. fast Automatic Retransmission Request – ARQ – mechanisms).

Unlike fixed TDMA access mechanisms, in dynamic packet scheduling mechanisms, each slot is not uniquely reserved to an attached connection, but is dynamically assigned to admitted Mobile Stations (MSs) by a central scheduler running on the Base Station (BS). By suitably designing the scheduling discipline, heterogeneous multi rate and/or variable rate traffic sources, with different delay requirements and priorities, can be accommodated on the same shared medium. Channel-dependent scheduling strategies [4], [5] have been also considered to cope with varying transmission error conditions.

This paper is motivated by the observation that a flexible slot assignment allows to trade-off cell throughput with Quality of Service (QoS) perceived by the admitted calls. In fact, fast packet retransmission can be used to reduce the packet loss ratio. However, the number of retransmissions that can be attempted to improve the packet loss ratio depends on the channel utilization, i.e. the greater the number of available idle slots, the

greater the number of retransmission opportunities. It has been shown [2][8] that, in the assumption of short error bursts, even traffic sources with tight real-time delay requirements may reach a tolerable quality of service over severely degraded channels (e.g. meet a 1% packet loss target over channels where the packet corruption probability is of the order of few tens percent [2]), provided that enough retransmission opportunities are available.

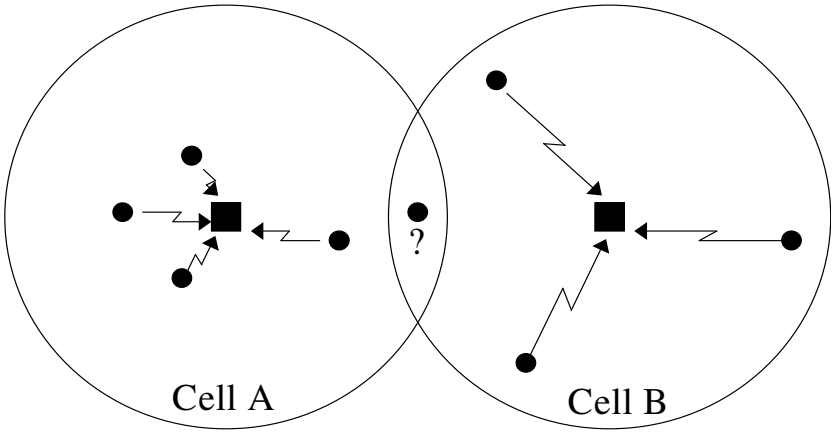


Figure 1: The problem of selecting the less loaded BS

We argue that the above described throughput/QoS trade-off affects the design of load balancing mechanisms. The new challenges in load balancing algorithms are enlightened in the simple example scenario depicted in figure 1. Two adjacent BSs serve, respectively, four and three calls in progress. At a given instant of time, a new MS wants to set up a new connection (or, equivalently, requires handover into one of the two cells). On the assumption that the MS measures a comparable channel quality to both target BSs, i.e. it is capable of attaching to either BSs, the ultimate decision to which BS to select may be taken by a load balancing algorithm. In traditional circuit switched networks, based on fixed channel assignment, load is expressed in terms of busy circuits per cell [9], [10]. In the reference example, this implies that the MS would select cell B as destination since it results the less loaded in terms of number of

admitted calls.

Suppose now that dynamic slot assignment and fast retransmission of corrupted packets is supported at the MAC layer. In the assumption that channel condition degradations are coped by retransmissions, in order to reduce co-channel interference, no form of power control is adopted. In this scenario, the number of admitted stations is no more representative of the overall packet load experienced in the cell. Since the MS's in cell B are placed at a much larger distance from the BS in comparison with the MS's in cell A, they will likely suffer of worse channel conditions. Thus, cell B will envision an extra load generated by packet retransmissions larger than cell A. Hence, it may result that, in terms of resulting packet error ratio, a much better admission strategy might consist in admitting the incoming MS to cell A, despite the fact that this would create a load unbalancing at the call level.

In conclusion, the thesis carried out in this paper is that load balancing algorithms should exploit additional information regarding the channel quality perceived by each already admitted call, rather than being limited to use "call level" information (number of accepted calls per cell). The rest of the paper is organized as follows. Section II discusses related work regarding dynamic slot assignment and load balancing in wireless packet networks. Section III proposes two specific channel-dependent BS selection metrics. Section IV describes the simulation model adopted in this paper. In section V, numerical results are presented. Finally, conclusive remarks are drawn in section VI.

2 Related Work

According to the authors' knowledge, the present paper represents the first attempt to account for the not uniform channel conditions perceived by the users in the description

of the load offered to a cell, and use such information for load balancing purposes. In fact, existing literature considers these two problems as separate issues.

A large amount of work has been done to include channel quality information in the scheduling algorithm devised to share the common wireless resource among the users of the same cell. *Channel State Dependent* packet scheduling has been originally proposed in [4] to account for the on/off behavior of the radio channel. The idea is to defer sending packets for a mobile station during error bursts and resume transmission only when the link quality improves to an acceptable level. Further studies have proposed channel dependent scheduling algorithms devised not only to intelligently cope with error bursts, but also to provide fair resource sharing among users experiencing long-term (i.e. location-dependent) different channel quality. In such a scenario, the key point of every resource assignment scheme is the tradeoff between individual user performance and total system throughput. In order to give the same service rate to all the users, more resources can be assigned to the users that perceive bad channel conditions, for example using a differentiated error correction coding [11]. On the other hand, it may be considered unfair that users experiencing good channel conditions suffer from performance degradation because of the waste of resources made by users experiencing bad channels. In addition, cell throughput is maximized by giving more resources to "good" users, at the expense of "bad" users. A thorough investigation of the throughput-fairness tradeoff in wireless packet networks has been done in [5], [12]. Note that in this paper we do not deal with throughput-fairness tradeoffs and channel dependent scheduling (as a matter of fact, the scheduling approach adopted has been kept very simple). Indeed, our proposed load-balancing mechanisms may rely on these sophisticated scheduling approaches for an underlying fair resource allocation within

each cell.

Unlike our paper, which deals with packet networks and dynamic resource assignment schemes within each cell, the problem of how to share resources in a multi-cell network scenario has been deeply investigated for circuit networks. In this case, each user receives the same amount of resources (circuit), and the QoS experienced by an admitted call is not considered to be a related issue (the underlying assumption is that an MS capable of attaching to a cell will receive a tolerable QoS). Thus, resource allocation algorithms in a multi-cell scenario is just driven by the goal of improving of the outage probability, i.e. the probability that a call is refused or subject to a forced interruption. Conversely, in our scenario, the admission of a call to a given cell not only affects the outage probability, but also affects the QoS experienced by all users within the cell

The resource sharing problem among neighboring cells have been faced in literature according to two different strategies: load balancing and channel borrowing. The first strategy can be adopted when the coverage areas of different base stations overlap. Whenever a MS can attach to more than one BS, the idea is to direct the call to the BS with the greatest number of available channels [9]. If call redirection is allowed, the optimal policy is to rearrange all the calls in the network upon every arrival and departure (call repacking). It has been proven [10] that this operation minimizes the probability that future incoming calls (either newly originated or handovering ones) will be blocked because of lack of resources. Furthermore, since frequent repacking of calls may not be feasible or may cause significant performance degradation due to handover overhead [13], more sophisticated algorithms, which trigger load balance only if load exceeds a certain threshold, have been developed [14]. The second strategy is based on

channel (frequency) repartition rearrangements among the cells [15]. The idea is to allow channel borrowing from one cell to another, in order to allocate a new call when the set of native channels is exhausted. Different borrowing policies have been proposed to limit co-channel interference. Examples of such policies are the lock of borrowed channel in all the cells within the required channel reuse distance from the borrowing cell [9] and the transmission power limitation on borrowed channels [16]. The problem of fairly allocating channels to users placed at different distance from the BS has been also considered in [17]. Channel borrowing is not dealt with in this paper.

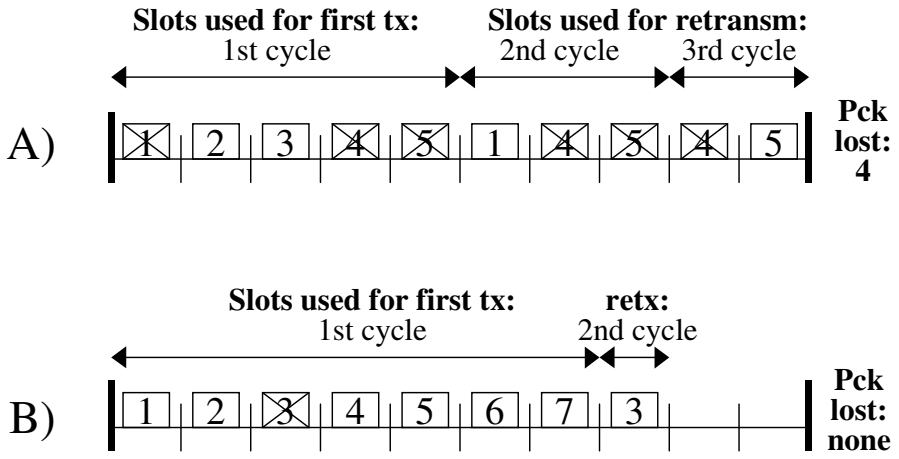


Figure 2: Usage of idle slots for retransmission

Most of the literature papers regarding load balancing do not consider mobility in their study. The rationale is that a suitable mobility model needs to account for an extremely large number of supplementary parameters, such as statistical description of velocity and direction of MSs [19], mobility prediction and related decisions [20], [21], [22], hysteresis thresholds for handover algorithms [23], etc. Eventually, this leads to affect, in an unclear manner, the performance results, i.e. it results hard to distinguish the beneficial effects of a load balancing strategy from the effects of the specific mobility models employed. Therefore, in this paper, we have conformingly decided to

limit our investigation to the case of fixed stations, leaving to future investigation the particular application of our proposed metrics to specific mobility scenarios.

3 CHANNEL-DEPENDENT LOAD BALANCING

The reference scenario is a wireless cellular packet network. Cells comprising the network have the same capacity, and in particular each cell can accommodate up to a maximum of M calls. Within each cell, access to the channel is managed via dynamic TDMA. For convenience of presentation, we can simplify[1] the description of our access scheme, by considering each channel divided into frames composed of M fixed size slots (the case M=10 is depicted in figure 2).

Different from fixed TDMA access mechanisms, channel access is coordinated by the BS scheduling operation, so that contention among transmitted packets cannot occur. Via polling commands [2], the BS scheduler grants, at the beginning of the frame (figure 2), one transmission opportunity to each admitted call. However, a packet may be corrupted by physical channel impairments. Thus, the remaining slots of the frame (*idle slots*) are granted for supplementary transmission opportunities. A packet is ultimately lost when all the granted transmission attempts fail, and no more idle slots are available within the considered frame.

The example shown in figure 2 highlights that the probability that a packet is lost depends on either (i) the channel quality encountered by the transmission of the considered packet, (ii) the spare frame capacity, represented by the number of idle slots M-N, being N the number of attached MSs, and (iii) the channel quality perceived by the other admitted calls. This latter point is motivated by the fact that other calls suffering of high transmission error probability require additional transmission opportunities

within the frame, thus reducing the number of idle slots that can be dedicated to retransmit the considered packet. In conclusion, the example highlights that the information related to the call-level load, namely the number of attached MSs, is only partially representative of the load in terms of packets per frame and of the related packet loss probability encountered by admitted calls.

In the following, we propose two different metrics to quantify the information related to packet level retransmission load. The first metric is based on the computation of the average number of packet transmissions within a cell. The second metric is more complex, and attempts to directly estimate the packet loss performance, which in turns represents an indirect measure of the packet load. Our metrics are introduced and motivated on the assumption that transmission error probability depends on the MS distance to the target BS and the result of consecutive packet transmissions for the same MS are independent (uncorrelated errors). However, we will verify that the conclusions drawn from our study are still valid when error correlation is considered (simulations have been also obtained using the realistic error correlation model [18] described in section 4.3).

We describe packet successes and failures due to fading and noise with a propagation model that takes into account Rayleigh fading, due to multipath, and η-th power loss law. The power P_R, received at the BS from a MS located at distance r, is given by:

$$P_R = \alpha^2 A\, r^{-\eta} P_T \qquad (1)$$

where α^2 is an exponentially distributed random variable with unit mean that represents the fading, $Ar^{-\eta}$ accounts for the power loss law, η typically takes values in the range [3, 4] and P_T is the transmitted power, which is considered the same for all

[1] As detailed in section 4.2, where the scheduling algorithm adopted is presented, there is actually no need to have

stations. Assuming a threshold model for the packet success, the probability $p_i(r)$ that a packet transmitted by a station at distance r is incorrectly received at the BS is given by:

$$p_r(r) = \Pr\left[\frac{P_R}{W} \le b\right] \tag{2}$$

where W is the total noise power at the BS (assumed constant), and b is the capture ratio. This model implicitly considers that the value of the power signal attenuation is constant throughout the duration of the packet transmission. Given W, the parameters A and P_T, and recalling that α^2 is exponentially distributed, we have:

$$p_r(r) = 1 - e^{-br^{\eta} / SNR_0} \tag{3}$$

where $SNR_0 = AP_T/W$ is the average Signal To Noise Ratio at the cell border. We assume that different users experience different multipath and noise effects, so that their transmission processes are independent.

3.1 The "Gross Load" Metric

The "Gross Load" (*GL*) metric stems from the observation that, assuming unlimited resource availability, the number of packet retransmissions necessary to successfully transmit a single packet is a geometrically distributed random variable, i.e:

$$P_{ES_i}(j) = (1 - p_i)\, p_i^{\,j} \tag{4}$$

is the probability of a successful packet transmission after j failures (in other words, the probability that the i-th MS requires j Extra Slots to complete the packet transmission). In the above equation, p_i is the transmission error probability suffered by the i-th MS, placed at distance r_i from the serving BS, given by equation (3). This implies that, in average, the i-th MS needs g_i slot to transmit its packet successfully, where:

---

explicit frame delimiters on the channel.

$$g_i = 1 + E[extra\ slots] = 1 + \frac{p_i}{1 - p_i} \tag{5}$$

We define g_i to be the Gross Load offered by MS i. Being N the number of attached MSs, the total Gross Load GL offered to a cell is given by:

$$GL = \sum_{i=1}^{N} g_i = N + \sum_{i=1}^{N} \frac{p_i}{1 - p_i} \tag{6}$$

GL is given by the sum of two terms. The first is the number N of attached MSs, that is a static value when no MS attaches/detaches. The second term is a dynamic value, depending not only on the number of MSs, but also on their physical position, since p_i depends on the (time-varying) MS distance r_i through equation (3).

3.2 The "Packet Loss" Metric

The "Packet Loss" metric is motivated by the observation that the best possible load balancing criterion is the minimization of the expected packet loss percentage after the addition of a new MS in the network. In the simplistic case of all MSs suffering the same packet error probability p (i.e. placed at the same distance from the BS), a closed form expression of the probability to lose l packets ($l > 0$) is given by:

$$\Pr\{lost\ packets = l\} = p^{l+M-N}(1-p)^{N-l}\binom{M}{N-l} \tag{7}$$

where M is the number of available slots, and N is the number of stations. Thus, the average packet loss in a cell of capacity $M=N+k$, with N stations connected, is given by:

$$P_{loss}(N,k) = \frac{1}{N}\sum_{l=1}^{N} l(1-p)^{N-l}\, p^{l+k}\binom{N+k}{N-l} \tag{8}$$

Although, an explicit loss formula for MS having different packet error probabilities appears non trivial, we have verified that an excellent approximation is to compute the "Packet Loss" (PL) formula (8) using, instead of p,

$$p_{equ} = 1 - \frac{N}{GL} \tag{9}$$

The rationale at the basis of our approximation is to substitute the actual MS distribution with another one, in which we consider an equal number of equidistant stations that offer the same gross load *GL*.

4 SYSTEM MODEL

4.1 Network and Traffic Model

To avoid border effects in the simulation results, we have developed a toroidal network model (a similar model has been used in [2]). The network is depicted in figure 3, and consists of *25=5x5* hexagonal cells[2] lying on a torus surface. As shown in the figure, there is no network border, and each cell always has 6 adjacent cells - e.g. cell (5,5) is adjacent to cells (1,4), (1,5), (5,4), (5,1) (4,5) and (4,1).

The BSs are located at the center of the hexagonal cells and operate with omni-directional antennas. This implies that the coverage area of each one is circular and contiguous areas overlap. We call overlap-factor (*OF*) the ratio between the maximum distance d_{max} at which an MS can attach, and the measure *R* of the hexagonal cell side. Unless otherwise specified, the value d_{max} = 1.5 *R* has been adopted. As illustrated in figure 3, for a given BS, we define the coverage area within the circle inscribed in the hexagonal cell as the inner coverage region, and the coverage area external to the circle of radius *R* as the outer coverage region.

[2] We have verified via simulation that more computationally expensive results obtained with larger network sizes do not show any significant difference with respect to the 25 cells case, which has been thus adopted as reference network scenario for all the simulation runs.

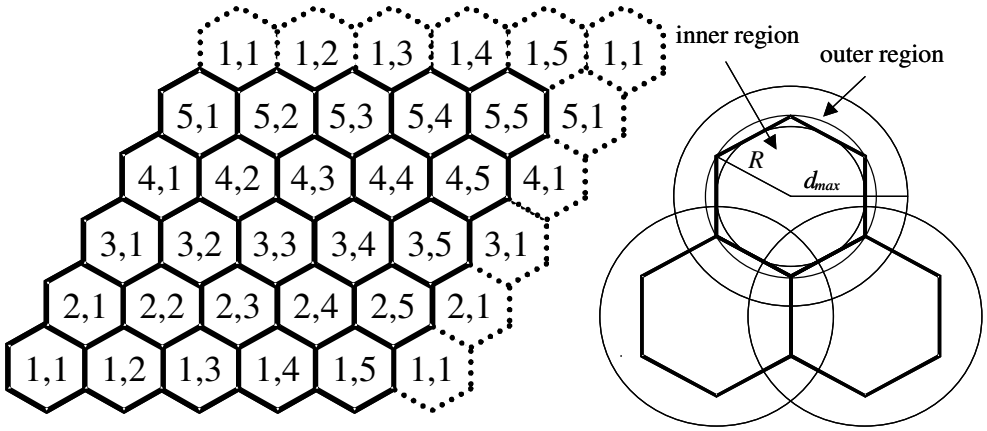


Figure 3: Cellular network scenario (toroidal topology and cell overlapping areas) used in the simulation program

Note that if a station is located in the inner region of a cell it is attached to the closest BS, while if it is located in the outer region of a cell it is not attached to the closest BS. For this reason, we call a station in the inner region as a close station, and a station in the outer region as a remote station. The signal to noise ratio SNR_0 is measured at a distance equal to R. Unless otherwise specified, very critical channel scenarios have been considered, by setting $SNR_0/b = 5$, which (owing to eq. 3) results in about 18% slot corruption ratio at border cell.

Our investigation has been carried out for a dynamic scenario in which homogeneous constant rate traffic sources activate and deactivate. New calls arrive to the network according to a Poisson process. For each arriving call, a random position for the originating MS is uniformly drawn on the whole torus surface. Every arriving call undergoes a very simple admission control test: A call is accepted by the network if it finds at least one BS in its coverage area (i.e, for sake of simulation, whose distance is lower than d_{max}) for which the number of already admitted calls is lower than M (in what follows, we'll refer to M with the name cell capacity). Otherwise the call is

blocked[3]. If more than one available BS is found, the MS selects the BS target as the one that minimizes the selected metric, computed including the contribution of the incoming MS.

When accepted, each call emits fixed size packets at a constant rate of one packet every χ=30 ms. Accepted calls last for an exponentially distributed time with mean value set to 150 s (5000 packets). Give these traffic assumptions, the capacity M of each cell depends on the channel rate, and, ultimately, on the time slot duration σ. Unless otherwise specified, we have assumed σ=1.5 ms, yielding M=20.

Although no MS mobility has been considered, we assume that the process of handover from one cell to another occurs if, at a given instant of time, a better QoS is offered in an adjacent cell. In fact, since the cells significantly overlap (figure 3), a user placed in the cell periphery can be served by two or more BSs. In principle, an handover event to a different cell may be triggered by the load-balancing algorithm even for stations placed very close to a BS, for example when an adjacent cell is very low loaded. However, this is not recommended, as future calls attaching to the destination cell will very soon trigger an inverse handover. Thus, we have introduced a minimum threshold, called d_{trig}, that we have set equal to the radius of the inner region (see Figure 3). Stations whose distance from the serving BS results lower than d_{trig} cannot perform further handovers.

4.2 Channel access and scheduling scheme

Within each cell, the channel access scheme is based on a polling mechanism coupled with a scheduling strategy that allows prompt retransmission of erroneous

[3] As a consequence, packet loss is never due to cell capacity overflow, but only to radio channel impairments. Note that the definition of a call admission control scheme aimed at providing QoS guarantees is a complex problem, out of the scopes of this paper.

packets. A simplified framed operation of the scheduling strategy was graphically shown in figure 2. Upon arrival, each packet is given a first transmission opportunity. Packets whose transmission has failed are queued, and a second, third and so on retransmission opportunities are given, until the frame space exhausts.

In the simulation program, we have implemented the same scheduling logic, but without relying on explicit channel frames. In particular, the scheduler is composed of a bank of priority FIFO queues *Q[i]*. As soon as a new MS has a new packet to transmit, a transmission request is inserted in queue *Q[0]*. When the first transmission fails, a retransmission request for the considered MS is inserted in queue *Q[1]*, and so on. At each time slot, the scheduler grants transmission for the first non null head of the line request found starting from queue *Q[0]*: a grant for queue *Q[i]* is assigned only if queues *Q[0], …, Q[i-1]* are empty.

A packet originated from a given MS is lost if it is not successfully transmitted by the time χ of arrival of the subsequent packet. In such a case, the scheduler deletes the transmission request of the expired packet, and inserts the transmission request for the new packet in queue *Q[0]*.

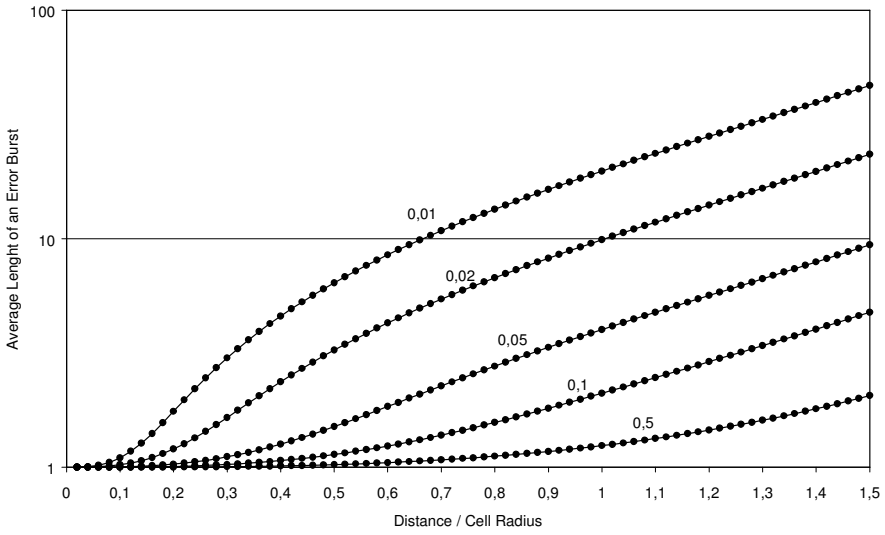
4.3 Correlated Error Model

In addition to the uncorrelated error model, described in section 3, we have implemented a more realistic correlated error model [18]. This model assumes that the channel fluctuates between two states: the good state in which no packet error occurs, and the bad state in which every packet transmission fails. The model is specified by the following two parameters:

- average packet error rate p_r
- average length of an error burst q

The parameter p_r measures how often, in average, a packet transmission fails. It depends on the MS's distance r from the BS according to (3). The second parameter q indicates how clustered the errors tend to be. It depends on the packet error rate (i.e. the distance r), as well as on the product $f_d\tau$, where f_d is the Doppler frequency (mobile velocity divided by carrier wavelength), and τ is the time slot duration[4]. An analytical expression for $q(p_r, f_d\tau)$ is non trivial, and can be found in [18]. For convenience of the reader, Figure 4 reports q, measured in time slots, versus the distance r of the MS from the BS, for a signal to noise ratio at border cell equal to 5 and for several values of the correlation coefficient $f_d\tau$.

When $f_d\tau$ is small (<0.1) the error process is very correlated (slow fading), as testified by the large error burst duration for remote stations. Conversely, for larger values of $f_d\tau$ (>0.5), two consecutive transmissions are in practice independent (fast fading).



[4] The described correlation model is introduced to validate the robustness of our proposed metrics for different channel scenarios. In this sense, the product $f_d\tau$ must be considered as a pure parameter for tuning the channel correlation, rather than, literally, intended as a model depending on terminal mobility (we remark that our MSs are fixed).

Figure 4: Average length of an error burst versus the normalized MS distance from the serving BS, varying the fading velocity (correlation coefficient $f_d\tau$)

5 NUMERICAL RESULTS

In this section, we compare the performance of our proposed packet-level load balancing metrics, namely the "*Gross Load*" metric (*GL)* and the "*Packet Loss*" metric (*PL*), with the performance provided by the following two traditional approaches:

- *Minimum Distance (MD)*: the target BS is the closest one. According to this metric, no load balancing is implemented, and the cell selection criterion is purely driven by channel quality optimization (distance from the BS).

- *Nominal Load (NL)*: This is the traditional approach to load balancing. Among the set of available cells, the target BS is the one accommodating the lower number of admitted calls. Therefore, no channel quality optimization is considered.

The goal of our comparison is to evaluate the performance improvements, in terms of QoS and throughput, achievable by the combined usage of channel quality and cell load information, versus the traditional *MD* and *NL* schemes, which are based on either the former or the latter information. Unless otherwise specified, the following results have been obtained with the system parameters summarized in table 1.

Table 1. System parameters

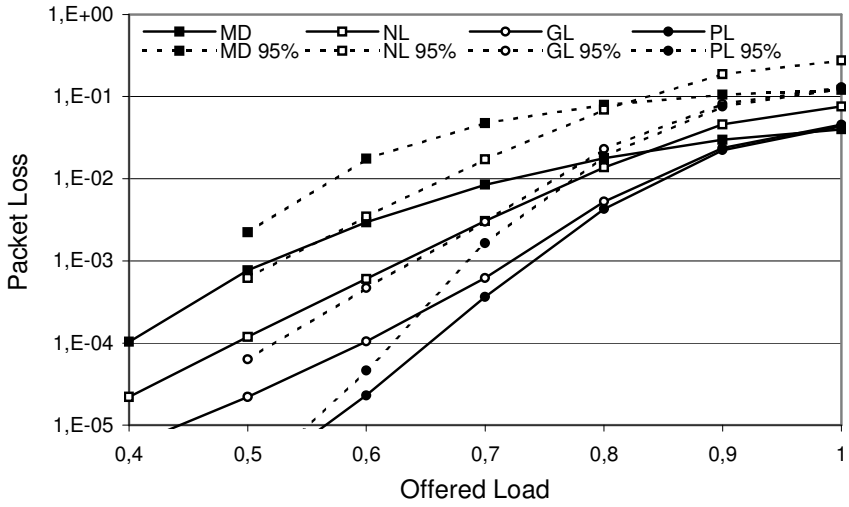


Figure 5. Comparison of packet loss figure for different BS selection criteria

Figure 5 plots, for each of the four considered BS selection criteria, the average packet loss ratio encountered by admitted calls (straight lines), as well as the 95% percentile of the packet loss ratio (dashed lines), versus the normalized offered load. The figure shows that the channel-dependent *GL* and *PL* metrics, targeted to jointly optimize channel quality and retransmission effectiveness, consistently provide loss performance improvement in comparison with the traditional *MD* and *NL* metrics, in almost all load conditions. Specifically, for offered loads leading to an average packet loss lower than 10^{-2}, our metrics achieve an improvement of at least one order of magnitude.

A comparison between *PL* and *GL* shows that *PL* provides better performance, especially when light load conditions are considered. The reason is that the *PL* metric directly optimizes the expected packet loss, while the *GL* metric is only indirectly related to this performance figure. From the figure, we note also that, when the normalized offered load is less than 0.8, the *NL* metric is more performing than the *MD*

metric. This appears motivated by the fact that the uniform user distribution among the cell improves the packet recovery capability via retransmission. However, although *NL* balancing improves the retransmission effectiveness, admitted calls suffer of worse channel conditions, since they are, in average, more distant from the serving BS, with respect to the *MD* selection criterion. Thus, in high traffic conditions, when frame space left to retransmissions is minimal, we observe that the *NL* metric gives the worst performance.

In addition, the 95-th percentile curves reported in figure 5 show that the *GL* and *PL* metrics are effective not only in terms of average conditions perceived by the users, but they provide a reduced packet error ratio even for MSs which experience worst-case conditions.

We have run simulation results in which we have tuned the cell load to achieve a packet loss ratio lower than 10^{-2} for 95% of the admitted stations. Figure 6 reports the resulting accepted load versus the cell capacity for each of the four metrics. The figure shows that channel-dependent metrics provide a large increase in the sustainable network throughput (i.e. throughput constrained to a QoS requirement) when compared with the *MD* and *NL* metrics.

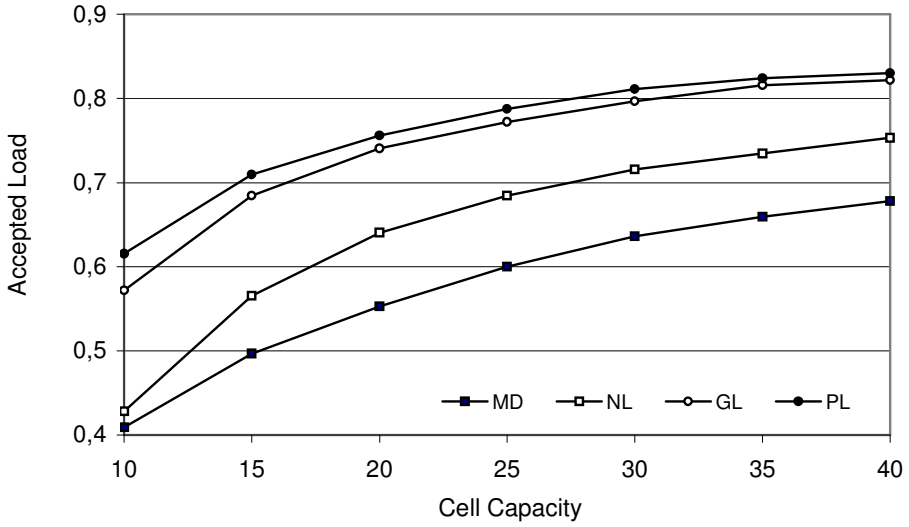


Figure 6. Maximum accepted load to guarantee a packet loss ratio less than 0.01 for the 95% of the stations. SNR/b=5; OF=1.5.

The following two figures visualize how our load balancing mechanisms work, in terms of geographical distribution of the packet transmissions and load distribution among the cells. Figure 7 shows the distribution of the packet transmission attempts versus the distance from the serving BS. Curves have been obtained by quantizing the x-axis in 0.1 steps. Each curve shows a slope variation starting from interval 0.8-0.9 *R*, where *R* is the cell radius. The reason for this phenomenon is purely geometric: as the distance grows from the inner region to the outer region of a cell (see Figure 3), due to the hexagonal geometry, the density of users reduces.

As expected, the figure shows that packets can be transmitted from a distance greater than *R*, and that this probability depends on the balancing criterion employed. The *MD* metric minimizes the number of transmissions in the outer region. In fact, an MS attaches to a remote BS (i.e. to a BS whose distance is greater than *R*) only if all the other BSs in its coverage area are busy. Conversely, the other three load balancing

schemes allow MSs to attach to remote BSs, provided that these are low loaded. From the figure we see that, in the case of the *NL* metric, as much as 3.3% of the packet transmissions occur at maximum distance from the BS (1.4 *R* - 1.5 *R*). Moreover, the number of packet transmissions originated in the outer region is almost constant, since the station position is not considered in the load balancing. *GL* and *PL* metrics represent an intermediate case between *MD* and *NL* results. Since channel condition is included in load balancing, the probability that very far MSs connect to a given BS is very low.

An important target of our load-balancing policy is to distribute users among cells to leave spare capacity for retransmissions. Figure 8 quantifies the distribution of the number of simultaneous users within a cell. From the figure, we see that, in the case of the *MD* metric, cells are full in the 7.4% of the cases. In this case, no idle slots are available for retransmissions, and loss performance coincide with the radio channel error rate given in (3), which is considerably high in our simulation scenario (SNR/b=5). Conversely, the *NL* metric is devised to concentrate as much as possible the distribution of the number of simultaneous users around its mean value (in our case, 14 users), and therefore to maximize the spare capacity in each cell. The figure shows that the *GL* and *PL* metrics closely emulate the *NL* metric in reducing the spread of the distribution of the number of stations per cell around its mean, without suffering of the disadvantages of the *NL* metric, concerning remote MSs, as discussed for figure 5 and figure 7.

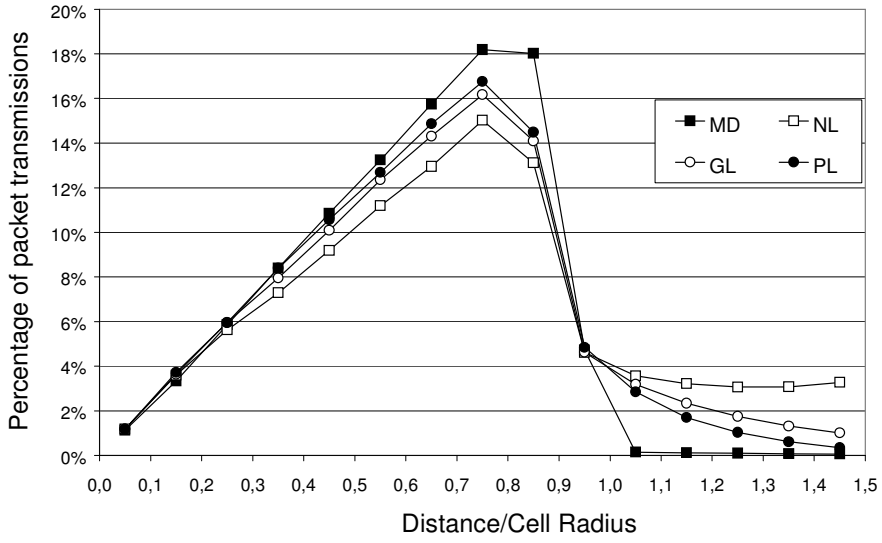


Figure 7. Distribution of the transmitted packets versus the MS distance from the serving BS (quantization step: 0.1). Offered Load=0.7; Frame size=20 slots; SNR/b=5.

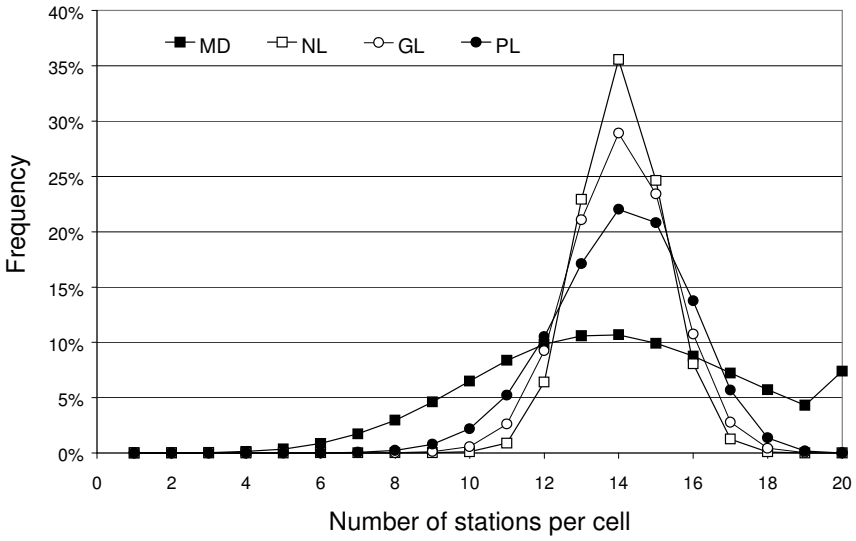


Figure 8. Cell occupation distribution - Offered Load=0.7; Frame size=20 slots; SNR/b=5.

Finally, we want briefly investigate some effects of different modeling and design choices on the performance of our load-balancing schemes. A possible question is the

effect of channel temporal correlation on the performances provided. Clearly, the slower the fading process, the less likely a failed packet transmission can be recovered before the next packet arrival. In fact, for very correlated error channels, retransmission-based scheduling loses most of its effectiveness. This result is put into light in figure 9, which shows that, as the correlation coefficient $f_d\tau$ decreases (i.e. the fading becomes slower) the packet loss performance of the system gets worse regardless of the specific metric considered. It is straightforward to note that, for highly correlated channels (e.g. $f_d\tau <$ 0.02), the most performing metric becomes the *MD*.

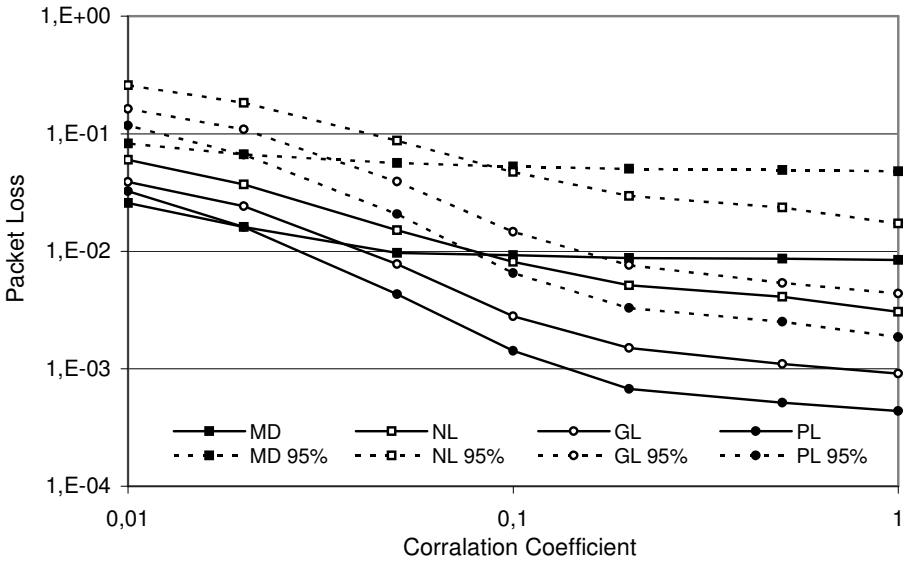


Figure 9. Packet loss versus the correlation coefficient $f_d\tau$. Offered Load=0.7; Frame size=20 slots; SNR/b=5; OF=1.5.

In fact, for highly correlated channels, each packet transmission has, in practice, one single opportunity (if the first packet transmission fails, it is extremely likely that the subsequent retransmission fails again). In such conditions, the obvious strategy is to minimize the transmission error probability, i.e. maximize the radio link quality, as provided by the *MD* metric. Indeed, it is quite interesting and surprising to note that the

PL metric remains superior not only for slowly correlated channel ($f_d\tau$ close to 1), but also for mildly correlated channel ($f_d\tau > 0.02$).

Figure 10 shows the sensitivity of the performance (packet loss and 95-th percentile) on the cell overlapping-factor *OF*, i.e. to the maximum distance at which an MS can attach. This factor is a critical design parameter. In fact, as *OF* increases, the call block probability, due to the lack of available channels, is reduced, while the average transmission conditions of the attached MSs degrade. An opportune *OF* dimensioning represents a tradeoff between capacity and quality of service provided by the network. From the figure we note that, when the overlapping-factor is small (*OF* < *1.2*), all metrics give similar performance. In fact, when the transmission error probabilities are not critical, gross load and nominal load are almost equivalent. Moreover, the calls that can be directed to more than one BS, represents a small fraction of the overall traffic: thus the effect of the load-balancing, in terms of spare resource distribution, is marginal. Indeed, as the *OF* increases, this effect becomes more relevant, but, at the same time, the average transmission conditions get worse, because very far MSs, subject to large transmission error probabilities, may be admitted. When the first effect prevails, the performances improve; conversely, when the second effect prevails, the performances degrade. This explains the minimum that we observe in the *NL, GL* and *PL* curves. The most interesting result is that, in the *PL* case, the performance is almost independent of the overlap factor. This is a very important practical result, as the increase of the overlap factor is a mean to increase the network capacity.

Figure 11 shows the packet loss probability encountered for different values of the SNR_0/b at distance *R*. Obviously, the curves are decreasing and monotone. In all the cases, our metrics give the best performance. Note that the *NL* and *MD* curves present a

crossing point (an explanation has been given while discussing Figure 5). As expected, for large SNR_0/b, the curve shows that, in general, load balancing mechanisms (i.e. *NL*, *PL*, *GL*) become a very effective mean to improve the performance with respect to the *MD* criterion (no load balancing).

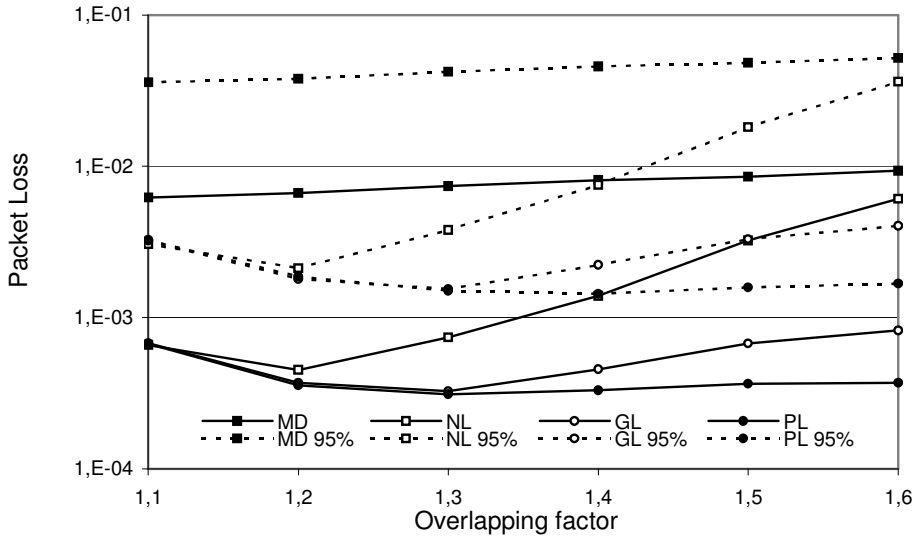


Figure 10. Packet loss observed varying the BS coverage. Offered Load=0.7; Frame size= 20 slots; SNR/b=5.

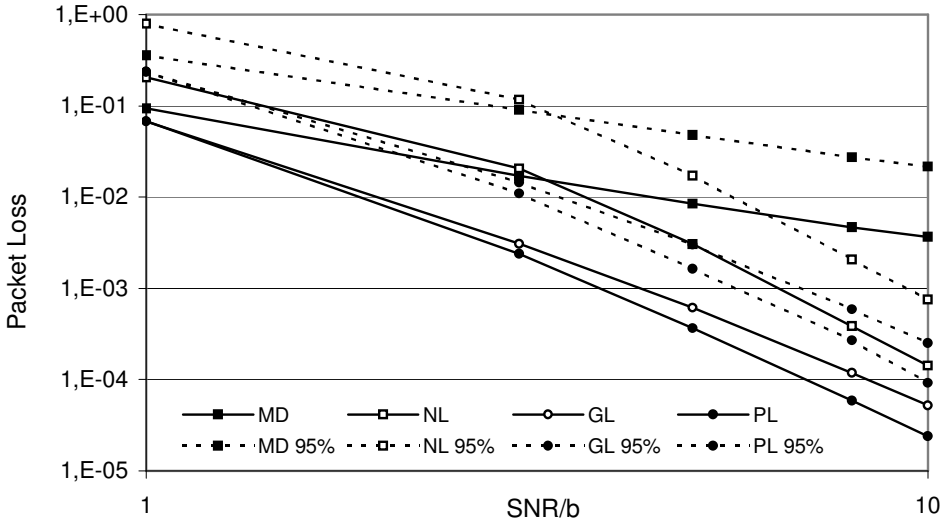


Figure 11. Packet Loss observed varying the signal to noise ratio. Offered Load=0.7; Frame size=20 slots; Overlapping factor=1.5

6 CONCLUSIONS

In this paper we have considered wireless cellular packet networks based on dynamic resource assignment. For such schemes we have shown that cell load is only loosely related to the number of attached connections, as it depends on the amount of retransmissions needed to successfully transmit a packet. We have shown that traditional load balancing schemes are not effective to operate in this context. Hence, we have proposed two novel load balancing metrics to drive the selection of the best BS candidate to connect to. The first metric, called "Gross Load" (*GL*), accounts for each connection in terms of its average channel slots consumption per successful packet transmission. Since the transmission error probability varies with the stations' position, the *GL* metric dynamically adapts to any cell configuration. A second metric, called "Packet Loss" (*PL*), was then introduced to overcome some limits shown for the *GL* approach. The *PL* metric attempts to directly estimate the packet loss performance experienced in a cell, and consequently allows MSs to always select the cell for which the best QoS is expected.

Extensive simulation results show that our proposed metrics provide the advantages of traditional load balancing schemes, while in the same time attempting to minimize the distance of each station from the selected cell. Moreover, our proposed schemes appear to seamlessly adapt to various traffic and channel conditions, and thus they appear to be valid candidates for next generation wireless packet networks.

REFERENCES

1. M. J. Karol, Z. Liu, K. Y. Eng. An Efficient Demand-Assignment Multiple Access Protocol for Wireless Packet (ATM) Networks. *ACM Journal on Wireless Networking* 1995, **1**(3):267-280.

2. G. Bianchi, F. Borgonovo, L. Fratta, L. Musumeci, M. Zorzi. C-PRMA: a Centralized Packet Reservation Multiple Access for Local Wireless Communications. *IEEE Trans. on Vehicular Technology* 1997, **46**(2):422-436.

3. D.Raychaudhuri, L.J.French, R.J. Siracusa, S.K. Biswas, R. Yuan, I. Narasimhan, C. A. Johnston. WATMnet: a prototype wireless ATM system for multimedia personal communications. *IEEE Journal of Selected Areas in Communications* 1997, **15**(1):83-95.

4. P. Bhagwat, P. Bhattacharya, A. Krishna, S.K. Tripathi. Enhancing throughput over wireless LAN's using channel state dependent packet scheduling. *Proc. IEEE INFOCOM '96*, **3:**1133 -1140.

5. S. Lu, V. Bharghavan, R. Srikant. Fair Scheduling in Wireless Packet Networks. *IEEE/ACM Transactions on Networking* 1999, **7**(4):473–489.

6. Bluetooth SIG groups. Specification of the Bluetooth system: Core, version 1.0. July 1999.

7. IEEE 802.11, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, ISO/IEC-8802-11, Geneva, Switzerland, 1999.

8. F. Borgonovo, A. Capone, L. Fratta. Retransmissions Versus FEC Plus Interleaving for Real-Time Applications: A Comparison Between CDPA and MC-

TDMA Cellular Systems. *IEEE Journal on Selected Areas in Communications* 1999, **17**(11):2022 - 2030.

9. J. Karlsson, B. Eklundh. A cellular mobile telephone system with load sharing – an enhancement of directed retry. *IEEE Trans. Commun.* 1989*,* **37:**530-535.

10. M. Alanyali, B. Hajek. On simple Algorithm for Dynamic Load Balancing. *Proc. IEEE INFOCOM '95*, **1:**230-238.

11. M.Elauod, P.Ramanathan. Adaptive use of error correction codes for real time communication in wireless networks. *Proc. IEEE INFOCOM '98*, **2:**548 -555.

12. T.S.E. Ng, I. Stoica, H. Zhang. Packet Fair Queueing Algorithms for Wireless Networks with Location-Dependent Errors. *Proc. IEEE INFOCOM '98*, **3:**1103–1111.

13**.** A. Acampora, M. Naghshineh. Control and Quality of Service Provisioning in High Speed Microcellular Networks. *IEEE Personal Communication Magazine* 1994, **1**(2)**:**36-46.

14. M.H. Willebeck-Le Mair, A.P. Reeves. Strategies for Dynamic Load Balancing on Higly Parallel Computers. *IEEE Trans. Parallel and Distributed Systems* 1993, **4:**979–993.

15. T. Chu, S. Rappaport. Overlapping Coverage with Reuse Partitioning in Cellular Communication systems. *IEEE Trans. on Vehicular Technology* 1997, **46**(1):41-54.

16. H. Jiang, S.S. Rappaport. A new channel assignment and sharing method for cellular communication systems. *IEEE Trans. on Vehicular Technology* 1994, 313–322.

17. X. Lagrange, B. Jabbari. Fairness in Wireless Microcellular Networks. *IEEE Trans. on Vehicular Technology* 1998, **47**(2):472-479.

18. M. Zorzi, R. Rao, L.B. Milstein. Error statistics in data transmission over fading channels. *IEEE Trans. on Communications* 1998, **46**:1468–1477.

19. K.K. Leung, W.A. Massey, W.A.; W. Whitt. Traffic models for wireless communication networks. *IEEE Journal on Selected Areas in Communications* 1994, **12**(8):1353–1364.

20. M. Hellebrandt, R. Mathar. Location tracking of mobiles in cellular radio networks. *IEEE Trans. on Vehicular Technology* 1999, **48**:1558–1562.

21. D.A. Levine; I.F. Akyildiz; M. Naghshineh. A resource estimation and call admission algorithm for wireless multimedia networks using the shadow cluster concept. *IEEE/ACM Transactions on Networking* 1997, **5**(1):1 –12.

22. W. Su; M. Gerla. Bandwidth allocation strategies for wireless ATM networks using predictive reservation. *GLOBECOM '98*, **4:**2245–2250.

23. M.D. Austin; G.L. Stuber. Direction biased handoff algorithms for urban microcells. *Veh. Tech. Conference* 1994 , **11:**101–105.