ARMAP - An Energy Conserving Protocol for Wireless Multimedia Communications

Paramvir Bahl

Microsoft Research
Microsoft Corporation, Redmond, Washington, USA
bahl@microsoft.com

Abstract

Adaptive Reservation Multiple Access Protocol (ARMAP) has been designed to provide explicit support for integrated services over wireless radio networks. It allows terminals to communicate with multiple traffic types, including data, voice, and digital video, providing quality of service (QoS) guarantees to video connections and a high priority to voice connections. The regularity in the video packet generation process is exploited in the protocol to provide timely and contention free channel access for dynamic reservations. An adaptive reservation-slot scheduling algorithm ensures near-optimum bandwidth usage and near-optimum power consumption by the radio terminal. Simulation with realistic parameters reveals that ARMAP achieves a promising combination of bandwidth efficiency, and quality of service for time bounded isochronous traffic

1 Introduction

Current second-generation channel access schemes for PCS systems are biased towards integrated packet voice and data communications. Packet video has generally been ignored during the design, analysis and simulation of these systems. This is because visual communications, though desirable, was not considered essential for the success of these systems. In contrast, the objectives guiding the development of third generation PCS systems are grander. Integration of digital video with voice and data is a design goal and a required feature [5]. In-fact, the success of these systems will depend on how well they are able to support broadband services as the demand for applications with embedded multimedia components is expected to increase manifold in the near future. Web access, visual communications, and tele-training are often cited and typical examples which require efficient management and timely delivery of video data. Thus, explicit network support for streaming video is essential to the success of these emerging third-generation systems [8].

While it takes a number of system components working in harmony with each other to truly support digital video [2], our coverage in this paper is limited to the support provided by the medium access control (MAC) protocol. Although MAC protocols have been one of the favorite topics for researchers during the last two decades, and while there have been numerous proposals in this area most of these while providing respectable quality for voice and data communications fall short when evaluated for real-time video communications [1], [12], [11], [13], [15], [16]. In this paper we describe ARMAP, a mobile-to-base station channel access protocol that inherits the virtues of previously proposed voice-centric protocols and adds to these explicit support for timely transmission of compressed video. The protocol is designed to operate in a managed network only and under the control of a base station. An intelligent *slot scheduling algorithm* running in this base station orchestrates a contention-free reservation process between various on-going connections. By dynamically adapting to the time-varying system load and the characteristics of the different communicating terminals this algorithm achieves a balance between the needs of the system (efficient management of radio resources), and the needs of the applications (QoS). A distinguishing feature of ARMAP, which is key to its improved performance for visual communications, is the statistical multiplexing gain that it obtains from providing optimal number of reservation-slots in a timely manner. This is in contrast to existing protocol proposals which either do not provide such dedicated reservation slots [12], [13] or those which provide these but at uniformly distributed frame intervals without regard to the underlying traffic characteristics [11], [15], [16]. Reservation slots provided at times when they are not needed result in bandwidth wastage.

2 Protocol Operation

Adaptive Reservation Multiple Access is a TDM-based reservation multiple access protocol in which time is divided into frames and frames are divided into slots. The number of slots per frame and the size of these slots are system-wide constants that are known to all terminals. Agreement on the location of frame boundaries by communicating terminals is not important. Frame duration is equal to the voice codec packet generation period, thus exactly one voice packet is generated in one frame time for any one voice connection. Two types of slots are defined: *Reserved Slots* are slots that have been assigned to on-going connections and *Reservation Slots* are slots in which requests for new slots

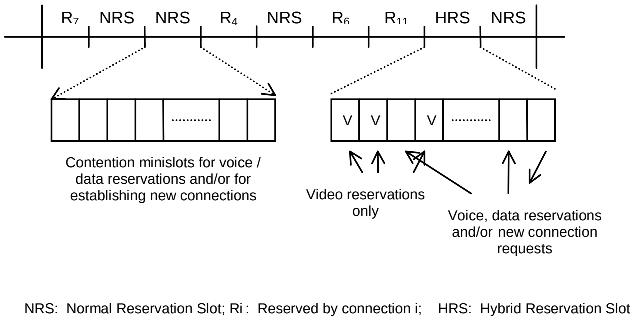

R7
NRS
NRS
R4
NRS
R6
R11
HRS
NRS
V
V
V
Contention minislots for voice / data reservations and/or for establishing new connections
Video reservations only
Voice, data reservations and/or new connection requests
NRS: Normal Reservation Slot; Ri : Reserved by connection i; HRS: Hybrid Reservation Slot

Figure 1 NRS and HRS within a ARMAP Frame

are made. Reservation Slots are further classified as *Normal Reservation Slots* (or NRS) and *Hybrid Reservation Slots* (or HRS). Both NRS and HRS are divided into minislots and reservation requests are made on minislot boundaries. Minislots within the NRS are accessed in a uncontrolled manner leading to a non-negative probability of collision between requesting terminals. Usage of these minislots is limited to terminals with on-going voice and data connections and to terminals establishing new connections. On the other hand, minislots within HRS can be used for all types of requests including future reservations for video connections. Minislots within the HRS are partitioned so that some are accessed in a controlled and contention free manner, while the rest are accessed in the usual uncontrolled manner. Figure 1 illustrates the frame and slot structure for ARMAP. The distinction between these different types of slots is made clear in the ensuing discussion.

2.1 Connection Establishment

Any terminal that wishes to establish a new connection contends for slots in one of the minislots within the Reservation Slots. Terminals employ the S-ALOHA protocol to contend on minislot boundaries. Feedback from the base station on the downlink channel at the end of the reservation slot provides information to the terminal on the success or failure of its attempt. Once the base station is aware of the terminal's connection request, it initiates a dialog with the terminal and executes the necessary steps to establish a connection. A unique connection identifier allocated by the base station and used by the terminal serves to identify the source and destination addresses for all subsequent packet transmissions to and from this terminal.

2.2 Reservations

After a connection has been established, terminals are required to reserve slots before transmission can commence. Depending on the type of connection (voice, data or video) two types of reservations are defined (1) Static Reservations and (2) Dynamic Reservations.

2.2.1 Static Reservations

Only video connections can make static reservations, and these are made at connection establishment time only. Data and voice connections can make dynamic reservations only. Static reservations are slot reservations that under normal operation are maintained for the lifetime of a connection. Reservations are canceled by explicit or implicit disconnection messages. Implicit disconnection is assumed if the terminal fails to use the reserved slots for a period of time, pre-determined by a system-wide threshold (design parameter). Explicit disconnection is made by transmitting a EOT (End-of-Transmission) sequence.

2.2.2 Dynamic Reservations

Dynamic Reservations are slot reservations that vary with each request and with the traffic class. Non real-time data connections are considered lowest priority for making such reservations. Data connections contend for slots with a permission probability that is a design variable and a system-wide constant. For such connections a success during the contention phase in a Reservation Slot, insures that a transmission slot will be available (reserved) for the connection shortly. Feedback from the base station informs the terminal whether or not it succeeded in its contention. If it did, the terminal "listens" to messages from the base station to determine which slot is available for its use. Once the data packet has been transmitted and there is more data to send the terminal returns to its contention state.

Real-time voice connections are considered highest priority. Voice connections contend for slots with a pre-determined permission probability that is higher than both the data and video permission probabilities. A success during the contention phase insures that the base station will provide the terminal with one slot per frame. Delay restrictions associated with voice packets further insure that the base station will only indicate a success if it is able to provide a transmission slot no later than one time frame. As in R-ALOHA [14], once a slot is assigned by the base station, it is reserved for the connection in subsequent frames until the terminal has no more voice packets to send. Reservations are canceled if a empty reserved slot arrives at the base station.

For real-time video connections, dynamic reservations are more complex but are made without contention. Unlike data and voice connections, video connections can make reservations only in the Hybrid Reservation Slots. Like NRS, HRS are made up of minislots but unlike NRS some of the minislots are reserved exclusively for on-going video connections. The number of minislots reserved for video connections within the HRS can range from one to all minislots and is determined by the number of on-going video connections. Collisions are avoided since each video connection is assigned a minislot in the HRS. Minislots within the HRS that are not reserved for video connections are open for contention to data, voice and new connections. At the start of each frame, the base station indicates on the downlink channel the reservation slots that are to be interpreted as HRS in the current frame. The base station also broadcasts the position of the minislot within the HRS for each video connection.

Reservations for on-going video connections are made at the end of each *video compression cycle* defined as a cycle that includes the capture, compression and packetization of a single image-frame within the video sequence. Packetization includes fragmenting the compressed image-frame into fixed size packets, adding to these appropriate header bits and error correcting codes. Video compression cycles of individual terminals are followed closely by the base station software and this information dictates the HRS generation frequency.

2.3 Scheduling Hybrid Reservation Slots

The base station utilizes knowledge of the video packet generation process to compute the periodicity of HRS. There is an inherent underlying regularity in when video packets are generated. This regularity is set by the frame capture and compression rate of the video encoder. For example, at full speed the capture and compression rate is 30 frames per second. Assuming negligible packetization delay, on the average, a video compression cycle will be 33 millisecond long and consequently, on the average, an HRS will be needed every 33 millisecond.

In practice however, due to bandwidth limitations, encoder complexity and power limitations, the actual video frame generation rate may be less than 30 frames per second. Furthermore, different terminals tend to have different capabilities. Terminals with built in hardware support for video compression usually have a smaller video compression cycle than terminals that rely on software-only solutions. Also since different terminals may have different CPUs with varying processing power, even within the group of software-only solutions, video compression cycles tend to be of different lengths. Finally, different terminals generally transmit different video streams and since the speed of video compression algorithms depends on the content of the video sequence, video compression cycles tend to be of varying length. It is thus reasonable to conclude that different terminals will need a HRS at different times.

To accommodate terminals with different video compression cycles the base station monitors the request rate of individual video connections. It then uses this information to dynamically control the frequency of the HRS and the allocation of minislots within the HRS to match the various reservation request rates. The base station thus adapts to the requirements of the communicating terminals in a manner that insures optimum usage of the radio resource. The algorithm used by the base station for determining the optimum HRS frequency and the optimum minislot allocation within the HRS is as follows:

No HRS is generated if there are no video connections and all reservation slots are treated as NRS. When this happens, ARMAP degenerates to DRMA [13]. When the first video connection is requested, the HRS generation frequency is set to the maximum possible rate (i.e. 30 HRS per second or if the system is heavily loaded to the maximum allowable by the system). The video connection is allocated one minislot within each HRS. As time progresses, a *Connection Scheduler* algorithm running in the base station monitors the number of requests the terminal made in the past second (or any other time quantum, pre-determined at design time). Depending on the request rate the HRS frequency or the reserved minislot frequency within the HRS is adjusted to match the request rate of the terminal. As new video connections are initiated within the network, the algorithm monitors the request rate of each connection and adjusts the HRS frequency and the minislot allocation frequency within each HRS to match the request rate. In the case when some request rates cannot be matched due to resource problems, the terminals slow-down their video compression cycles and adapt to the system.

Time (sec.)	HRSfreq (# of HRS / sec)	Minislots Allocated to VC1 @ 5 fps	Minislots Allocated to VC2 @ 9 fps	Minislots Allocated to VC3@ 8fps
1	30	30	x	x
2	5	5	x	x
3	6	6	x	x
4	5	5 ⟵	x	x
5	6	5	6	x
6	5	5	4	x
7	6	5	5	x
8	6	5	6	x
9	7	5	7	x
10	8	5	8	x
11	9	5	9	x
12	10	5	10	x
13	9	5	9 ⟵	x
14	9	5	9	x
15	9	5	9	9
16	9	5	9	8
17	9	5	9	9
18	9	5	9	8 ⟵

Figure 2 Minislot allocation within HRS for three video connections

Figure 2 illustrates the minislot allocation process within the HRS for three video connections originating at different times having different video compression cycles. For simplicity we have assumed the time quantum for updating *HRSfreq* and minislot allocation within the HRS to be once every second. Examination reveals that using the allocation algorithm leads to oscillations in the minislot allocation process. These oscillations if not checked result in a bandwidth wastage of $N_v/2$ minislots per second (where N_v is the number of on-going video connections). Oscillations can be detected by noticing that they begin to occur when the quantity (*MslotAllocated - MslotsUsed*) equals 1 just after it has been equal to 0 in the previous iteration. With this observation oscillations can be avoided if the following assumption is made $\mapsto$ communicating terminals maintain their capture and compression rate for the lifetime of the connection.

If all minislots within the current HRS are already reserved by on-going video connections, a existing NRS is converted to an HRS. If no NRS is available, the terminal's connection establishment request is refused.

2.4 Energy Conservation

Terminals that are capable of transmitting video at rates faster than what the channel can handle can conserve power by monitoring the number of minislots allocated to them by the base station and then adapting accordingly. For example a terminal capable of transmitting at a rate of 20 frames per second may be allocated only 10 minislots per second due to heavy traffic. By monitoring the number of reservations it was able to make, this terminal can adapt to the available radio resources and reduce its capture and compression rate to 10 frames per second. This feature has two good effects. First, power wastage is avoided. This is a big win since battery power in mobile devices is a premium resource that needs to be conserved and saved when possible. Second, bandwidth wastage is avoided. When the terminal adjusts to the allocated radio resource, it allows the compressor to adjust accordingly. Thus the compressor generates only those video frames that can be transmitted and decoded successfully. Without this adjustment the network would have to understand the compression algorithm to avoid dropping frames. This is important since in popular motion-compensated video compression algorithms, key-frames are necessary for decoding non-key frames. Without knowing what not to drop, the network may drop video frames arbitrarily and if a key frame is dropped, non-key frames already sent that rely on this dropped frame cannot be decoded. Thus, in effect valuable bandwidth is wasted since video-frames that are un-decodable were transmitted.

3 Simulation and Performance

We restrict our attention to a single cell environment with radio terminals communicating with a single base station.

3.1 Assumptions

We make the following assumptions in our simulation:

1. The number of data-only, voice-only and video-only terminals in the cell are N_d, N_a and N_v respectively. The number of ongoing data, voice and video connections at any one time is variable.
2. Each data terminal generates packets of fixed sizes with a Poisson arrival rate of λ packets per second.
3. Each voice terminal is equipped with a *voice activity detector* (silence detector). The packet rate from each on-going voice connection is constant and is modeled as in section 3.2
4. Each video terminal carries real-time connections only. The video frame generation frequency is determined randomly from a uniform distribution (between 1 and 15, bandwidth and power limitations make a 30 fps rate unrealistic in the near future). This rate is decided at connection establishment time and remains constant for the duration of the connection.
5. The permission probabilities for channel reservation for voice, video, and data are variable but the following rules are observed: real-time voice packets are given a higher priority than real-time video packets and data packets; real-time video packets are given a higher priority than data packets.
6. Bit errors occur randomly or in clusters as in Section 3.2. All control packets are free of transmission errors except for collision of reservation requests (We assume that a powerful error control technique such as a combination of RCPC and RS encoding is employed for the control packets).
7. The raw channel transmission rate is set at 2 Mbps with an effective throughput of about 1.2 Mbps.
8. The channel round-trip propagation delay is in the range of a few microseconds. Consequently, it is possible for terminals to receive instantaneous feedback from the base station.
9. The size of each information packet, time frame, slot and minislot is fixed (i.e. the number of slots in each ARMAP frame is constant and the number minislots in each ARMAP slot is constant).
10. The upper bound for the delay of voice packets and video packets is 100 millisecond. Voice and video packets are dropped if the packet is delayed beyond this value.
11. Hidden terminal and exposed terminal problems are not considered.

Connection setup and release phases of ARMAP are not considered in the simulation. Also, the overhead due to error correction is not considered.

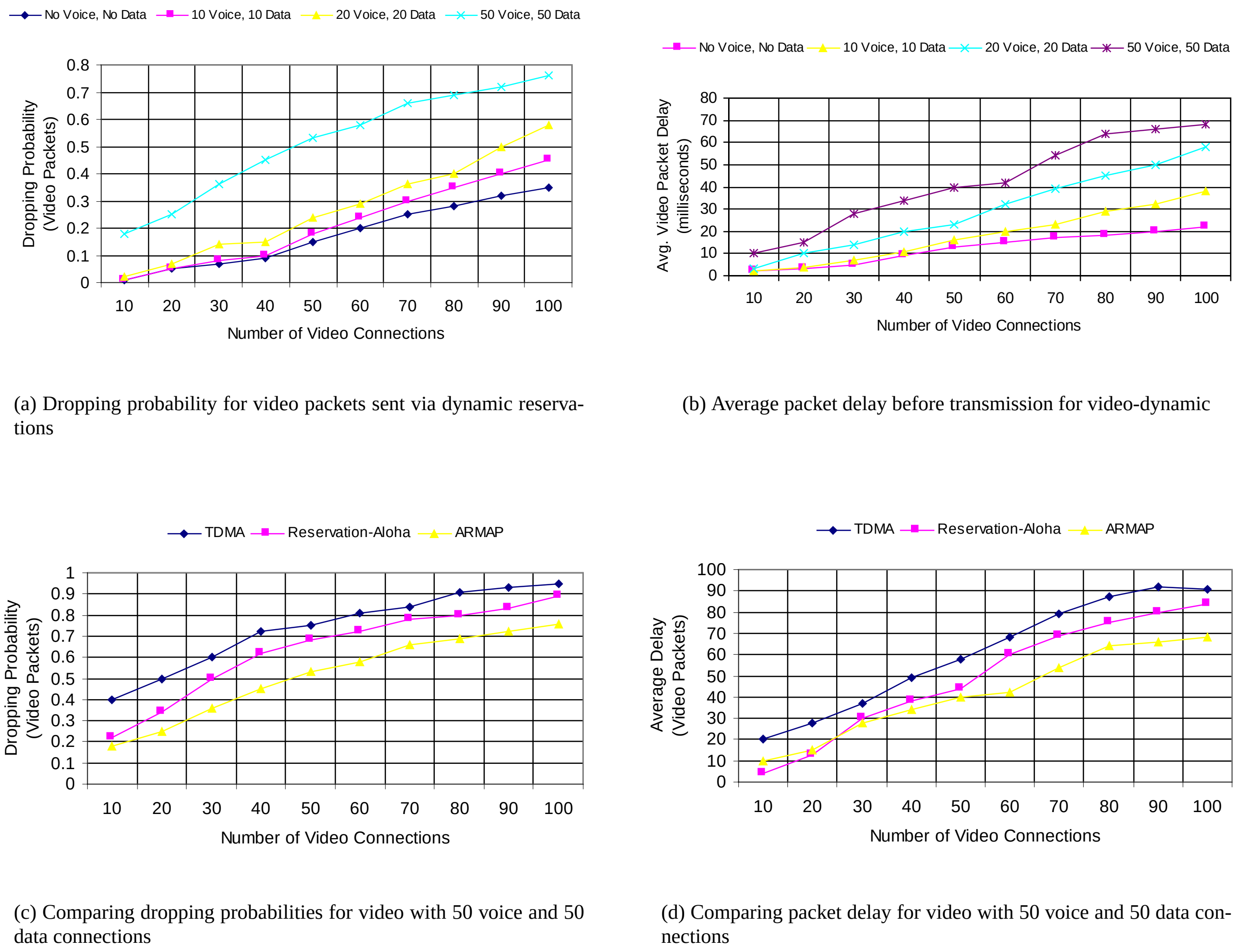
3.2 Traffic Models

1. *Voice Conversation Model*: The classical On-Off model is used [6]. The talkspurt and silent periods are independent and distributed geometrically with means of 1.5 seconds and 2.25 seconds for a voice activity factor of 40%. It is assumed that during the talkspurt the voice coder digitizes voice at a rate of 13 Kbps to form 48-byte voice packet every 3 msec.
2. *Visual Conversation Model*: For video, we use both trace driven simulation and model based simulation (see Table II). For model-based simulation we use the Markov Renewal Process (MRP) video model proposed in [10]
3. *Error Generation Model*: Errors are modeled as a modified Elliot-Gilbert model [7] and [9]. The two states represent the *burst error state (BE)* and the *random error state (RE)*. In each state bit errors are geometrically distributed with a mean bit error rate of $\lambda_{BE} = 2.1 \times 10^{-2}$ and $\lambda_{RE} = 0.5$ respectively. The transition from *RE state* to *BE state* and vice-versa are also Poisson distributed with a mean transition rate of λ_{RB}/sec and λ_{BR}/sec where $\lambda_{RB}^{-1} = 10$ sec. and $\lambda_{BR}^{-1} = 1 \cdots 3$ sec.

3.3 Results

Figure 3 shows the results obtained when ARMAP is compared to Dynamic-TDMA [16] and Reservation-ALOHA [14].

Additional performance results have been presented in [3] and [4]. In particular, in these papers we showed that when compared to classical TDMA systems, video connections with ARMAP exhibited better average video frame display rate with increasing



(a) Dropping probability for video packets sent via dynamic reservations

(b) Average packet delay before transmission for video-dynamic

(c) Comparing dropping probabilities for video with 50 voice and 50 data connections

(d) Comparing packet delay for video with 50 voice and 50 data connections

Figure 3 Simulation Results

Table I Simulation Parameters - Channel

Parameters	**Value**
Wireless MAC	ARMAP
Duplex	FDD
Minislot Size	4 bits
Slot Size	424 bits (53 bytes)
Payload Per Slot	384 bits (48 bytes)
Channel Rate	2068 Kbps
Frame Duration	2.85 msecs.
Slots per Frame	141
Bit Error Rate	see Section 3.2
Range	Approx. 40 meters
Error Correction	RS + interleave + 16-bit CRC

Table II Simulation Parameters - Video

Parameters	**Value**
Video Codec	ITU's H.263+
Packet Rate	VBR
Peak Rate	24 kbps
Frame Rate	1 to 15 Hz (varying)
Frame Size	QCIF
Permission Prob.	0.4
Packet drop threshold	100 msecs.

load, and the frames skipped over a fixed time quantum was much smaller. Additionally, the perceived quality of the video was better since the display rate with increasing bit error rate and average pear signal-to-noise ratio with increasing load was better with ARMAP.

4 Conclusions

We have introduced a TDM based reservation multiple access protocol that simultaneously supports voice, video and data in managed networks. The interesting aspect of the protocol lies in the scheduling algorithm which provides timely, contention-free access during dynamic reservations for on-going real-time video connections. Terminals are able to conserve power and bandwidth by monitoring allocated resources and avoiding unnecessary work.

Simulation results corroborate that the complexity introduced by the scheduling algorithm is offset by clear gains in bandwidth efficiency and quality of service provided by the algorithm.

References

[1] N. Amitay and L. J. Greenstein. Resource Auction Multiple Access RAMA in the cellular environment. In *IEEE International Conference on Vehicular Technology*, pages 1175–1179, 1994.

[2] P. Bahl. Supporting digital video in a managed network. *IEEE Communications Magazine*, 36(6):94–102, June 1998.

[3] P. Bahl and I. Chlamtac. H.263 based video codec for real-time visual communications over wireless radio networks. In *IEEE International Conference on Universal Personal Communications*, pages 773–779, San Diego, California, USA, October 1997.

[4] P. Bahl, I. Chlamtac, and A. Farago. Resource assignment for integrated services in wireless ATM networks. *International Journal of Communication Systems*, pages 29–41, 1998.

[5] P. W. Baier, P. Jung, and A. Klien. Taking the challenge of multiple access for third generation cellular mobile radio systems - a european view. *IEEE Communications Magazine*, 34(2):82–89, February 1996.

[6] P. Brady. A model for generating on-off speech patterns in two-way conversations. *The Bell Systems Technical Journal*, 48(2):2445–2472, 1969.

[7] E. O. Elliott. Estimates of error rate for codes on burst error channels. *The Bell Systems Technical Journal*, 42, September 1963.

[8] J. S. Dasilva et. al. European third-generation mobile systems. *IEEE Communications Magazine*, 34(10):68–83, October 1996.

[9] E. N. Gilbert. Capacity of a burst-noise channel. *The Bell Systems Technical Journal*, 39:1253–1266, September 1960.

[10] D. M. Lucantoni and M. F. Neuts. Methods for performance evaluation of vbr traffic models. *IEEE/ACM Transactions on Networking*, 2(2):176–180, April 1994.

[11] N. M. Mitrou, T. D. Orinos, and E. N. Protonotarios. A Reservation Multiple Access Protocol for microcellular mobile-communication systems. *IEEE Transactions on Vehicular Technology*, 39(4):340–351, November 1990.

[12] S. Nanda, D. J. Goodman, and U. Timor. Performance of PRMA: A packet voice protocol for cellular system. *IEEE Transactions on Vehicular Technology*, 40(3):584–598, August 1991.

[13] X. Qui and V. O. K. Li. Dynamic Reservation Multiple Access (DRMA): A new multiple access scheme for Personal Communication Systems (PCS). *ACM Journal on Wireless Networks*, 2(2):117–128, 1996.

[14] S. Tasaka. Stability and performance of R-ALOHA packet broadcast system. *IEEE Transactions on Computers*, C-32:717–726, August 1983.

[15] J. De. Vile. A reservation based multiple access scheme for future, Universal Mobile Telecommunications System. In *Proceedings of the 7th IEE Conference on Mobile and Personal Communications*, pages 210–215, December 1993.

[16] N. D. Wilson, R. Ganesh, K. Joseph, and D. Raychaudhuri. Packet CDMA versus dynamic TDMA for multiple access in integrated voice/data PCN. *IEEE Journal on Selected Areas in Communications*, 11(6):870–884, August 1993.