

Comparison of Different Scheduling Algorithms for Packetized Real-Time Traffic Flows

Kenneth S. Lee¹ and Magda El Zarki²

¹ Department of Electrical Engineering
University of Pennsylvania
Philadelphia, PA 19104
ksl@ee.upenn.edu

² Information and Computer Science
University of California, Irvine
Irvine, CA 92697
magda@ics.uci.edu

Abstract

There has been an increasing interest in IP-based wireless networks in which all traffic, including real-time traffic such as voice, is delivered in the packetized form. One main obstacle to using packet-based transmission is the time-varying characteristics of the wireless channel and the stringent delay requirement of the real-time traffic. In this paper, we propose a scheduling algorithm for real-time packet data based on the proposed physical framework of the next generation CDMA-based wireless network and compare it to some existing methods. The scheduling algorithm utilizes both the channel condition and the delay requirement of the traffic to minimize the transmit power while meeting the hard deadlines of the real-time packets. Further, we use the delay information from RTP/RTCP commonly used in IP networks to improve the performance. The scheduling algorithm is compared to the constant-SIR approach used in circuit-switched voice networks and also to a channel state-dependent (but traffic-independent) scheduling algorithm proposed for wireless packet data networks. We show that at the expense of a small increase in complexity, significant improvement can be obtained.

Key words

Wireless, IP, real-time, packet scheduling, QoS.

1. Introduction

There is an increasing interest in having a single unified network that can handle a large variety of multimedia services. The wireless network of the future should be a part of the unified network. The current packet networks have been designed for relatively delay-insensitive applications while the current cellular wireless networks have been designed primarily for voice calls. In order to have a packet-based wireless network, it is imperative that we have some means of supporting varied quality of service (QoS) requirements of different multimedia services. An especially challenging task is the support of real-time applications such as voice over packet and videoconferencing. These applications required bounded delay and guaranteed throughput but usually can

tolerate some errors (e.g. by using error concealment schemes [1]).

IP has emerged as the network platform for the unified wireline/wireless network due to its tremendous popularity and ubiquity. To promote application independence and decrease costs for transport and switching, it is highly attractive to extend IP over the air interface to the end user equipment rather than terminating the IP transport at the base station (BS). This eliminates the dependencies between applications and the wireless access network, and the opportunity for more players to participate and develop new applications expands. This should be compared with present-day cellular services, which are vertically integrated and optimized, resulting in high spectral efficiency for voice calls but low flexibility in introducing new services. In spite of some remaining issues such as efficiency of using IP over the wireless link, the benefits of having a consistent network infrastructure would be significant, and such work is on-going [2].

Providing QoS guarantees on bandwidth, end-to-end delay, delay jitter, and packet loss in a wireless network is a challenging task due to unique physical problems of the wireless channel such as location-dependent bursty errors. The allocated bandwidth also tends to be limited compared to wireline networks. There have been numerous studies that sought to address the support of multimedia services over the wireless network. However, the emphasis of these works has been on the MAC protocol [3][4] or admission control [5][6], rather than on delivery of each packet to meet the QoS requirements.

In this paper, we propose a scheduling algorithm suitable for real-time packet data streams based on the proposed physical framework of the next generation direct-sequence code division multiple access (DS-SS) wireless networks and compare it to some existing methods. The scheduling algorithm utilizes both the channel condition and the delay requirement of the traffic to minimize the resource usage while meeting the hard deadlines of the real-time packets. Further, we use the delay information from RTP/RTCP [7] commonly used in IP networks to improve the performance. Only the downlink transmission from the BS to the mobile

terminal (MT) is considered in this paper. In addition to packet scheduling, admission control plays a crucial role in providing the desired QoS. Given the performance results of the packet scheduling policy, a suitable admission control can be developed.

2. System Model

A wireless network with a cellular architecture is assumed in this paper. IP is used in the radio access network (RAN) that includes the wireless link and connected to the Internet to form an end-to-end IP network. To ensure adequate performance to the mobile user, the wireline IP network must provide for the QoS needs of real-time applications. The QoS architecture likely to be deployed on the Internet is the Differentiated Services (Diffserv) [8] model. It uses the Type of Service field of the IP header to classify packets into different aggregates. The packets are then forwarded using a particular per-hop behavior (PHB) based on the classification. Of the PHBs that have been standardized by the IETF, only expedited forwarding (EF) seems suitable for real-time applications [9]. We use this architecture to simulate the packet delivery to the BS.

The downlink is shared by different MTs using a slotted DS-CDMA protocol. The advantages of CDMA for cellular voice communication have become well known in recent years [10]. The same advantages have been further exploited to support multimedia communication over the wireless channel [11][12]. Since the capacity of CDMA systems are typically interference-limited, the total transmit power must be carefully controlled. Otherwise, the signal-to-interference ratios (SIR) of on-going connections will degrade and result in unacceptably high error rates. In theory, it is possible, at least in the downlink, to use orthogonal signature sequences or codes for different MTs to avoid interference among different codes. In practice, however, multipath effects decrease the orthogonality, and the signal intended for one user becomes interference for other users.

Transmit power control is one method of controlling interference and is an integral part of CDMA operation. It has been used in present CDMA systems to equalize the received powers from MTs in the uplink. However, it can also be used to support different BER requirements of multimedia traffic by assigning different target power levels or SIR to different traffic classes [11][12]. In order to constrain the total transmit power, data may be sent only to a subset of MTs with on-going connections at any given time (i.e. in each time slot) while other MTs maintain synchronization contact with the BS using a low-rate channel. This is in contrast to current circuit-switched connections in which data is sent to all MTs at all times albeit at different data rates. Packet data mode in some third-generation systems utilizes the approach of maintaining low-rate connections in order to decrease the acquisition delay typically associated with discontinuous transmissions [13]. This type of transmission scheme is similar to hybrid CDMA/TDMA proposed in [14] and [15].

The maximum number of packets that can be transmitted simultaneously at any given time slot depends on the SIR requirements of on-going connections. This is the function of channel conditions of different MTs due to different amounts of power required under different channel conditions.

Equation (1) shows the condition that must be satisfied for each connection i to achieve the desired level of SIR, and Equation (2) shows the limitation on total transmit power due to intercell interference requirements.

$$\left(\frac{E_b}{N_0}\right)_i \geq \frac{W}{R_i} \frac{h_i P_i}{I_{\text{intercell}} + \sum_{j \in S, j \neq i} h_j f_j P_j + \eta} \quad (1)$$

$$\sum_{i \in S} P_i < I_0 \quad (2)$$

S = set of all active MTs in the cell

W is the chip rate, R_i is the information rate, P_i is the transmit power allocated to MT i , h is the channel gain, f is the orthogonality factor, η is the background noise, and I_0 is the maximum transmit power level that has been defined a priori by the network operator in order to limit the maximum intercell interference. If condition (1) is violated for certain users, then they will not have the desired level of SIR, and if condition (2) is violated, then users in neighboring cells may not achieve the desired levels of SIR. In this paper, perfect power control is assumed, i.e. the forward channel condition is accurately determined, and the transmit power is adjusted accordingly to meet the target SIR at each MT. Different bit rate requirements and variable rate encoded streams are supported using both variable spreading gain and multi-code CDMA [16][17].

3. Packet Scheduler

Packet scheduling problem for conventional wireline networks has been studied extensively over the years (for a comprehensive survey see [18]). The studies had been done mostly for the network layer of the OSI or Internet protocol architecture. As such, the scheduling was done independently of other layers. Adapting these algorithms to the wireless domain has not been a trivial task due to unique problems of the wireless channel that had not been anticipated. In this work, we propose a scheduling algorithm that is tied closely not only to the applications layer (i.e. RTP/RTCP) but also to the physical and data link layers.

3.1 Conventional scheduling algorithms

Among the many scheduling algorithms that have been studied for the wireless channel, we describe two simple algorithms that describe the general approaches for comparison. The first one is a simple first-come-first-served (FCFS) scheduling. The scheduler (i.e. the BS) chooses the packet that has arrived before any other packets in the queue and transmits that packet with appropriately high transmit power to achieve the target SIR. In a DS-CDMA system, multiple packets may be transmitted simultaneously. In this case, the scheduler chooses the first n packets in the queue as long as the power required to transmit those packets does not violate the intracell and intercell interference requirements.

While scheduling is a packet switching-based concept, FCFS is comparable to the circuit-switched connections used in current systems. For example, in a DS-CDMA system, a persistent connection is maintained between the BS and the MT, and data is transmitted at the power required to receive

the signal with desired SIR. Since at any given time all connections must be served, admission control is used to limit the portion of time where either intracell or intercell interference requirement is violated. The difference between the circuit-switched connection and FCFS is that instead of transmitting data to all MTs at all times at low rates, data is transmitted to only a subset of MTs at higher rates in the order packets arrived at the BS. This allows the possibility of statistical multiplexing in which during periods of inactivity by one user, the bandwidth may be used by other users. Thus, FCFS should at least be equal in performance to the circuit-switched connection.

Another common approach to wireless packet scheduling is one that is channel-dependent. Since channel conditions vary widely among users and are often uncorrelated, the limited transmission resources may be better utilized by serving users with better channel conditions who require less resource, i.e. transmit power. Instead of expanding a large amount of transmit power to serve a MT with poor channel condition, the same amount of power can be used to serve more MTs or transmit more data at a higher rate. A simple channel-dependent approach is to schedule and transmit packets in order of their channel conditions (best-channel-first). The BS selects MTs with the best channel conditions in each time slot and serves the packets destined for them first. One obvious disadvantage of this scheme is that if an MT is subject to long-term fading due to its location, for example, its data may not be served for an extended period of time.

The focus of the packet scheduling algorithm proposed is on the late packet probability, i.e. the portion of packets that is not delivered to the receiver by the end of the packet's usefulness (i.e. the deadline). The data is transmitted with high-enough SIR such that random bit errors may be recovered, or application-level error concealment schemes may reduce the effects of such errors. The main objective of the proposed scheduler is to transmit each packet before its deadline, and if this is not possible, to minimize the number of packets that violate the deadline. The deadline of the packet is upper bounded by the tolerable end-to-end delay and must account for the coding/decoding delay, the propagation delay, and the queuing delay over the network that leads to the BS. We initially assume a fixed deadline appropriate for the type of traffic. However, it is possible to determine or estimate the "real" deadline of each packet.

3.2 Proposed scheduling algorithm

The proposed scheduling policy is based on the earliest-deadline-first (EDF) policy. EDF policy has been shown to be delay-optimal in wireline networks [19] and also for certain restricted cases in wireless channels [20]. Because of variable packet lengths and maximum transmission rates, the deadline of each packet is normalized by calculating a priority metric. The priority of packet i at time t is defined as follows:

$$\Phi_i(t) \equiv \frac{\lceil (L_i(t)/M_i)/\tau \rceil}{\lfloor R(t)_i/\tau \rfloor}. \quad (3)$$

L_i is the (remaining) length of packet i , M_i is the maximum transmission rate of the destined MT, R_i is the remaining time until the deadline of packet i is reached, and τ is the length of

a time slot. Longer packets may be broken into shorter segments, and Φ corresponds to the fraction of remaining slots (until deadline) in which a segment should be transmitted. Note that if Φ is greater than 1, then the packet cannot be delivered before its deadline and thus should be dropped at the BS. If Φ equals 1, then the packet must be transmitted in every slot until the entire packet is transmitted to meet the deadline. In each time slot, the scheduler calculates Φ for each packet and schedules packets in decreasing order of Φ .

The throughput of the wireless channel can be divided into two regions: intracell interference limited and intercell interference limited. A graphical depiction of the two regions is shown in Figure 1 for a single-user system. In the intracell interference limited region (the flat portion of the plot), most of the interference is caused by the cross-correlation among non-orthogonal codes and self-interference. In this region, no amount of transmit power can increase the throughput without degrading the SIR of on-going connections. In the intercell interference limited region, the maximum transmit power limit has been reached due to the need to transmit to an MT in poor channel condition and reaching the maximum intercell interference limitation. If the throughput is intercell interference limited, and if we can delay the transmission to the packet that is causing the throughput to be intercell interference limited, we may be able to transmit more data using the same amount of resource.

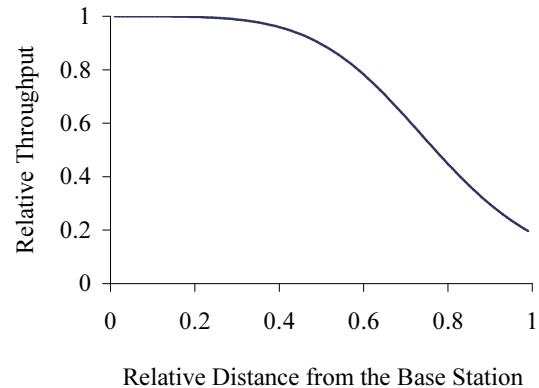


Figure 1: Throughput of the downlink as a function of distance from the base station

The proposed scheduling algorithm uses Φ_i to determine if packet i is considered urgent, i.e. is Φ_i greater than some threshold? If the packet is not considered urgent, then the scheduler delays the transmission of the packet until the next scheduling epoch at which time the channel condition of the destined MT may have improved. In the mean time, by refraining transmission to the MT with poor channel condition, more data may be transmitted, perhaps to multiple MTs (by moving to the intracell interference limited region). However, if the packet is considered urgent (i.e. has large Φ), then the packet is transmitted (with appropriately high transmit power) despite the fact that it may use disproportionately large amount of the transmission resource. The parameter that determines the urgency of the packet is denoted `priority_threshold`.

3.3 Deadline determination

The other ends of on-going connections may be as close as the same cell or as far as across the country. Hence, packets from those ends may experience significantly different amounts of delay to the BS. If we could measure such delays, then instead of treating all packets arriving at the BS equally, we could give preference to those packets have experienced more delay and are closer to the “real” deadline. We propose using RTP and RTCP for this purpose. Timestamps are used in RTP primarily for intra- and inter-media synchronization. By synchronizing the clocks at the sender and the receiver (including the BS in this case) using e.g. GPS, we can measure the one-way delay and the deadline of each real-time packet can be determined. The overhead from this operation is small if IP/UDP/RTP header compression [21] is performed. One of the problems of extending IP over the wireless link to the MT is the overhead associated with the different protocol headers. For example, IP/UDP/RTP headers are at least 40 bytes long, and a data packet using G.729 vocoder contains only 20 bytes of data. It is almost necessary to utilize some form of header compression in order to decrease the inefficiency of using a packetized transmission.

The compression is done on a link-by-link basis, and the state of each connection must be maintained at each router. The need for per-flow state and per-flow processing raises the scalability concern in large networks. On the other hand, in a wireless IP network, the BS (which also functions as a router) must keep track of each active MT (i.e. maintain state), and the header compression scheme can readily be accommodated. It is necessary for the BS to examine the packet headers in order to perform the compression over the wireless link. Thus, it is only an incremental step to extract the timestamp and maintain the deadline information for each packet.

Only the scheduling of real-time packets (specifically voice and videoconferencing) is studied in this paper. Other issues, such as fair use of unused bandwidth and admission control, will be addressed in further work. For the purpose of this study, the bandwidth allocated to the real-time traffic is assumed to be fixed (premised on the IP Premium Service).

4. Performance Evaluation

The performance of the proposed scheduling algorithm was simulated using the parameters in Table 1. The traffic source modeled a compressed voice source with voice activity detection (G.729 parameters) using a Markov-modulated stochastic process. The average bit rate produced by the vocoder was 6 kbps. The packet arrivals at the BS were derived from simulating an EF PHB using a non-preemptive priority queue. IP/UDP/RTP header compression was used over the wireless link.

Two distinct channel conditions were considered in the simulation. In the first case, only fast fading was considered. It is assumed that other techniques such as slow power control and forward error correction are used to normalize the channels such that all MTs experience identical channel statistics (Rayleigh distribution). For average-sized cells, the differences in attenuation due to different distances from the base station can be orders of magnitude larger than fading

Table 1: Simulation parameters and their values

Item	Value
Carrier frequency	2 GHz
Chip rate	1.25 Mcps
Time slot length	0.625 ms
Distance loss exponent	4.0
σ of shadow fading	6 dB
Cell radius	1000 m
Maximum transmission rate of an MT	156 kbps

effects leaving MTs further away from the base station at much disadvantage in terms of requiring more resources to receive the same service. The situation incorporating distance attenuation and shadow fading in addition to Rayleigh fading was simulated in the second part.

4.1 Fast fading only

FCFS was simulated for multiple MT speeds corresponding to different fading rates (Figure 2). The main metric we measure is late packet probability, i.e. the fraction of packets that had not been delivered to the destination before their respective deadlines. The deadline of each packet had been set to the arrival time plus 50 ms. Recall that the throughput is either intercell interference or intracell interference limited. If packet(s) at the head of the queue are to be transmitted over poor conditions, then the throughput is degraded. In faster-fading conditions, the channel conditions for the packets at the head of the queue are more likely improve in the next scheduling epoch and no longer limit the throughput.

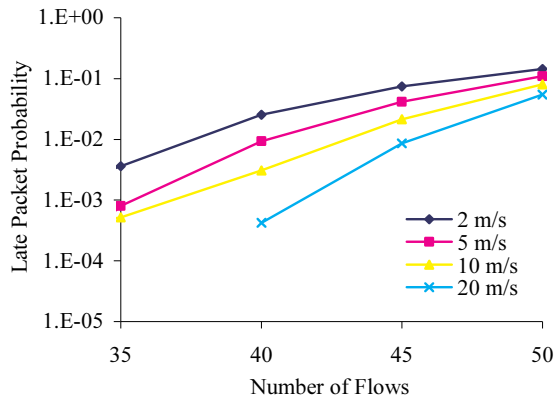


Figure 2: FCFS scheduling for different MT speeds

Figure 3 shows the performance of the proposed scheduler compared to those of the FCFS and best-channel-first algorithms described previously. FCFS performed worse than best-channel-first for most cases except for extremely low loads. Since best-channel-first scheduling is independent of arrival time or deadline, it does not take advantage of the fact that in lightly-loaded conditions, even intercell interference limited capacity may be sufficient to handle the load as long as the packets with earlier deadline are served first. The performance of the proposed scheduling algorithm varied depending on the value of

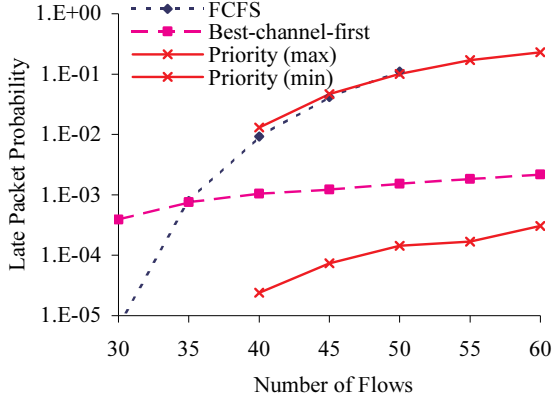


Figure 3: Performance for different scheduling algorithms (MT speed = 5 m/s)

priority_threshold (Figure 4). The maximum and the minimum values of late packet probability are shown in Figure 3. The priority_threshold value of 0 (corresponding to the “max” line in Figure 3) is equivalent to FCFS case since the same value of deadline was used for all flows. The proposed scheduling algorithm performed better in terms of timely deliver of packets for varying traffic loads. While the best-channel-first scheme obviously would have the largest throughput for delay-insensitive data, our results show that for real-time data, its performance may suffer when compared to a scheduling algorithm that is more sensitive to the delay requirements of the traffic as well as the channel condition.

The late packet probability as a function of priority_threshold is shown in Figure 4 for varying traffic loads. The priority_threshold value at which late packet probability reaches the minimum increases for higher loads. This is likely due to the fact that at higher loads, one would like to conserve power and transmit those packets requiring less resource before transmitting those requiring more. After a certain priority_threshold value, the performance is relatively flat as a typical MT goes through multiple fades during the 50 ms a packet exists at the BS queue. Thus, it is not always necessary to transmit to an MT with poor channel condition even as the priority of the packet increases.

Significantly better performance was observed if the “real” deadline of a packet was used instead of treating all voice packets as if they have the same accumulated delay to the BS. The propagation delays from the other ends of flows to the BS were varied so that the life times of packets varied from 20 to 100 ms. As mentioned previously, the task of determining the “real” deadline can be performed without significant overhead in the architecture assumed. The results of the simulation are shown in Figure 5.

4.2 Full channel model

When simulation was performed with a channel model that accounts for both propagation loss and shadow fading in addition to Rayleigh fading, performance benefits similar to those for the Rayleigh fading only channel were observed (Figure 6). However, there was no clear relationship between

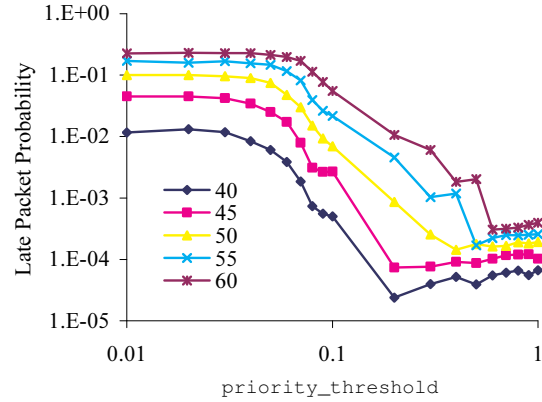


Figure 4: Performance of the proposed scheduling algorithm as a function of priority_threshold

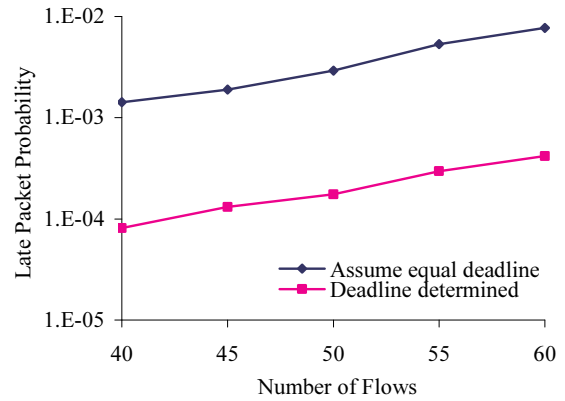


Figure 5: Performance improvement due to inclusion of deadline information

late packet probability and the value of priority_threshold. This is probably due to the fact that propagation loss and shadow fading differences between MTs can be several orders of magnitude larger than Rayleigh fading, and relative channel conditions do not change very much over the lifetime of a packet at the BS queue. Thus, priority_threshold seemed to have little effect on the performance as it had in the channel model that accounted only for fast fading. The effect does return when packets can tolerate longer delays (e.g. video streams) or the channel conditions of the MTs change quicker (e.g. move directly to/from the BS at a fast speed).

Finally, we looked at the fairness of the scheduling algorithm for the two different channel models. Because of the differences in propagation loss, it may be difficult to ensure that all users receive a similar QoS. Figure 7 shows that in the full channel model case, relatively few users experience very low late packet rates while others experience much higher rates. The differences are not as severe for the Raleigh fading case. Thus, it seems necessary that some provisions be made to compensate some users for the disadvantage of being located far away from the BS.

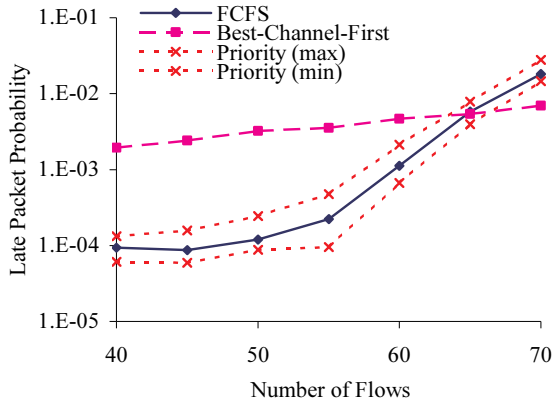


Figure 6: Performance for full channel model

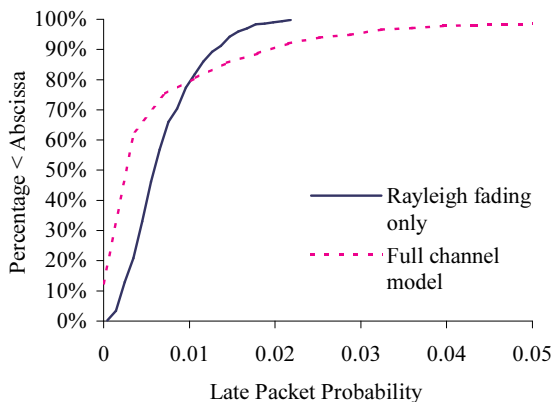


Figure 7: Fairness of two different channel models

5. Conclusion

We have presented a scheduling algorithm suitable for real-time data in packetized cellular wireless network based on DS-CDMA. The scheduling algorithm utilizes both the channel condition and the delay requirement of the traffic to minimize the transmit power while meeting the hard deadlines of the real-time packets. We compared the performance, via simulation, to two previously used approaches for wireless packet data. We also showed that utilizing delay information from RTP/RTCP can greatly improve the performance of real-time flows.

6. References

- [1] S. Shirani, F. Kossentini, and R. Ward, "Error Concealment Methods, A Comparative Study," in *Proceedings of CCECE*, Edmonton, Alberta, Canada, May 1999, pp. 835-40.
- [2] R. Ayala, K. Basu, and S. Elliott, "Internet Technology Based Infrastructure for Mobile Multimedia Services," in *Proceedings of WCNC*, New Orleans, LA, September 1999, pp. 109-13.
- [3] J. Sanchez, R. Martinez, and M. W. Marcellin, "A Survey of MAC Protocols Proposed for Wireless ATM," *IEEE Network*, vol. 11, pp. 52-62, November/December 1997.

- [4] F. Akyildiz, J. McNair, L. C. Martorell, R. Puigjaner, and Y. Yesha, "Medium Access Controls for Multimedia Traffic in Wireless Network," *IEEE Network*, vol. 13, pp. 39-47, July/August 1999.
- [5] J. S. Evans and D. Everitt, "Effective Bandwidth-Based Admission Control for Multiservice CDMA Cellular Networks," *IEEE Transactions on Vehicular Technology*, vol. 48, pp.34-46, January 1999.
- [6] W.-B. Yang and E. Geraniotis, "Admission Policies for Integrated Voice and Data Traffic in CDMA Packet Radio Networks," *IEEE Journal on Selective Areas in Communications*, vol. 12, pp. 654-64, May 1994.
- [7] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP," IETF RFC 1889, January 1996.
- [8] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An Architecture for Differentiated Service," IETF RFC 2475, December 1998.
- [9] V. Jacobson, K. Nichols, and K. Poduri, "An Expedited Forwarding PHB," IETF RFC 2598, June 1999.
- [10] K. S. Gilhousen, I. M. Jacobs, R. Padovani, A. J. Viterbi, L. A. Weaver, and C. E. Wheatley III, "On the Capacity of Cellular CDMA System," *IEEE Transaction on Vehicular Technology*, vol. 40, pp. 303-31, May 1991.
- [11] L. C. Yun and D. G. Messerschmitt, "Power Control for Variable QOS on a CDMA Channel," in *Proceedings of MILCOM*, Fort Monmouth, NJ, October 1994, pp. 178-82.
- [12] J. T. Wu and E. Geraniotis, "Power Control in Multimedia CDMA Networks," in *Proceedings of VTC*, Chicago, IL, July 1995, pp. 789-93.
- [13] T. Ojanperä and R. Prasad, "An Overview of Air interface Multiple Access for UMTS/IMT-2000," *IEEE Communication Magazine*, vol. 36, pp. 82-95, September 1998.
- [14] E. Brand and A. Hamid Aghvami, "Multidimensional PRMA with Prioritized Bayesian Broadcast – A MAC Strategy for Multiservice Traffic over UMTS," *IEEE Transactions on Vehicular Technology*, vol. 47, pp. 1148-61, November 1998.
- [15] T. Ojanperä, "FRAMES – Hybrid Multiple Access Technology," in *Proceedings of ISSSTA*, Mainz, Germany, September 1996, pp. 320-4.
- [16] C.-L. I and K. K. Sabnani, "Variable Spreading Gain CDMA with Adaptive Control for Integrated Traffic in Wireless Networks," in *Proceedings of VTC*, Chicago, IL, July 1995, pp. 794-8.
- [17] C.-L. I and R. D. Gitlin, "Multi-Code CDMA Wireless Personal Communications Networks," in *Proceedings of ICC*, Seattle, WA, June 1995, pp. 1060-4.
- [18] H. Zhang, "Service Disciplines for Guaranteed Performance Service in Packet-Switching Networks," *Proceedings of the IEEE*, vol. 83, pp. 1374-96, October 1995.
- [19] S. Panwar, D. Towsley, and J. Wolf, "Optimal Scheduling Policies for a Class of Queue with Customer Deadlines to the Beginning of Services," *Journal of ACM*, vol. 35, pp. 832-44, October 1988.
- [20] S. Shakkottai and R. Srikant, "Scheduling Real-Time Traffic with Deadlines Over a Wireless Channel," in *Proceedings of WoWMoM*, Seattle, WA, 1999, pp. 35-42.
- [21] S. Casner and V. Jacobson, "Compressing IP/UDP/RTP Headers for Low-Speed Serial Links," IETF RFC 2508, February 1999.