A Contention/Reservation Access Protocol for Speech and Data Integration in TDMA-Based Advanced Mobile Systems

Giuseppe Anastasi*, Davide Grillo**, Luciano Lenzini*, and Enzo Mingozzi*

* *Dept. of Information Engineering, University of Pisa, Via Diotisalvi 2, I-56126 Pisa (Italy)*

** *Fondazione Ugo Bordoni, Via Baldassarre Castiglione, 59, I-00142, Roma (Italy)*

Abstract. The performance of third generation mobile systems is greatly influenced by the multiple access protocols used in the radio access system. The paper introduces a multiple access protocol, SIR (Service Integration for Radio Access), which has the potential for accommodating the requirements of speech and bursty data traffic in an efficient way. SIR is evolved from an access protocol (PRMA++) studied within the framework of the TDMA-based version of the European evolving standard for third generation mobile systems. In particular, SIR uses the same frame structure and in-band signalling but introduces a contention-free handling of data bandwidth requests while meeting speech service requirements via basic PRMA++ mechanisms.

1. Introduction

One key requirement on the radio access design of advanced, third generation mobile systems is the ability to accommodate a variety of service types via flexible and efficient multiple access protocols. For TDMA-based systems, PRMA, [1], and more recently PRMA++, [2], access protocols have been proposed. The latter has been studied in the framework of the ATDMA-based version of UMTS (Universal Mobile Telecommunication System), [3], the European developing standard for third generation mobile systems. In particular, aspects related to bandwidth allocation and management - as opposed to bandwidth request handling - need close examination to ensure both efficient use of available spectrum and meeting of quality requirements. These requirements relate to the diverse components of the traffic mix resulting from the service integration scenario envisaged for advanced mobile systems. While the performance of PRMA and PRMA++ has been analyzed mainly for speech traffic, e. g. [4] - [5], relatively less effort has been devoted to their assessment for mixtures of speech and data, e. g. [6] - [7]. With PRMA, information packets contend for bandwidth request/allocation according to a slotted ALOHA mechanism. Depending on the service type, allocated bandwidth may be retained until no longer needed or contention may occur on a packet-by-packet basis. With PRMA++, contention is restricted to packets dedicated to notify the need for bandwidth, so-called "reservation" packets; once requests have been notified, bandwidth is centrally allocated at the base station. These protocols are apt to exploit the bursty nature of "interactive" applications such as speech (statistical multiplexing), but could also be used so as to degenerate to "circuit-switching" mode - in the latter case discontinuous transmission would still be used for reducing co-channel interference. In addition to speech, other applications could in principle profitably be handled according to statistical multiplexing. However, this should not be at the cost of excessive request activity which would result in inefficient use of bandwidth and outweigh the PRMA(++) advantage.

Operation stability problems and inefficient use of bandwidth resulting from a straightforward use of PRMA++ for a mixture of speech and a particular type of data applications (Telnet, [8]) have been shown in [9] for a macrocell environment. To overcome these shortcomings, a new access scheme, SIR (Service Integration for Radio Access), evolved from PRMA++ has been introduced, [9]. This evolution consists essentially in adopting the PRMA++ (access contention) protocol for speech and a Round-Robin algorithm for collecting bandwidth requests of the data connections. Backlogged data workload is then carried by using both dedicated bandwidth and bandwidth left over by speech traffic, giving data connections a chance to transmit according to a Round-Robin algorithm. This enables contention-free and flexible use of (reserved) bandwidth for accommodating data traffic demand while ensuring guaranteed bandwidth to speech applications. The SIR access mechanism adopts the same frame structure as PRMA++, in particular it uses the same in-band signalling for reservation, acknowledgement and fast paging tasks. In this sense, SIR is compatible with such radio transmission techniques as the ones proposed for the TDMA-based UMTS.

This paper extends the analysis in [9] to a microcell environment and provides a firm basis for assessing the traffic performance of SIR. This is achieved by exercising the protocol under a wide range of operation and workload conditions. Since stable operation of PRMA++ is a prerequisite for sensible assessment of SIR, particular emphasis is given in the paper to the identification of the stable operating region and its dependency on the retransmission activity as a means for resolving contention situations. This fills a gap in previous analysis on PRMA++.

In presenting the SIR access protocol, the paper focuses on transmission of user information from mobiles to the base station and the related in-band signalling on both up- and downlink. It does not consider user information transmitted in the reverse direction. Moreover, the paper considers those aspects which are key from a traffic performance perspective. Technical aspects related to monitoring/restoration of the channel quality which have a bearing on the implementation of a radio access protocol such as interleaving, error recovery, power control, etc. are beyond the scope of the paper.

The rest of the paper is organized as follows. Section 2 briefly recalls the essential aspects of PRMA++ from a traffic performance point of view and highlights issues related to stable system operation when handling speech traffic. Section 3 gives a detailed description of SIR access protocol. Section 4 analyzes the performance of SIR for a basic scenario and a range of combinations of operating parameters; the analysis is then extended to variations of the basic scenario chosen so as to test the sensitivity of the access protocol to different offered load characterization and apportionment of bandwidth for speech and data. Finally, section 5 draws some conclusions.

2. PRMA++ Speech Performance and Operation Stability

As anticipated, PRMA++ is based on a distinction between reservation and user data slots in a TDMA frame. The former are used by the mobiles to notify the base station about bandwidth needs, the latter are used for conveying (contention-free) information from the mobiles. The distinction between reservation and traffic slots was first introduced in [10], whereas a description of the PRMA++ protocol is found in [2] and [5].

2.1 PRMA++ Basic Mechanisms

The PRMA++ physical frame structure is shown in Fig. 1. Uplink and downlink frames are synchronized. The slots on the uplink are separated into reservation slots (R-slots) and user information slots (I-slots). Access contention is confined to R-slots only, which are dedicated to control packets used for call set-up, handover procedures and for bandwidth reservation. I-slots are used for collision-free transmission of information packets. Likewise, the slots on the downlink are separated into acknowledgement slots (A-slots), paired with uplink R-slots, and information slots (I-slots), for downlink information transmission.

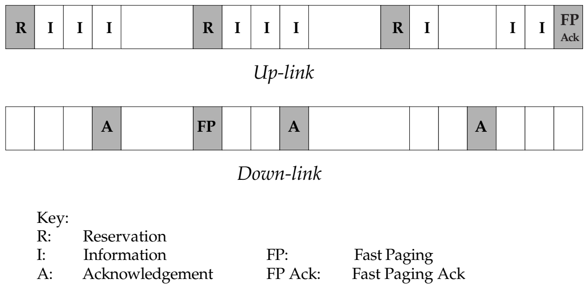

Fig. 1. PRMA++ frame structure (key elements).

Contention vicissitudes on R-slots are notified to mobiles through acknowledgements on downlink A-slots. Furthermore, since allocation is performed for downlink I-slots too, mobiles are informed about which downlink channel they have to listen to via special fast paging slots (FP-slots), paired with fast paging acknowledgement slots on the uplink (FP/Ack slots).

In order to gain a reservation for one or more I-slot on the uplink channel, a mobile has to transmit a *request* to the base station by means of the reservation channels. If the request in unsuccessful because of collision with other requests, then the request is re-attempted on the next available R-slot with probability p, "permission probability". Once such a transmission is successful, an I-slot allocation process is initiated.

If the number of requested slots is not available, the mobile is informed by the base station via an A-slot, either to wait until bandwidth is available or to reduce the bandwidth demand according to some service dependent policy. If the (re-negotiated) request can be satisfied, the base station allocates the slots required, and the allocation process terminates, otherwise requests are queued at the base station.

One of the distinguishing features of PRMA++ is the central role played by the base station in managing bandwidth resources. While waiting for bandwidth to become available, excessive delay may be incurred in. In the case of speech, this may lead to packet dropping at the source since too long delayed packets would negatively affect time integrity of signal reconstruction at the sink. Performance analysis of PRMA++ as used for speech service, see [5] for example, has been focusing on the maximum number of speech connections that can share radio channels while meeting a target dropping probability, P_{drop}, consistent with speech quality objectives, e. g. $P_{drop} < .01$.

2.2 Stability of Operation for Speech Service

While retransmission of collided requests helps in recovering from momentary congestion situations, it is clear that to ensure stable operation and meet service requirements the PRMA++ operation parameters can not be set independently. In fact, mismatch among operation parameters may run PRMA++ into instability, i. e. reservation channels become congested due to excessive collisions and retransmission activity, while information channels tend to be idle since bandwidth can not be allocated to mobile stations. In [9] stability problems and protocol inefficiency for a use of PRMA++ in the case of speech and data (Telnet) traffic have been identified. Nevertheless, in the literature on PRMA++ performance, little attention has been given to the identification of a stable operation region.

For a given frame structure, the stability region for PRMA++ depends on both the permission probability and the number of R-slots. Given the inadequacy of the simulation analysis for studying problems related to system stability, we defined a PRMA++ model for speech traffic[1] and solved it (see Appendix) by using the Equilibrium Point Analysis (EPA) [11] method. Basing on this model and the operation parameters of Table 1, results useful for identifying the stability region are given in Fig. 2.

Table 1 . Operation parameters for the SIR protocol.

Slots per frame	72
R-slots	2 - 16
I-slots	55 - 69
FP-Ack slots	1
Talkspurt Mean Length	1 s
Silence Mean Length	1.35 s
Packet Dropping Delay Thres.	10 ms
Permission probability	0.1 - 0.35
Carrier Bitrate	1.8 Mbit/s

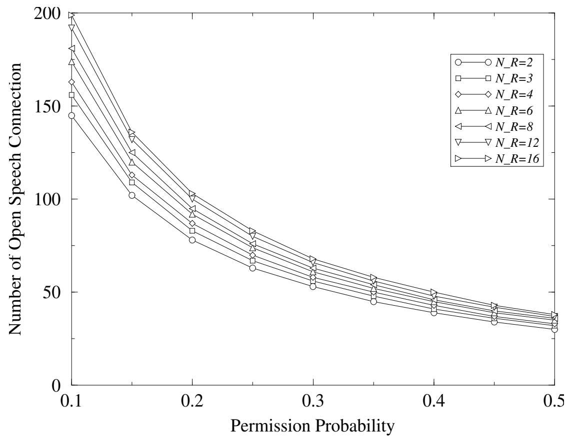

Fig. 2. Stability boundaries for PRMA++ speech operation.

Although the actual results apply to a microcell environment, the indications they convey have general validity. Fig. 2 shows the

[1] Speech sources are modelled according to [12].

region of stable operation for PRMA++ (the area below each curve) as a function of the number of R-slots (N_R) when speech is the only service. The region boundary is identified as the maximum number of speech connections that can be supported for any given value of p.

From Fig. 2 it is realized that by increasing the number of R-slots, the maximum number of sources still enabling the system to operate in a stable region increases, although not linearly with the number of R-slots. The figure also indicates that there is an inverse relationship between the number of allowed sources and the permission probability, i. e. the higher the number of speech connections the lower the upper bound of the range for the permission probability. While Fig. 2 delimits the stability region, it does not provide indication about the system performance or optimal setting of operation parameters. In [5] it is shown (via simulation analysis) that from the packet dropping probability point of view, four R-slots is the optimal choice. Furthermore, our simulation analysis has indicated that P_{drop} is practically insensitive to p provided the system operates in the stability region since in that case the probability of a request collision is generally very low. In other words, the operation of PRMA++ is not critically dependent on the value of the permission probability provided this is confined to the allowed range.

In particular, when the other operation parameters are as indicated in Table 1, setting the number of R-slots to 4 (and the number of I-slots to 67), the maximum number of speech sources corresponding to P_{drop} = .01 is 145. Entering Fig. 2 with this number of sources, the maximum allowed value for p slightly exceeds .1. Based on these considerations, we chose to assume throughout the paper p = .1 and four R-slots as operation parameters for speech handling.

3. The SIR Access Protocol

3.1 Protocol Rationale

In previous studies, [12]-[14], on the integration of voice and data with PRMA, voice and data connections contend for all the free slots. Whether data connections contend on a packet basis ([12],[14]) or they obtain reserved bandwidth for a burst transmission ([13]), a major disadvantage of these schemes is that voice traffic performance is conditioned by the data traffic load. More recently, a solution to this problem has been proposed in [15] by associating voice and data transmission with distinct bandwidth regions, while the boundary between the regions is dynamically varied according to an optimization algorithm. However, no protocol is explicitly defined for data transmission. In [16], a voice/data channel access protocol has been proposed whereby data traffic is transparent to speech connections and does not affect speech traffic performance. The paper defines a hybrid reservation protocol for data traffic: although a collision-free reservation scheme is proposed, a random access component is preserved to enhance system performance at low loads. Bandwidth reservation for data traffic is performed by allocating a fixed number of slots (possibly more than one) to each data connection on a frame basis.

As currently defined, PRMA++ leaves ample freedom as to how allocate radio resources when handling mixtures of speech and data. In [9] the performance of PRMA++ is analyzed for a mix of speech and Telnet data traffic. In the reference, the PRMA++ operation initially assumed is that I-slots would be rigidly partitioned into slots reserved to speech and slots reserved to data, whereas R-slots would indifferently be used for speech and data. While I-slots would be associated with speech connections for the whole duration of a burst (according to canonical PRMA++ prescriptions), data connections with packets to be transmitted would be served by making use of the dedicated I-slots in a Round-Robin fashion. The reasons for poor performance of the proposed use of PRMA++ are essentially: i) equitable handling of speech and data bandwidth requests; ii) inefficient use of bandwidth consumed in attempting to notify frequent requests associated with data traffic and resolve incurred collisions. Moreover, use of common signalling channels for speech and data bandwidth requests exacerbates poor performance.

The above results do not challenge PRMA++ as such, but rather its straightforward use for accommodating mixtures of speech and data. PRMA++ remains an efficient and flexible mechanism for accommodating speech traffic requirements and candidates as an essential component of an evolved protocol for handling integrated speech and data applications.

Based on the above considerations, a protocol evolved from PRMA++ is proposed. The new protocol, SIR, tries to combat the mentioned deficiencies by:

- segregating reservation channels for speech and data;
- enforcing a scheduled (Round-Robin) access to data signalling channels;
- servicing data connections in a Round-Robin fashion by allocating bandwidth not currently used by speech, on top of a minimum (possibly zero) guaranteed bandwidth.

The details of SIR access protocol are given in the following.

3.2 Frame Format

SIR adopts the PRMA++ physical structure. However, in order to accommodate speech and data, SIR introduces a further classification of slots on the uplink. Specifically, the total number, N, of slots in a frame is partitioned into two sets. The first set has N_V slots which are allocated to speech, the other set has $N\_D = N\text{-}N\_V$ slots which are for exclusive use of data. Similarly as with PRMA++, the N_V-slots are organized into information (VI-slots) and reservation (VR-slots) slots and are used in a similar fashion. (VR-slots are also used for connection set-up, both speech and data, and handover handling). The N_D-slots are also organized into information (DI-slots) and reservation (DR-slots) slots. In addition, data can use VI-slots momentarily not used for speech transmission. As opposed to VR-slots which are accessed on a contention basis, DR-slots are accessed according to a Round-Robin algorithm and are thus contention-free. Moreover, also DI-slots (plus momentarily unused VI-slots) are accessed on a Round-Robin basis, although with a different pace as explained in the following Section.

Slots unused by speech and available to data are notified on the downlink in a dedicated slot at the start of each frame by means of an allocation bit pattern. The bits in the allocation pattern are in a one-to-one correspondence with the slots in the frame, i. e. the value (0/1) of each bit represents the status (used/unused) of the associated slot. An example of the SIR frame structure is shown in Fig. 4. Table 2 summarizes the use of slots in the SIR protocol.

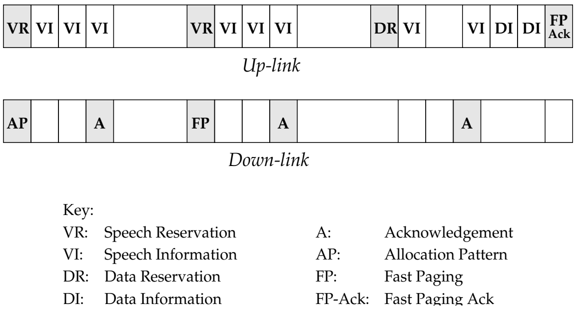

Fig. 3. SIR frame structure and slot types.

Table 2. Use of slots with the SIR protocol.

SIR phase	Speech	Data
Connection set-up	VR-slots, access contention	VR-slots, access contention
Bandwidth request notification	VR-slots, access contention	DR-slots, contention free (Round-Robin)
Transmission	VI-slots, contention free	Dedicated DI-slots, if any, plus unused VI-slots, contention free (Round-Robin)

3.3 Protocol Description

The concepts of "open connection" and "active connection" are recurrently used in the following description. An open connection is a connection set-up between a mobile terminal and the base station for the exchange of user data. During the time an open connection is maintained, traffic sources alternate "activity" and "silence" periods. When workload resulting from activity periods has to be transmitted (from the mobile to the base), a connection is said to be active.

3.3.1 Data Connection Set-Up and Tear-Down Phases

Mobile originated speech and data connections use the same set-up procedures. In particular for data, contention access with retransmission is performed in VR-slots to transmit a request control packet for connection set-up. Once successful, the request control packet is acknowledged by the base station, in the A-slots paired with VR-slots, with another control packet containing:

- an identifier, say {i}, assigned to the data connection just opened, which is unique within the cell controlled by the base station;
- the value of a variable indicating the number of open data connections (NOC). Note that the value of NOC includes the data connection just opened and is broadcast by the base station.

Mobile initiated tear-down is performed through the transmission of a control packet to the base station in reserved DR-slots (see below). Upon receiving this control packet, the base station updates NOC whose use by each station is described in detail in Section 3.3.3.1 on *Request Algorithm.*

3.3.2 Data Bandwidth Reservation

Since speech is handled according to PRMA++ and data handling will not affect speech traffic, the description of user information transport will focus on data only. Upon arriving at station *{i}* [2], a

[2]The same identifier *{i}* is used to address a mobile or the connection associated with it.

message is segmented into packets, the size of which fits the payload transmitted in one slot. Packets then join a local queue, L_Q(i), which is left when packets are scheduled for transmission after station *{i}* receives authorization from the base station. To obtain this authorization a station executes the *Request* algorithm which basically consists in sending to the base station a control packet (using a DR-slot) carrying the number of packets waiting in the local queue. The station then waits for a control packet from the base (in the A-slot paired with the DR-slot) informing on the current number of active connections, NAC. At the end of this information exchange the open data connection becomes *active* and remains such until the last authorized packet has been transmitted. When this occurs, the data connection becomes *inactive* if the local queue is empty, otherwise another request procedure is immediately initiated.

When a data connection becomes active, it shares DI-slots (plus momentarily unused VI-slots) with all the other active data connections. To achieve this, all the stations execute the Transmission algorithm which makes use of NAC. The number of active data connections may change at the end of each slot since an open connection may become active or an active connection may become inactive. For this reason, a variable local to each station, denoted by *Q*, is initialized to the NAC value when the data connection becomes active, and it is updated to keep track of the changes before the beginning of the subsequent slots on the basis of information broadcast by the base station.

The NOC data connections are then partitioned into two sets, one including the Q active connections, and the other containing the NOC-Q inactive connections.

To simplify the protocol description, we assume that each mobile is associated with only one (speech or data) connection.

3.3.3 User Information Transport

The processes associated with information transport are: i) a periodic scan of the open data connections to collect bandwidth requests (Request algorithm); ii) allocation of the available bandwidth to meet the collected requests (Transmission algorithm). By design, both processes are managed so as to avoid contention and are implemented in a distributed fashion. With the Request algorithm, DR-slots are cyclically offered for dedicated use to each of the open data connections for collecting the accumulated bandwidth request since last notification. With the Transmission algorithm, DI-slots are used cyclically by the active data connections: i.e., one DI-slot is offered for use to each of the active data connections for packet transmission.

To access DR-slots, station{i}, maintains a counter, denoted DR_POS(i), which is initialized to the NOC value at the time the data connection is opened. This counter is decreased by one every time station{i} observes a DR-slot. When DR_POS(i) reaches the value of one, the next DR-slot can be used by station{i} for transmission. After completion of transmission DR_POS(i) is initialized again to the current NOC value. This process is repeatedly executed during the time a connection remains open.

Depending on the current value of NOC and the number of DR-slots in a frame, a complete scan of all open data connections may take one or more frames. Clearly, the price paid for contention-free request collection is that not all DR-slots will necessarily be used at each scan, and the trade-off between bandwidth utilization and timeliness in identifying bandwidth requests is an operation decision.

With the Transmission algorithm, the DI-slots (plus those VI-slots not currently utilized by speech) are offered for use of active data connections. These slots, denoted as D-slots for conciseness' sake in the following, are allocated in a cyclic fashion, with one D-slot associated with each active data connection. Specifically, when the open connection at station{i} becomes active, station{i} updates a counter, T_POS(i), to the *Q* value. This counter is decreased by one every time station{i} observes a D-slot (all stations observe D-slots at the same time). When T-POS(i) reaches the value of one, the next D-slot can be used by station{i} for packet transmission. After completion of transmission T-POS(i) is initialized again to NAC value. This process is repeatedly executed during the time a connection remains active.

The Transmission algorithm is transparent to the frame begin/end demarcation, and basically allows to use any residual capacity.

The key parameters for SIR operation are summarized in Table 3. The technicalities of the Request and Transmission algorithms are illustrated in the following.

Table 3 - Key parameters for SIR operation.

NOC	Number of open connections
NAC	Number of active connections
Q	Number of active connections, i. e. with at least one packet in the transmission buffer
R_BKG	Number of packets in the request buffer
T_BKG	Number of packets in the transmission buffer
DR_POS	Counter for controlling transmission of requests for data bandwidth on DR-slots
T_POS	Counter for controlling transmission of data packets on reserved bandwidth, DI- or VI-slots

3.3.3.1 Request Algorithm

This algorithm is executed by those stations which have an open data connection any time they observe a DR-slot. The total number of packets contained in Q_L(i) will be denoted by R_BKG(i). The *Request Algorithm* is depicted in Fig. 4.

Depending on whether R_BKG(i)=0 or R_BKG(i)>0, mobile{i} can be in the *Wait_DR_Idle* or *Wait_DR_Active* state respectively. While in these states, station{i} looks for DR-slots. The counter DR_POS(i) is decreased by one, whenever a DR-slot is observed (Transitions 1 and 5). When the value of DR_POS(i) is one, mobile{i} has the right to access the DR-slot, and two different actions can be taken depending on the value of R_BKG(i). If R_BKG=0 (Transition 2) mobile{i} moves to the *Wait_A_Idle* state and then, when the A-slot paired with the DR-slot is observed on the downlink, it updates the DR_POS(i) counter (Transition 3) by using the value of NOC notified by the base station. If, on the other hand, R_BKG>0 (Transition 6), mobile(i) transmits into the DR-slot the *REQ(i,n)* control packet, where {i} is the station (or connection) identifier and *n*=R_BKG(i) is the backlog of the local queue at this time.

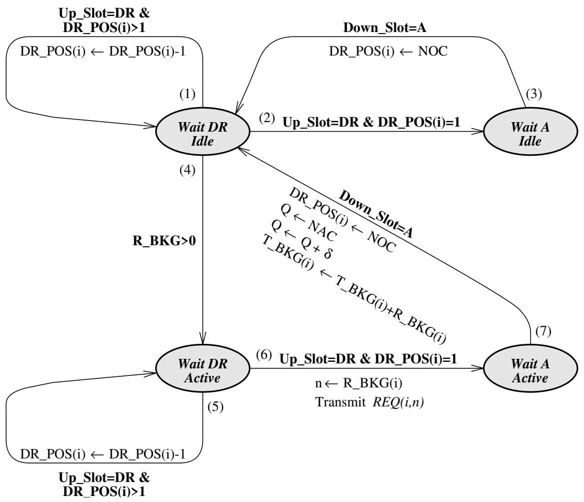

Fig. 4. SIR bandwidth request algorithm.

After transmission of *REQ(i,n)* , mobile{i} moves to the *Wait_A_Active* state. While in this state, mobile{i} waits for the paired A-slot on the downlink. Upon its arrival (Transition 7), station{i} performs the following actions:

- initializes the DR_POS(i) counter to the NOC value (the same as Transition 3);
- moves R_BKG(i) packets from the head of the local queue into another queue, named *transmission queue*, in which there may be other packets ready for transmission (i.e., for which D-slots were previously requested);
- initializes the local variable Q with the NAC value passed by the base station corresponding to the Number of Active Connections at the beginning of the A-slot.
- updates the value of Q by adding the NAC variations (d) occurred during the A-slot (see below).

Station{i} is now back to the *Wait_DR_Idle* state, from which it departs instantaneously (Transition 4) if any messages had arrived (R_BKG>0) when the station was waiting for the ACK in the *Wait_A_Active* state. Otherwise, station{i} remains *Wait_DR_Idle* until the next message arrives.

The number of packets contained in the transmission queue of station{i} is denoted as T_BKG(i), i. e. a connection is active if T_BKG(i)>0). The difference between R_BKG(i) and T_BKG(i) is that the former represents the number of packets queued at the local queue but still unknown to base station, while the latter indicates the packets notified to the base station via the Request algorithm and hence ready for transmission (see Transmission Algorithm).

3.3.3.2 Transmission Algorithm

This algorithm is executed by those stations which have an active data connection any time they observe a D-slot. The purpose of the *Data Transmission Algorithm* (Fig. 5) is to allocate D-slots to active connections in a contention free manner. To achieve this result, station{i} utilizes the T_POS(i) counter and the local variable *Q*.
As explained in the preceding section, Q is set by station {i} when it receives the NAC value in the A-slot paired with the DR-slot in which the *REQ(i,n)* was made. However, to keep with the changes of NAC following the bandwidth usage vicissitudes of the open connections, the base station also broadcasts the single variation, d (d=+1, 0, -1), in the number of active connections[3]. The value of d is embedded in each

[3] Specifically, d=+1 if an open connection becomes active (which can occur at the end of an A-slot carrying the request grant), d=0 if NAC remains unchanged, and

slot in the downlink. Upon sensing d in one slot, a mobile updates the value of *Q* before the beginning of the next slot. The counter T_POS(i) is decreased by one, whenever a D-slot is observed. When the value of T_POS(i) becomes one, mobile{i} has the right to transmit one data packet. Subsequently, if T_BKG(i)>0, T_POS(i) is reinitialized to the current number of active connections maintained by the variable *Q*.

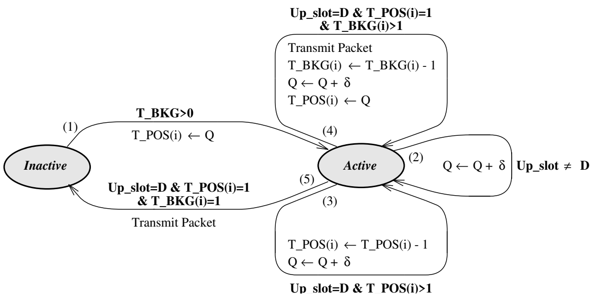

Fig. 5. SIR data transmission algorithm.

The Transmission Algorithm works as follows. When the open connection at station{i} becomes active (T_BKG(i)>0) mobile{i} updates T_POS(i) to the *Q* value and then it moves from the *Inactive* state to the *Active* state (Transition 1). While in the Active state, station{i} monitors the slots in the Uplink frame, and takes one of the following action depending on the slot type:

- if not a D-slot, only the value of *Q* is updated (Transition 2);
- if a D-slot, but it is not yet the station' s turn to transmit (T_POS(i)>1), *Q* and T_POS(i) are updated (Transition 3);
- if a D-slot, and it is the station' s turn (T_POS(i)=1) and, furthermore, there is more than one packet in the transmission queue waiting for transmission (T_BKG(i)>1), then: i) one packet is transmitted; ii) T_BKG(i) is decreased by one; iii) *Q* is updated; and iv) T_POS(i) is re-initialized to the current value of *Q* (Transition 4).

d=-1 if an active connection transmits the last packet in the transmission queue (which can occur at the end of any D-slot).

In the above three transitions the state of the connection remains the same. If, on the other hand, a D-slot is observed and it is the station' s turn to transmit (T_POS(i)=1) but there is only one packet waiting in the transmission queue (T_BKG(i)=1), then this packet is transmitted (Transition 5), T_POS(i) and T_BKG(i) are reset to zero, and the connection state switches to *Inactive*.

3.3.4 Variants for accommodating traffic mixes

Extensions/variations of SIR could be introduced in order to enhance the capacity of the protocol to accommodate the variety of service requirements resulting from handling complex mixtures of applications (e. g. speech, and classes of data and video with different quality of service targets) envisaged for advanced mobile systems. A few examples are given in the following.

Priority based request acknowledgement control

The base station could delay the acknowledgement to some *REQ(i,n)* basing on a priority indicator carried in the request packets. This would reduce the number of active connections, hence reducing the length of the Round Robin transmission cycle. Active connections characterized by a higher priority would then be favoured in getting bandwidth allocated.

Hybrid slot allocation

Data (or video) applications with bandwidth constraints could be accommodated by a hybrid slot allocation, i. e. some of the slots reserved for data transmission could be allocated to the data connections on a circuit switching basis, so as to guarantee a minimum bandwidth, whereas additional bandwidth could then be obtained via the SIR allocation mechanisms. Obviously, this would imply that the definition of SIR bandwidth allocation procedures at call set-up had to be revisited.

In order to investigate the potential of SIR, the analysis of the protocol is performed in the following with reference to the basic SIR capabilities described above.

4 Performance Analysis

4.1 Data Traffic Characteristics

While there are satisfactory models, e. g. [17], to represent the packet arrival process of a speech source (i.e., to model the behaviour of a speech activity detector), for a variety of data applications (e.g., file transfer, database interrogation, etc.) an equally satisfactory modelling is not easily available. Furthermore, a number of studies have shown that in a LAN/WAN environment the packet inter-arrival times of the aggregate traffic clearly differ from exponential [18], [8], [19]. On the other hand, a sound assessment of multiple access schemes for supporting the ongoing standardization activity on advanced mobile systems, [20], [21], needs to be based on realistic traffic characterization and scenarios.

In order to evaluate the SIR capability to handle a mixture of speech and data traffic, we chose to load the system with data traffic generated by interactive applications (e.g., remote terminal access to a data base and remote log-in). Our decision was motivated by: i) the expected importance of these applications in future mobile systems; ii) the packet arrival burstiness inherent to these applications representing severe operation conditions capable of giving conservative indications on SIR performance.

The analytic traffic model underlying our study, [8], fits wide-area TCP live traffic traces of interactive Telnet applications. Although the operation environment for this model is different from the context of mobile systems, it captures aspects of general validity for the user behavior associated with interactive data applications.

A complete model for Telnet traffic would include distributions for session inter-arrivals, message inter-arrivals within a session and message size (in packets). However, the time scale over which session arrivals are normally observed is largely beyond the time horizon of interest for the performance study of the SIR protocol. We therefore do not consider the session inter-arrival process and exercise the protocol with a fixed number of established speech and data connections. Further, we initially assume that messages associated with each session convey just one byte of user data, corresponding to a keystroke. Considering the overhead involved by formatting and coding, this translates into a few (and constant number of) packets. This assumption will then be relaxed, and messages of increased length will be considered.

Following [8], message inter-arrival times are distributed according to a Pareto distribution with parameters α and k:

$$P\{T \leq t\} = 1 - \left(\frac{k}{t}\right)^{\alpha} \quad k, \alpha \geq 0, \; t \geq k$$

The parameters in this model are k=0.1 s and a=0.85 up to the 97th percentile, while for the upper 3% tail the parameters are k=0.15 s and a=0.95. If a≤2, the distribution has an infinite variance, while if a≤1 the mean is infinite as well. Hence, the distribution above is heavy tailed, with infinite variance and mean[4]. This mathematical property accounts for the burstiness of the arrival process causing intense reservation bandwidth activity over signalling channels.

4.2 Performance Metrics

The main difficulty for the analysis of the SIR protocol via analytical models is the high degree of complexity and interdependence of the various processes that involved in the protocol operation. Therefore, to evaluate the performances of the protocol we designed an ad hoc simulator and implemented it in C++ language.

Although data integrity is a key requirement for data transmission, we assume an error free radio channel despite the fact that we are dealing with a very unreliable environment. This is because in the present paper we are focusing on multiple access aspects of SIR as measured by the access delay experienced by messages transmitted over data connections. The *message transport delay* is defined as the time elapsed from message available at the source site to successful transmission of the last packet of the message. Three major components of the transport delay are:

- *reservation delay,*: time elapsed from complete message available at the source site to bandwidth request acknowledgement received at the source on the A-slot paired to the DR-slot in which the request was made;
- *queuing delay:* time elapsed from bandwidth request acknowledgement to start transmission of the first packet of the message; and
- *transmission delay:* time elapsed from start transmission of the first packet of a message to completion of transmission of the last packet of the same message

Fig. 6 shows the breakdown of the message transport delay.

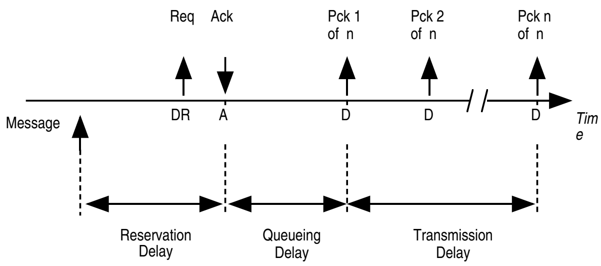

Fig. 6:. Breakdown of message transport delay.

Each component accounts for a particular aspect of the SIR access protocol: reservation delay is a function of the number of DR-slots and the number of *open* data connections; queuing delay is a function of the single data connection offered load and of the number of *active* connections, while transmission delay is a function of the number of *active* connections and of the message packet length.

[4] A distribution is said to be heavy-tailed if $P\{T \geq t\} \propto t^{-\alpha}$ for $t \to \infty$, $\alpha \geq 0$.

4.3 Operation Parameter Setting and Assessment Scenarios

The system parameters for a microcellular environment are inspired from [5], hence the transmission rate is 1.8 Mbit/s and each SIR frame is 5 ms long. Assuming a 13 kbit/s gross bit rate for digitized speech sources, 65 payload bits are needed for each packet transmitted in a slot. Accounting for 59 overhead bits, it follows that each frame has 72 slots. Further, it is assumed that cells are provided with one carrier.

The performance of SIR has been analyzed for various values of *N_D* (*N_D* = 8, 7, ..., 4), adjusting the number of slots reserved for speech transmission accordingly (i.e., *N_V* = 72 - *N_D*). As explained in Section 2, for each setting of *N_V* - by given *N_VR* , i. e. 4 - it is required that the stability region be identified, i. e. the limiting combinations between the maximum number of speech connections and *p* be determined. As a second step, for a choice of *p* (.1) consistent with stable system operation for a range of speech connections the behaviour of P_{drop} is analyzed and the number of speech connections for which P_{drop} = .01 determined. Results related to this second step are shown in Fig. 7 which is then the basis for the calibrating the number of the speech connections to the actual values of *N_D* in all subsequent experiments.

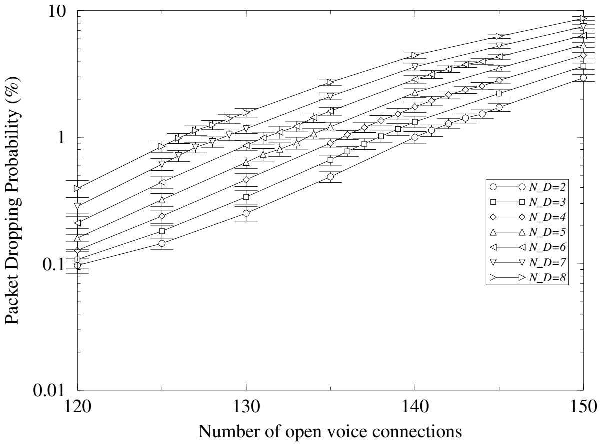

Fig. 7. P_{drop} vs the number of speech connections for several *N_D* values.

Due to the statistical characteristics of speech "activity" and "silence" periods, a fraction of VI-slots remains unused (on the average). This "unused" capacity could be exploited by data traffic with relaxed delay constraints. Table 4 shows the average number of slots left unused (E{VI free slots}) by speech connections. As it can be noted, the number of unused slots remains almost constant by varying *N_D* .

Table 4. Maximum number of allowed speech connectionsand average number of unused VI-slots as a function of *N_D*.

N_D	Max # of Voice Sources	E{VI free slots}
4	135	6.11
5	133	6.02
6	131	5.87
7	128	6.01
8	126	5.83

Each *N_D* value defines a set of scenarios, each one characterized by the number of slots allocated to data reservation (*N_DR)* and data transmission *(N_DI)* , obviously, *N_D* = *N_DR* + *N_DI.*

The methodology adopted to analyze the performance of SIR in handling data is as follows. We assess SIR starting from a basic set. Variations are then introduced to estimate the sensitivity of SIR to different load and operating conditions. Fig. 8 shows the relationship between the assessment scenarios.

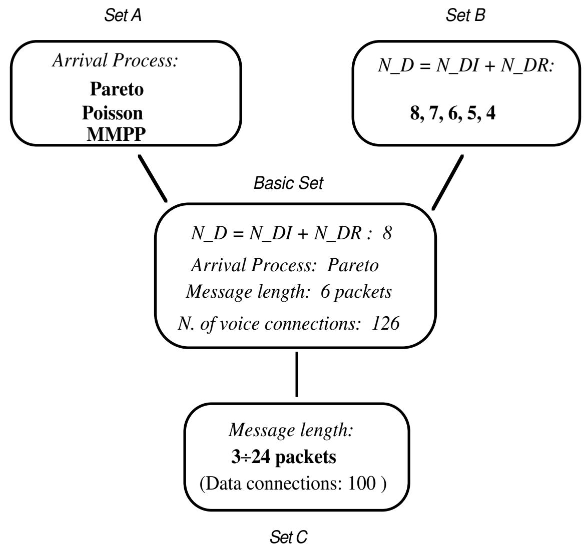

Fig. 8. Scenarios for SIR performance assessment.

For all experiments related to the scenarios shown in Fig. 8, the system is exercised for 150 seconds (including 30 seconds warm-up). The message arrival rate per connection related to each simulation experiment is approximately .8 messages/s[5].

4.4 Basic Set

The transport delays for different scenarios in the basic set as a function of the proportion between DR and DI slots, and varying number of data connections is given in Fig. 9.

[5] Since the Pareto process has infinite mean and variance, the arrival rate depends on the duration of the simulation run.

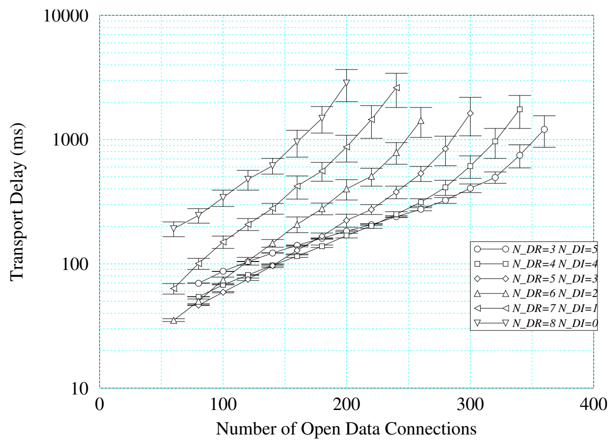

Fig. 9. Transport delay vs number of open data connections for the basic scenario.

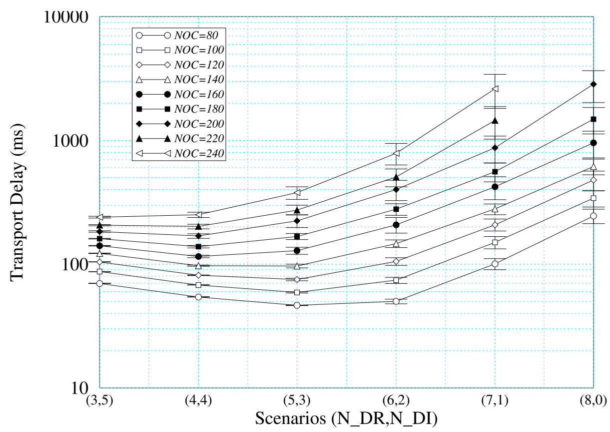

Fig. 10. Transport delay vs *(N_DR, N_DI).*

The figure shows that at low offered load (i. e. small number of data connections) the increase of *N_DR* over *N_DI* results in a better transport delay performance. This can be explained by considering that the increase in *N_DR* reduces the contribution of the reservation delay, which contributes to the transport delay in a substantial way. On the other hand, at higher loads the transport delay increases by increasing *N_DI* over *N_DR*. The existence of an optimal combination *(N_DR, N_DI)* for each number of data connection is indicated by Fig. 10. Hence, the optimal transport delay performance in the basic scenario is obtained by operating the system as indicated by theca envelope of the curves in Fig. 9.

Figures 11 to 14 highlight the influence of the delay components to the transport delay for the scenario associated with the envelope curve. Fig. 11 shows that the reservation delay grows linearly with the number of open data connections. Furthermore, as expected, the higher *N_DR* the lower the reservation delay. In particular, the reservation delay decreases linearly with the reservation bandwidth: i.e., *N_DR*.

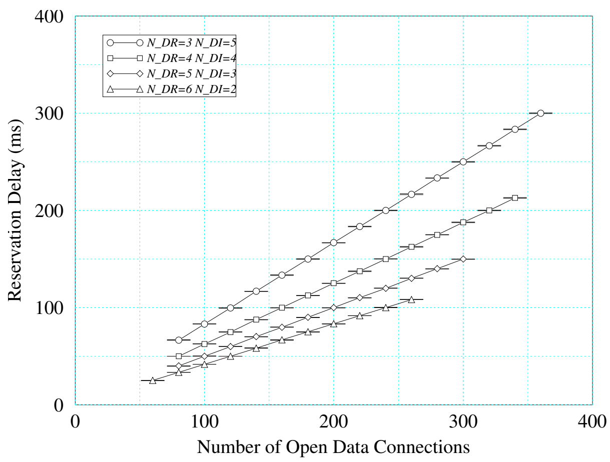

Fig. 11. Reservation delay vs the number of open data connections.

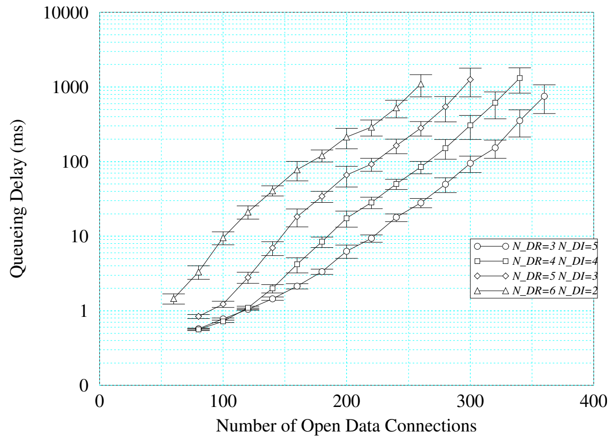

Fig. 12. Queuing delay vs the number of open data connections.

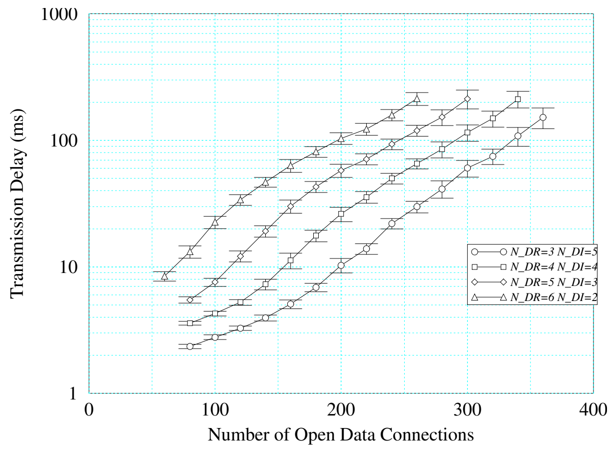

Fig. 13. Transmission delay vs the number of open data connections.

This can be explained by considering that, even if the interarrival time of each data source is distributed according to Pareto, the aggregate message arrival process results in a Poisson process[6]. Since the number of open data connections is kept constant throughout the simulation, the interval between two consecutive DR-slots assigned to a mobile increases linearly with the number of open data connections. Hence, if the arrival instant of a message (independently

[6] This is due to the well known results on the superposition of random processes, [22].

of the connection it is related to) is uniformly distributed between two consecutive DR-slots, then the delay to the next DR-slot grows linearly as well. Hence, the reservation delay is exactly half the time interval between two consecutive DR-slots assigned to a mobile.

Queuing (Fig. 12) and transmission delay (Fig. 13) show the typical throughput-delay behaviour. Furthermore, it is worth noting that delay curves do not intersect each other, and queuing and transmission delay increase with the reduction in the available bandwidth (i.e., *N_DI*).

To summarize: i) reservation and "queuing plus transmission" delays exhibit opposite variation with respect to *N_DR*, and there is an optimal value of *N_DR* which minimizes the transport delay (see Fig. 14); ii) the transport delay depends on the number of data connections and its optimization is achieved by a suitable choice of *N_DR* an a case by case basis.

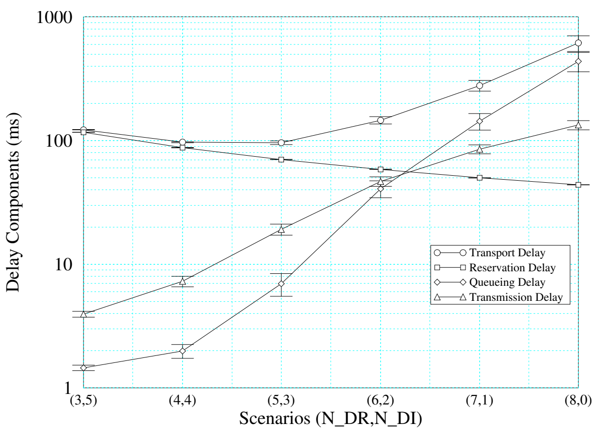

Fig. 14. Delay components vs *(N_DR, N_DI)*.

4.5 Set A

Experiments in the set A are intended to analyze the influence of the message interarrival distribution on the transport delay and its components. Specifically, two additional types of arrival processes are taken into consideration: Poisson and MMPP. In both cases the aggregate arrival rate related to any given number of data connections is the same as that generated under the Telnet arrival process (i.e., approximately 0.8 messages/s). The rationale behind the MMPP choice is to investigate the influence of the correlation between successive message arrivals on the various delay measures. Two 2-MMPP processes are considered and Table 5 indicates the related parameters. In both processes the arrival rate λ_2 in state 2 is assumed equal to zero and hence the processes alternate between activity and inactivity phases (see Fig. 15). However, the process MMPP2 exhibits a higher "burstiness" than MMPP1 since the ratio between the average inactivity period $1/\sigma_1$ and the average activity period $1/\sigma_2$ for the former is twice than the latter.

Table 5. MMPP arrival process parameters.

Arrival Process Parameters	σ_1 **[s^{-1}]**	σ_2 **[s^{-1}]**	λ_1 **[msg/s]**	λ_2 **[msg/s]**
MMPP1	0.306	0.0439	6.376	0.0
MMPP2	1.0	0.067	12.8	0.0

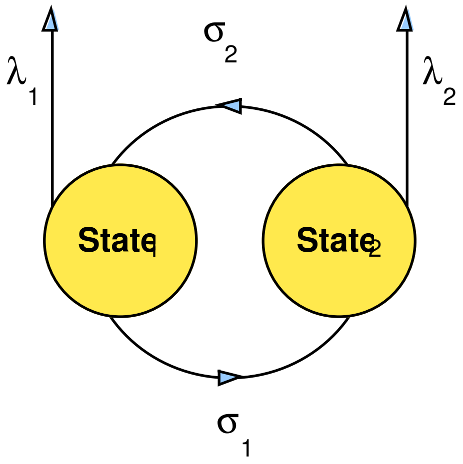

Fig. 15. MMPP arrival process.

For all the considered arrival processes, the delays for the various scenarios as a function of the number of open data connections exhibit the behaviour similar to those for the basic scenario. Fig. 16 shows the envelope curves corresponding to the optimization of the transport delay for all considered arrival processes.

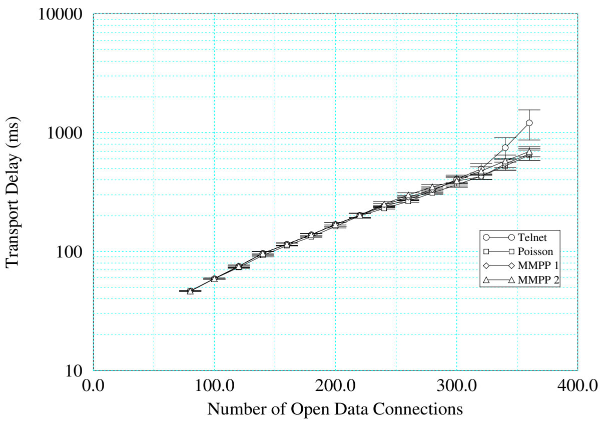

Fig. 16. Transport delay vs the number of open data connections.

As it is apparent fro the figure, the envelope curves are very close to each other for low loads while the Telnet envelope diverges from the others at high loads. In order to explain, this behaviour, Figures 18 and 19 show respectively the queuing and transmission delay for an optimal choice of the allocation parameters *N_DR* and *N_DI* at high loads, i. e. *N_DR = 3, N_DI = 5)*

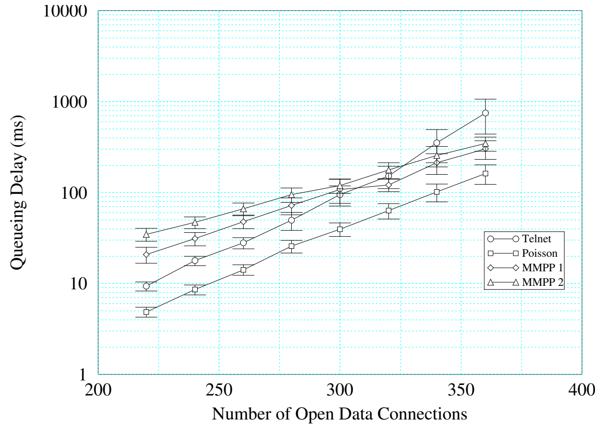

Fig. 17. Queuing delay comparison.

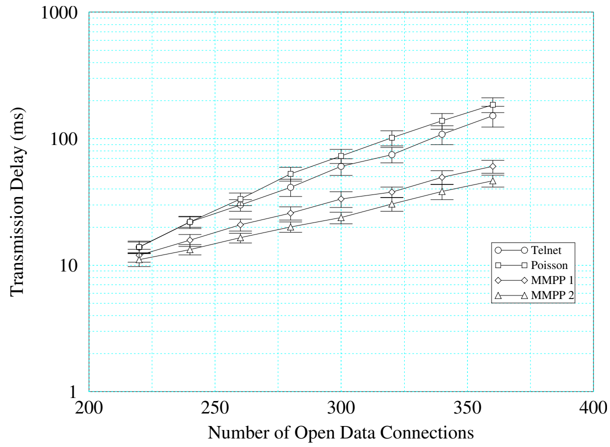

Fig. 18: Transmission delay comparison.

For the Poisson and MMPP arrival processes, Fig. 17 indicates that, over the whole range of considered loads, the queuing delay decreases when the burstiness decreases, while the opposite is true for the transmission delay (Fig. 18). This can be "intuitively" justified by observing that, when the burstiness decreases, the message arrivals are more uniformly distributed between DR-slots. It could then be argued that in the composite arrival process associated with data traffic sources, the number of request packets

REQ(i,n) increases while *n* (i.e., the number of messages per request) decreases and, consequently, the expected number of active data connections increases, too. The above reasoning is justified by Fig. 19 and Fig. 20 showing, respectively, the average value of *n* and the average number of active data connections as a function of the offered load.

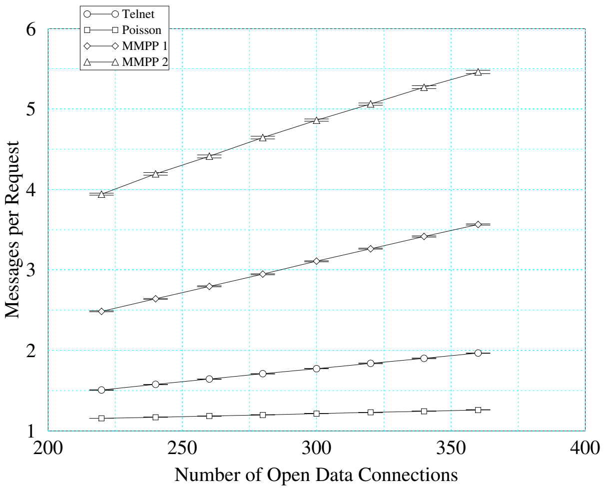

Fig. 19: Average number of messages per REQ(i,n) packet.

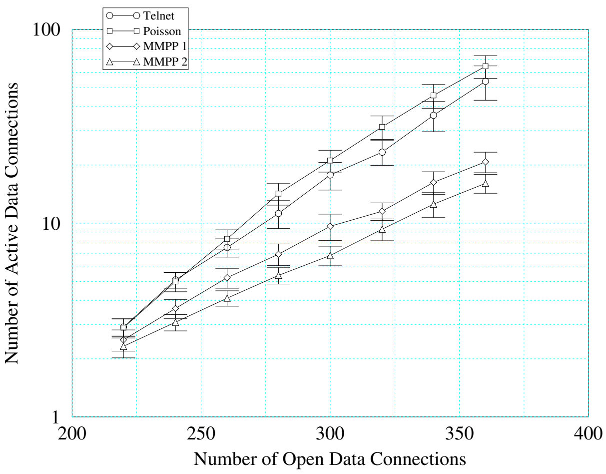

Fig. 20. Average number of active data connections.

From the above considerations it can then be concluded that when the burstiness decreases:

- the transmission delay increases and in accordance with the Round Robin scheduling property of increasing transmission delay when the number of users increases;
- the buffer occupancy decreases (see Fig. 21), since messages join the transmission queue with batches of decreasing size *n*.

By combining together the opposite trends related to queuing and transmission delays for MMPPs there is a sort of compensation and this results in a transport delay close to the delay related to the Poisson arrival process (see Fig. 16).

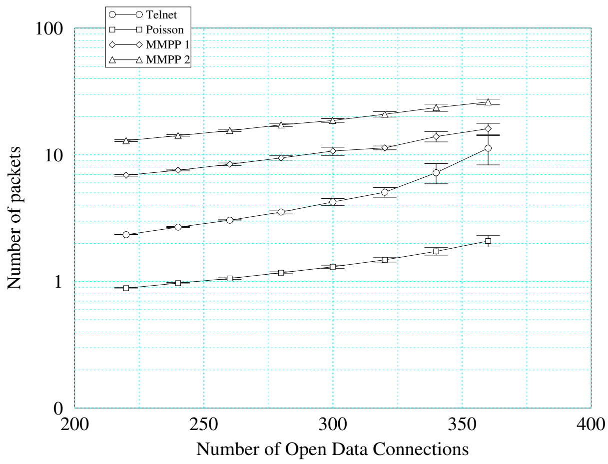

Fig. 21. Average queue length experienced by a message joining the transmission queue.

The above line of reasoning holds for Poisson and MMPPs. Since Pareto exhibits a burstiness which is in between Poisson and MMPP (see Fig. 19), it would be reasonable to expect that, for Pareto too, the transport delay would be close to that associated with Poisson and MMPP. However, from Fig. 16, it clearly appears that this is only valid for small to moderate offered loads, whereas for higher loads the transport delay diverges remarkably from the Poisson and MMPP cases. This is mainly due to the queuing delay component of the transport delay as shown in

Figures 17 and 18. In fact, with the Pareto arrival process, a not negligible fraction of connections do not transmit within the observation interval in the scope of our experiments (because of the effect of the heavy tail of the related distribution). This is not the case with the other arrival processes. Therefore, with Pareto, the aggregate offered load is then related to a number of connections lower than the number of open data connections. This results effectively in an increase of the queuing component of the transport delay.

4.6 Set B

The aim of the experiments in the set B is to analyze how the amount of bandwidth dedicated to data affects the transport delay. For each value of *N_D,* the optimum operating arrangement (corresponding to the optimum apportionment of *N_DR* and *N_DI*) is represented in Fig. 22 in terms of the envelope delay characteristics. For comparison purposes, the figure also shows the envelope curve related to the basic set. Since the curves do not intersect, it can be concluded that once the system operation is optimized for a given *N_D* a better performance can only be achieved by increasing the amount of bandwidth dedicated to servicing data.

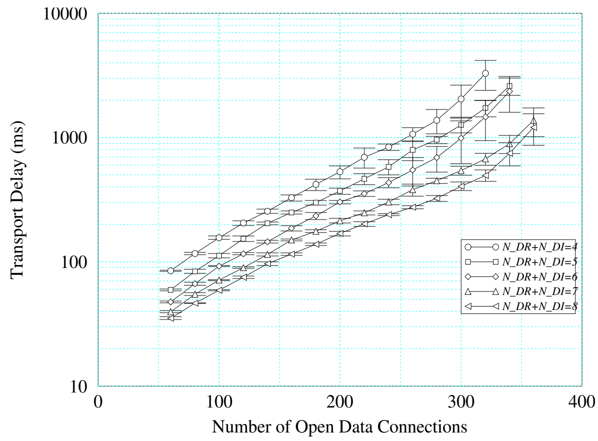

Fig. 22. Envelope comparison for scenarios in set B.

4.7 Set C

With set C, the aim is to study how the message size affects the SIR performance. To do this, the number of open data connections is kept constant (100) while the number of packets per message varies (from 3 to 24 in steps of 3). Results are depicted in Fig. 23 for several *(N_DR, N_DI)* combinations, i.e., those combinations contributing to the envelope curve for the basic set of Fig. 9.

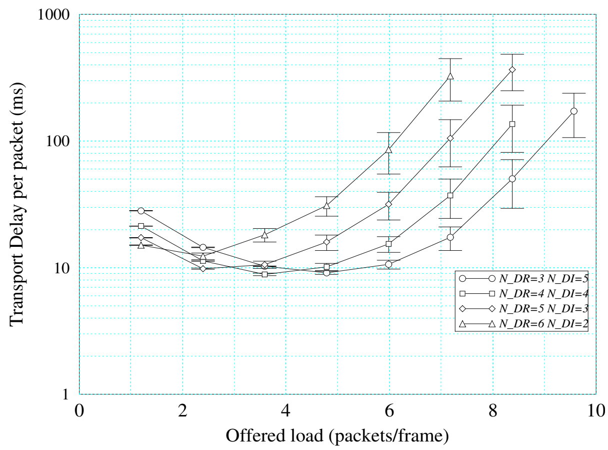

Fig. 23. Transport delay per packet vs message length.

The results in the figure can be explained by observing that:

- the reservation delay depends on the number of open data connections and not on the associated traffic. Since the time interval between two consecutive DR-slots is constant, the reservation delay per message is constant too. Therefore, the packet reservation delay decreases hyperbolically as long as the number of packets per message increases.
- the queuing and transmission delay components of the transport delay increase as the (aggregate) offered load increases.

From the above observations it follows that at low loads (see Fig. 23) the transport delay decreases since the reservation delay prevails over the queuing and transmission delay components. On the other hand, when the offered load increases the transport delay increases too, since the queuing and transmission delay components prevails over reservation delay. It is worth noting that, at moderate and high loads the optimal curve results from the combination *(N_DR = 3, N_DI = 5)*, i.e., the same combination leading to the optimal envelope curve, at moderate and high loads in the basic set.

Fig. 24 shows a comparison, at high load conditions, between the optimal transport delay curves for the basic and C set[7].

[7] In the basic class the number of packets per message is fixed (6) while the number of open connections varies. In the optimal curve for class C, the number of open connections is fixed (100) while the number of packets per message varies.

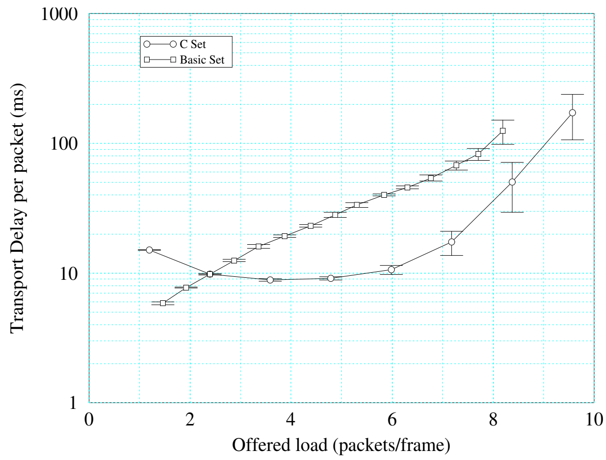

Fig. 24. Transport delay comparison.

From Fig. 24 follows that the SIR MAC protocol performs better with few sources producing a large amount of data packets rather than more sources with a few data packets per message.

5. Conclusions

Traffic capacity is a major evaluation aspect being considered in the global and regional standardization activity on the selection of the radio transmission technology(ies) for use in third generation mobile systems. Multiple access mechanisms in the radio access subsystem have a key impact on traffic capacity and are being designed so as to reconcile meeting the requirements of integrated services and efficient use of spectrum. In line with these objectives, the paper has introduced a multiple access protocol, SIR (Service Integration Radio Access), which has the potential for achieving those desirable properties while ensuring stable operation over a wide range of traffic scenarios. The protocol, evolved from a former protocol studied in the context of the European UMTS, accommodates data traffic in a contention-free manner. Since the protocol is introduced with reference to a data traffic characterization derived from live traffic traces of highly interactive applications, it is believed that the study gives conservative indications for a variety of other applications with relaxed interaction between end-users.

Appendix

The dynamic behaviour of the PRMA++ protocol with speech sources has been investigated by means of the Equilibrium Point Analysis [11]. The system under consideration comprises M speech connections which access a common uplink channel using PRMA++ to transmit packets generated by speech sources. Each source is modeled by a two state discrete time Markov chain as shown in Fig. A.1.

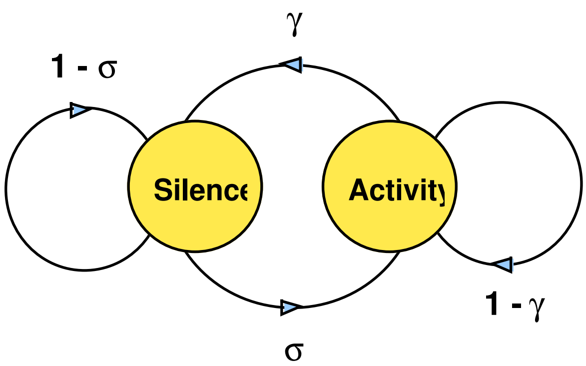

Fig. A.1. Speech source model.

Transition instants of this chain occurs immediately before the beginning of each R-slot. Hence, defining σ as the one-step transition probability from Silence to Activity state and γ as the transition probability in the opposite direction, we have:

$$\sigma = 1 - e^{-\tau/t_S}$$
$$\gamma = 1 - e^{-\tau/t_T}$$

where t_S and t_T are, respectively, the mean sojourn time of a source in Silence state and in Activity state, and τ is the interval between two consecutive embedding instants. Since frame duration is constant (5 ms) and R-slots are uniformly distributed in a frame, it follows that τ is inversely proportional to the number of R-slots in a frame.

We model the whole system as a multidimensional discrete state discrete time Markov chain. Transition instants of the chain are located at the beginning of each R-slot (as a consequence, transitions of the system state are always consecutive to source state transitions).

According to this model, each speech connection can be in one of the following four states:

- *silence* (S), speech source is inactive and there are no backlogged packets to transmit;
- *first request* (FR), the speech connection enters this state whenever the underlying source becomes active. It remains in this state for just one step, independently of the other connections state, since allocation request can be transmitted with permission probability one for just one time (the first one) in a R-slot;
- *contention* (C), the connection enters this state if the first request transmission has collided, and leaves it the first time request transmission is successful;
- *wait-for-allocation/transmission* (W/T), the connection is either waiting for the allocation of the I-slot by the base station, or allocation has been gained and speech packets are being transmitted (for the purposes of this study, it is

not necessary to distinguish between this two logically different connection states).

The state of the whole system, comprising a set of M speech connections, is represented by the random vector:

$$\mathbf{X} = [S, FR, C, WT],$$

where the components S, FR, C and WT are random variables. representing the number of connections in states silence, first request, contention and wait-for-allocation/transmission, respectively.
State transitions related to a single connection, along with the associated probabilities, are depicted in Fig. A.2.

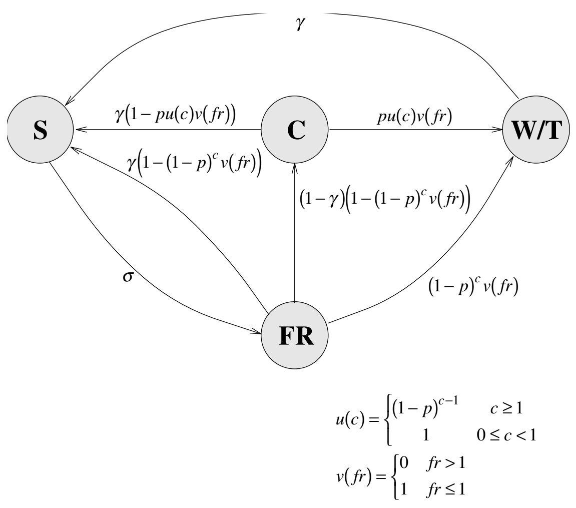

Fig. A.2. State transitions associated with a speech connection.

According to the EPA approach, equilibrium points are derived by imposing the condition that the expected increase in the number of connections at each state is null. From the above, it follows that an equilibrium point of the system is a set of four values $x = [s, fr, c, wt]$, where s, fr, c and wt represent the equilibrium values for the number of connections in states S, FR, C, and W/T respectively. By equating the expected outflow and inflow at each state, we obtain the following equations:

$$\sigma \cdot s = fr \qquad \text{(A.1)}$$

$$(1-\gamma)\left(1-(1-p)^c v(fr)\right)fr = \left(\gamma(1-pu(c)v(fr)) + pu(c)v(fr)\right)c \qquad \text{(A.2)}$$

$$(1-p)^c v(fr)\cdot fr + pu(c)v(fr)\cdot c = \gamma \cdot wt \qquad \text{(A.3)}$$

$$\gamma \cdot wt + \gamma(1 - pu(c)v(fr))c + \gamma\left(1-(1-p)^c v(fr)\right)fr = \sigma \cdot s \qquad \text{(A.4)}$$

Only three of the four equations are independent. The fourth independent equation translates the condition that the total number of connections is kept constant:

$$s + fr + c + wt = M. \qquad \text{(A.5)}$$

From A.4 and A.5, we obtain:

$$s = \frac{\gamma}{\gamma+\sigma}M - \frac{\gamma}{\gamma+\sigma}v(fr)\left(p\cdot c\cdot u(c) + (1-p)^c fr\right)$$

and then, from A.1:

$$fr = \frac{\sigma\cdot\gamma}{\gamma+\sigma}M - \frac{\sigma\cdot\gamma}{\gamma+\sigma}v(fr)\left(p\cdot c\cdot u(c) + (1-p)^c fr\right) \qquad \text{.(A.6)}$$

Under the assumption that $fr \le 1$, the above set of equations (A1-A4) simplifies. To show this, we substitute $v(fr)$ with its value (which is equal to one) in all equations. From A.2, we obtain:

$$(1-p)^c \cdot fr = fr - \frac{1}{1-\gamma}\left(\gamma(1-pu(c)) + pu(c)\right)\cdot c \cdot$$

Substituting this expression in A.6, after some algebra, we finally obtain:

$$fr = \frac{1}{1+\frac{1}{\sigma}+\frac{1}{\gamma}}\left(M+\frac{\gamma}{1-\gamma}c\right) \qquad \text{(A.7)}$$

This expression is valid if and only if the initial assumption ($fr \le 1$) is true. Since the equilibrium value for c is always lower than M, maximizing the expression for fr we obtain a sufficient condition for the validity of A.7:

$$M \le (1-\gamma)\left(1+\frac{1}{\sigma}+\frac{1}{\gamma}\right) \qquad \text{(A.8)}$$

For a typical configuration (microcellular environment, 4 R-slots per frame), A.8 results in:

$$M \le 1879.65$$

which is a condition met for the considered range of M. Hence, in the following, in order to obtain the equilibrium points of the system, we can solve the simplified set of equations obtained from A.1-A.5 assuming that $fr \le 1$.

For this set of equation, the following holds:

- there is always a solution for which $c = 0$. In fact, if we use this value in equations A.1-A.5, the system becomes linear, with just one solution for the other three variables s, fr and wt.
- on the other hand, if $c \ne 0$, then, by substitution, we derive from A.1-A.5 the following non-linear equation:

$$M = pu(c)\cdot c + \left(\frac{1}{\sigma}+\frac{1}{\gamma}+(1-p)^c\right)\cdot \frac{\gamma(1-pu(c))+pu(c)}{(1-\gamma)\left(1-(1-p)^c\right)}c \qquad \text{(A.9)}$$

Solutions of A.9, each of them an equilibrium value of c, can be derived graphically with the help of Fig. A.3, where the right-hand side of A.9 is shown as a function of c. Different curves are obtained by varying the value of p, while the number of R-slots per frame, and then σ and γ, is constant and equal to 4.

By intersecting a curve with a horizontal load line corresponding to a fixed number of speech connections, we obtain the corresponding solutions of A.9. For the sake of generality, in Fig A.3 we have also shown a vertical line (bold) which accounts for the solution corresponding to $c = 0$.

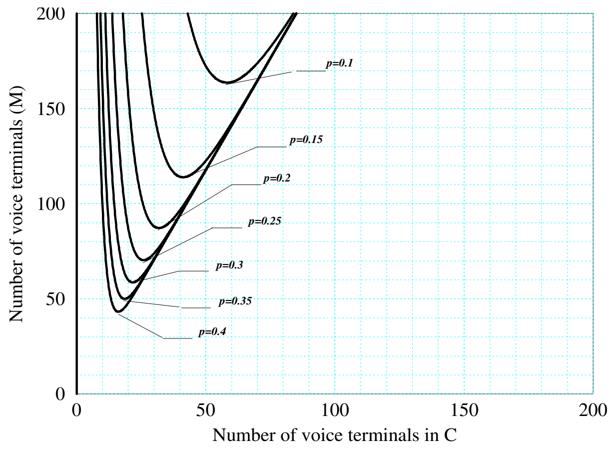

Fig. A.3. Solution of the equilibrium state equations.

The system operation is stable if and only if there is only one (stable) equilibrium point. By looking at Fig. A.3, we observe that, for a given system configuration, there is a critical value of the number of speech connections, over which the system has multiple equilibrium points, and hence is unstable. By considering different system configurations, we have derived the stability curves of Fig. 2.

References

[1] D. J. Goodman, R. A. Valenzuela, K. T. Gaylard, and B. Ramamurthi, "Packet Reservation Multiple Access for Local Wireless Communications", *IEEE Trans. Commun.*, vol. COM-37, pp. 885-890, August 1989.

[2] J. De Vile, "A Reservation Multiple Access Scheme for an Adaptive TDMA Air Interface", in *Proc. of Fourth WINLAB Workshop on Third Generation Wireless Information Networks*, New Jersey, USA, Oct. 1993.

[3] A. Urie, M. L. Streeton, and C. Mourot, "An Advanced TDMA Mobile Access System for UMTS", *IEEE Personal Communications*, vol. 2, pp. 38-47, February 1995.

[4] S. Nanda, D. J. Goodman, and U. Timor "Performance of PRMA: A Packet Voice Protocol for Cellular Systems", *IEEE Trans. Veh. Tech.*, vol. VT-40,, pp. 584-598, August 1991.

[5] J. Dunlop, J. Irvine, D. Robertson, and P. Cosimini, "Performance of Statistically Multiplexed Access Mechanisms for a TDMA Radio Interface", *IEEE Personal Communications*, vol. 2, pp. 56-64, June 1995.

[6] H. Y. Chung, and D. J. Goodman, "Transmission of Speech and Data Using Packet Reservation Multiple Access", in *Proc. ICC '91*, Denver, June 23 - 26, 1991.

[7] D. Grillo, and G. Subiaco, "A Combined LLC/MAC Traffic Analysis of the Radio Interface in TDMA-Based Third Generation Mobile Systems", in *Proc. VTC '92*, Denver, May 11 - 13, 1992.

[8] V. Paxson, and S. Floyd, “Wide Area Traffic: The Failure of Poisson Modelling”, in *Proc. SIGCOMM '94*, London, 1994.

[9] G. Anastasi, D. Grillo, L. Lenzini, and E. Mingozzi, "A Bandwidth Reservation Protocol for Speech/data Integration in TDMA-Based Advanced Mobile Systems, in Proc. *INFOCOM '96*, San Francisco, 1996.

[10] N.M. Mitrou, TH.D. Orinos, and E.N. Protonotarios, "A Reservation Multiple Access Protocol for Microcellular Mobile-Communication Systems", *IEEE Trans. Veh. Technol.*, vol. VT-39, pp. 340-351, Nov. 1990.

[11] S. Tasaka, “Stability and Performance of the R-ALOHA Packet Broadcast System, “ *IEEE Trans. Comput.*, vol. C-32, pp. 717-726, Aug. 1983.

[12] S. Nanda, “Stability evaluation and design of the PRMA joint voice data system,” *IEEE Trans. Commun.*, vol. 42, no. 5, May 1994.

[13] W. C. Wong and D. J. Goodman, “A packet reservation multiple access protocol for integrated speech and data transmission,” *IEE Proc.-I*, vol. 139, Dec. 1992.

[14] G. Wu, K. Mukumoto, and A. Fukuda, “Analysis of an integrated voice and data transmission system using packet reservation multiple access,” *IEEE Trans. Veh. Technol.*, vol. 43, pp. 289-297, May 1994.

[15] Narasimhan and R. D. Yates, “A new protocol for the integration of voice and

data over PRMA," *IEEE Journ. Select. Areas Commun.*, vol. 14, no. 4, pp. 623-631, May 1996.

[16] F. Li and L. F. Merakos, "Voice/Data channel access integration in TDMA digital cellular networks," *IEEE Trans. Veh. Technol.*, vol. 43, no. 4, pp 986-996, Nov. 1994.

[17] P. T. Brady, "A Model for Generating ON-OFF Speech Patterns in Two-way Conversation", *Bell. Syst. Tech. J.*, pp. 2445-2472, Sept. 1969.

[18] R. Gusella, "Measurement Study of Diskless Workstation Traffic on an Ethernet", *IEEE Transactions on Communications*, vol. COM-38, pp. 1557-1568, Sept. 1990.

[19] V. Paxson, "Empirically Derived Analytic Models of Wide Area TCP Connections", *IEEE/ACM Transactions on Networking*, vol. 2, pp. 316-336, August 1994.

[20] ITU-R Draft New Recommendation FPLMTS.REVAL, "Procedure for Evaluation of Radio Transmission Technologies for FPLMTS", 8th TG 8/1 meeting, San Diego, February 7 - 17, 1995.

[21] ETSI Draft Recommendation 050402, "Radio Selection Procedures", Version 0.8.1, 12th SMG5 meeting, Rome, February 27 - March 3, 1995.

[22] D. R. Cox, and W. L. Smith, "On the Superposition of Renewal Processes", *Biometrica*, vol. 41, pp. 91-99, 1954.

Giuseppe Anastasi received the degree (Laurea) in Electronic Engineering and the Research Doctor degree in Computer Engineering from the University of Pisa, Italy, in 1990 and 1995, respectively. In 1991 he joined the Department of Information Engineering of the University of Pisa where he is currently an assistant professor. His main fields of interest include wireless and mobile networks, high speed networks, service integration, modeling and performance evaluation of computer networks. He is a member of the IEEE Computer Society.
Email: anastasi@iet.unipi.it

Davide Grillo received a degree in Statistics from the University of Rome in 1965. In the same year he joined the Fondazione Ugo Bordoni, Rome, Italy, where he is currently manager, Personal Communications. He was a Visiting Scientist at the Siemens AG Central Laboratory, Munich, and the University of Dortmund (Germany), and the IBM Zurich Research Laboratory (Switzerland). His areas of research have included: telephone network operation and control; switching exchange architecture; packet switching networks; LAN/MAN architecture, interconnection, and control; and resource allocation strategies in the radio sub-system of mobile networks. He has been on the program committees of numerous conferences and workshops, and served as chairman for the 8th ITC Seminar on Universal Personal Telecommunication in 1992. He was a Guest Editor of an issue on LAN interconnection of the IEEE JSAC, a special issue of PERFORMANCE EVALUATION dedicated to high-speed telecommunication systems, and a special issue of IEEE Personal Communications on the European research and standardization activities for advanced mobile systems. He has been an active member of SG2 of the ITU-T for questions on traffic engineering of telephone networks and ISDN, where, since 1989, he has been leading a group concerned with traffic engineering of networks supporting mobile and UPT services. He is involved in TG 8/1 (FPLMTS) of the ITU-R, where he acts as Liaison Rapporteur to ITU-T SG2. He has been involved in the RACE projects on mobile communications (RACE I 1043, and RACE II MONET and ATDMA) with responsibility in the co-ordination of traffic performance modelling activities.
He is currently on the Editorial Board of *IEEE Personal Communications* and a Guest Editor of an upcoming special issue of IEEE JSAC on Personal Communications.
He is a member of the IEEE Communications Society.
E-mail: grillo@fub.it

Luciano Lenzini holds a degree in Physics from the University of Pisa, Italy. He joined CNUCE, an institute of the Italian National Research Council (CNR) in 1970. Starting in 1973, he spent a year and a half at the IBM Scientific Centre in Cambridge, Massachusetts, working on computer networks. He has since directed several national and international projects including: RPCNET, the first Italian packet switching network; STELLA, the first European broadcasting satellite network; and OSIRIDE, the first Italian OSI network. His current research interests include integrated service networks, the design and performance evaluation of both Metropolitan Area Network MAC protocols and packet-based radio access mechanisms for third generation mobile systems. He is author or co-author of numerous publications in these fields. He is currently on the Editorial Board of *Computer Networks and ISDN Systems.* He has been on the program committees of numerous conferences and workshops, and served as chairman for the 1992 IEEE Workshop on Metropolitan Area Networks. At present he is a Full Professor at the Department of Information Engineering of the University of Pisa.
E-mail: lenzini@iet.unipi.it

Enzo Mingozzi received the Laurea Degree in Information Engineering from the University of Pisa, Italy, in 1995 where he is currently a PhD student. His research interests focus on mobile communication systems and wireless networks, design and performance analysis of computer communication protocols and traffic modeling.
E-mail: mingozzi@iet.unipi.it

extra page