

Scheduling Real-Time Traffic in IP-Based Cellular Network

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

The increasing use of real-time applications over IP networks will require better use of available resources, especially in the wireless links. Priorities of different real-time packets can be differentiated by exploiting the information available from a higher layer of the protocol stack (RTP/RTCP). Utilizing the fast power control feature that is expected to be available in CDMA-based wireless networks, a packet scheduling algorithm for real-time traffic over an IP-based wireless network was designed and evaluated. As the demand for multimedia services over mobile wireless networks begins to grow, it will be important to exploit all available information, even from other layers of the protocol stack, when considering ways to deliver the quality of service that end users expect.

INTRODUCTION

The next generation cellular wireless network must be able to support a large variety of multimedia services. As the use of portable computers, personal digital assistants, and smart wireless phones increases, the demand will likely overwhelm the capabilities of current wireless networks. Unlike the voice-only cellular system characterized by homogenous traffic, the new multimedia network will have varying traffic characteristics and quality of service (QoS) requirements. Some of the traffic will be real-time (e.g. videoconferencing, voice calls, and interactive video gaming) while others will be less sensitive to delay (e.g. web browsing and email). For example, applications such as FTP do not have strict delay requirements but require virtually loss-free transmission and some minimum throughput over the lifetime of the connection. Real-time applications, on the other hand, require bounded delay and guaranteed rate but usually can tolerate some errors (e.g. using error concealment schemes [1]).

Because of the tremendous popularity and ubiquity of IP networks, there has been an ongoing effort to construct end-to-end IP networks that encompass wireless links.

Although there are some remaining issues with building wireless IP networks, including the efficiency of using IP over the wireless links, for maximum flexibility and compatibility with existing IP applications, the IP connection should extend to the mobile device [2]. The consistent network infrastructure may also simplify the development and deployment of different QoS schemes. Applications requiring better than best effort (BE) service over the Internet have been increasing over the years, and the effort to provide the necessary QoS guarantees has been an active area of research [3][4][5]. Providing similar guarantees on bandwidth, end-to-end delay, delay jitter, and packet loss in a wireless network is more challenging due to unique physical problems of the wireless channel such as limited bandwidth and high bit error rates (BER). Numerous MAC protocols have been proposed to deal with the expected demand for multimedia services over the wireless network [6][7]. However, in these approaches, the wireless link was studied separately from the rest of the network. The focus of these studies has been on the link-level performance rather than on end-to-end performance.

While the third generation mobile network developments such as WCDMA and CDMA2000 have increased the available capacity of the wireless network [8][9], the rapidly growing demand for bandwidth will continue. Furthermore, the increased burstiness of data traffic would render the performance of traditional circuit-switched approach used in the 2nd generation wireless networks inadequate. The challenge, then, is to suitably allocate the limited resource to contending sessions, depending on each session's desired QoS. This paper will present a new packet scheduling algorithm that is designed to support real-time applications in an end-to-end IP wireless network. Only the downlink is considered in this paper. The scheduling algorithm relies on information available from different layers of the protocol stack, namely the CDMA-based physical and data link layer and RTP/RTCP [10], which are widely used in transporting real-time data over the Internet. By taking advantage of information available from different layers of the protocol stack, better performance can be achieved. 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.

SYSTEM MODEL

A cellular system architecture with an end-to-end IP transport is assumed in this paper. To ensure adequate performance, the wired IP network must provide for the QoS needs of real-time applications. The QoS architecture most likely to be deployed on the Internet is the Differentiated Services (DiffServ) [5] model. It uses the Type of Service (TOS) field of the IP header for packet classification into different aggregates. The packets are then forwarded using a particular per-hop behavior (PHB) based on the classification. Of the PHB's that have been standardized by the IETF, only expedited forwarding seems suitable for real-time applications [11].

Mobile stations (MS's) in a cell share a common RF channel for downlink communication using a slotted direct-sequence code division multiple access (hybrid TDMA/CDMA) protocol. The advantages of CDMA for cellular voice communication have become well known [12]. These advantages have been exploited to support multimedia communications over the wireless channel **Error! Reference source not found.**[14]. Since the (downlink) capacity of CDMA systems are interference-limited, the total transmit power emitted by the base station (BS) must be carefully controlled. Otherwise, the signal-to-interference ratios (SIR) of on-going connections will degrade and result in unacceptably high error probability. Data will be sent only to a subset of MS's with on-going connections at any given time (i.e. in each time slot) while others maintain synchronization contact with the BS using low-rate channel. Acquisition delays typically associated with discontinuous transmission can be decrease by this approach. This transmission scheme is very similar to hybrid CDMA/TRMA proposed in [15] and [16]. Power control is an integral part of CDMA operation since it is used to control interference. The power control can also be used to support different bit error rate (BER) requirements of multimedia traffic by assigning different target power levels (i.e. SIR) to different traffic classes. To accommodate different bit rate requirements, both variable spreading gain and multi-code CDMA are used [17][18].

One problem with a wireless IP network is the overhead associated with transmitting IP and other protocol headers with each data packet. The problem is more severe for real-time applications since multiple layers of protocol are typically used to transport data packets. Our model uses IP, UDP, and RTP to ensure delivery and playback of real-time data packets (as is done e.g. in H.323 and SIP). Together, these headers are at least 40 bytes long. In comparison, a G.729 vocoder using 20

millisecond frames generates a payload of 20 bytes during speech activity. The overhead problem has been addressed by a header compression scheme [19]. It reduces IP/UDP/RTP packet headers to 2 bytes most of the time when UDP checksums are not used. However, the compression is done on a link-by-link basis, thus the state of each connection must be maintained. 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 mobile (i.e. maintain state) and header compression can readily be accommodated.

PACKET SCHEDULER

The scheduling problem for conventional wired networks has been studied extensively over the years (for a comprehensive survey see [20]). However, adapting these algorithms to the wireless domain has not been a trivial task due to unique problems in the wireless channel. In these adaptations, a two-state Markov chain is commonly used to model the wireless channel for designing and analyzing different scheduling algorithms. In such a model, the channel state for each MS is divided into "good" and "bad" states, and transmission is allowed only to those MS's with good channel states. The main premise is that channel goodput can be increased improved by swapping the channel usage between the error-prone (bad channel) and error-free (good channel) flows. One limitation of this model is that if a MS is in a fade for an extended period of time, it may not receive any data even if there are no other active MS's. The transmit power control available in CDMA systems can overcome some of the deficiencies associated with the 2-state channel model. In a DS-SS-CDMA system, all active MS's are utilizing the entire bandwidth at all times. Thus transmit power is the real shared resource that must be shared. Because more transmit power can be allocated to a MS in a bad channel state, the bad channel state can be turned into the good channel state with sufficiently low BER. However, the resource required (i.e. the transmit power) could become disproportionately large. Thus transmitting to MS's with the bad channel state could limit the total amount of information that can be transmitted in the time slot. If one were concerned only with the overall throughput of the channel, then transmission should take place only to those MS's with the good channel state. Since real-time applications require timely delivery of packets, allowing transmission to take place only when the channel condition is good may decrease the percentage of packets being delivered in time. Here lies the tradeoff in which the scheduler must decide whether it is worthwhile to transmit to a particular MS even when that decision may decrease the instantaneous channel throughput. 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 the MS.

Several criteria can be used in designing an efficient packet scheduler, including throughput, packet loss rate, delay, and delay jitter. For real-time traffic, perhaps the most important measure is the packet loss rate, i.e. the percentage of packets that are not delivered correctly to the receiver by the end of the packet's usefulness (i.e. the deadline). The 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 algorithmic delay, the propagation delay, and the queuing delay over the network that leads to the BS. In the proposed packet scheduling policy, the deadline of each packet is assumed to be known or can be estimated with reasonable accuracy. This is not an unfounded assumption. Timestamps are commonly used in transporting real-time data, 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, the deadline of each real-time packet can be determined. As mentioned before, it is necessary for the BS to examine the packet headers in order to perform IP/UDP/RTP header compression over the wireless link. Thus, it is only an incremental step to extract 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).

The proposed scheduling policy is based on the earliest-deadline-first (EDF), policy. Because different MS's scan have different maximum transmission rates, the deadline for each packet is normalized by calculating a priority metric. The priority of packet i and time t is defined in the following way:

$$\Phi_i(t) \equiv \frac{\lceil (L_i(t)/M_i)/\tau \rceil}{\lfloor R(t)_i/\tau \rfloor} \quad (1)$$

Where L_i is the (remaining) length of packet i , M_i is the maximum transmission rate of the destined MS, R_i is the time remaining until the deadline of packet i , and τ is the length of a slot. Packets may require several slots to transmit completely, and Φ corresponds to the fraction of remaining slots (until deadline) in which the packet 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 should be scheduled in every slot until

the packet is completely transmitted. For each slot, the scheduler calculates Φ and schedules packets in decreasing order of Φ . Before a packet is allowed to be scheduled for transmission, the intra- and inter-cell interference requirements must be met. The maximum number of simultaneous transmission in each slot varies as the channel conditions of different MS's vary. Equation (2) below must be satisfied for each MS to achieve the desired level of SIR. W is the chip rate, R_i is the information rate, P_i is the transmit power allocated to MS i , and h_i is the channel gain to MS i . Setting a limit on the total transmit power by the BS controls the inter-cell interference. The maximum aggregate transmit power, I_0 , is determined a priority by the network operator to satisfy the BER requirements of a certain percentages of MS's in the cell (95% in this case). If condition (3) is violated, then excessive inter-cell interference will be generated, and the performance of MS's in other cells will degrade.

$$\left(\frac{E_b}{N_0} \right)_i \geq \frac{W}{R_i} \frac{h_i P_i}{\sum_{j \in S, j \neq i} h_j P_j + \eta_0 W} \quad (2)$$

$$\sum_{i \in S} P_i < I_0 \quad (3)$$

$S = \text{set of active MS's in the cell}$

If the order of Φ values directs that transmission should take place to a MS with bad channel condition, the throughput of the channel may be reduced. Thus, the proposed scheduler 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 epoch at which time the channel condition of the MS may have improved. In the mean time, by refraining transmission to the MS with the bad channel condition, the scheduler may be able to transmit more data to multiple MS's with better channel conditions. 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 uses a disproportionately large amount of transmission resource. The parameter that determines the urgency of the packet is `priority_threshold`. If Φ is above `priority_threshold`, the packet is transmitted even though it may limit the overall throughput of the channel.

PERFORMANCE EVALUATION

The following systems parameters were chosen to evaluate the performance of the packet scheduler:

Table 1 Parameters and their values used in the simulation.

Item	Value
------	-------

Carrier frequency	2 GHz
Chip rate	1.25, 3.75 Mcps
Time slot length	0.625 ms
Distance loss exponent	4.0
Standard deviation of shadow fading	6 dB
Cell radius	1000 m
Max transmission rate for voice terminal	156 kbps
Max transmission rate for video terminal	625 kbps

Both voice and videoconferencing traffic were modeled as Markov-modulated stochastic processes. The voice traffic modeled the traffic generated from G.729 vocoder with voice activity detection. The average bit rate produced by the vocoder was 6 kbps. The videoconferencing traffic was modeled after H.263 encoder. The average bit rate of the videoconferencing source was 56.6 kbps. The arrival statistics at the BS were derived by simulating an IP Premium Service using a non-preemptive priority queue. IP/UDP/RTP header compression was used over the wireless link. The wireless channel model incorporated distance attenuation, shadow fading, and Rayleigh fading.

The performance of the scheduler was simulated under moderately heavy traffic load. FCFS scheduler performs reasonably well compared to EDF-based scheduling when real-time traffic consisting only of voice calls (Figure 1). However, as much burstier variable bit rate video traffic is introduced, the limitations of FCFS become apparent (Figure 2). In Figure 3, the late packet ratio is plotted as a function of `priority_threshold` under multiple MS speeds. If Φ of the packet is higher than `priority_threshold`, the packet is transmitted even when the overall throughput of the channel is decreased. Regardless of the MS speed, the late packet ratio is minimized around the `priority_threshold` value of ???. This may be due to the fact that the average fade duration (for all 4 speeds simulated) is shorter than the typical amount of time remaining until deadline of a packet. Thus, the fading characteristics appear statistically similar over the duration of the packet's existence at the BS. If FCFS scheduling is used, an unacceptably high late packet ratio of ??? % is observed. Simulations were performed for two different values of bandwidth (1.25 MHz and 3.75 MHz). Increased statistical multiplexing possibility introduced by the wider bandwidth (and larger number of traffic flows) did not result in improved performance when FCFS scheduling was used. Similarly considerable performance improvement (late packet ratio) was observed when the proposed scheduling algorithm was used.

Although not specifically considered in the design, another performance metric of interest is fairness of

service among active MS's. Because propagation losses can vary several orders of magnitude for MS in the same cell, a MS farther away from the base station is more likely to experience poor channel condition, i.e. more transmit power is needed to achieve the desired level of SIR. Figure 4 shows the cumulative distribution function of the late packet drop probability (MS speed = 10 m/s). It shows that with `priority_threshold` value of 0.4, 90 % of mobiles experience packet drop ratio less than 0.3% while in the FCFS case, the 90-percentile value increased to 10.9%. Thus, much fairer service (i.e. most mobiles receive acceptable level of service) is observed for the proposed scheduler.

CONCLUSION

A scheduling algorithm for real-time traffic over a wireless link was presented in this paper. The link was the last hop of an end-to-end IP network. The real-time applications were supported using RTP over UDP/IP. Fast power control of the physical/link layer was used in conjunction with deadline information available from RTP/RTCP to design a data link layer/network layer scheduling algorithm appropriate for real-time traffic. Exploiting information available from different layers of the protocol stack (link layer through transport layer) resulted in packet scheduling algorithm with much improved performance over FCFS scheduling. As the demand for multimedia services over mobile wireless IP networks continues to grow, it will be important to exploit all available information, even from other layers of the protocol stack, to meet the demand on bandwidth and other QoS.

REFERENCES

- [1] S. Shirani, F. Kossentini, and R. Ward, "Error Concealment Methods, A Comparative Study," in *Proceedings of CCECE'99*, 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'99*, New Orleans, LA, September 1999.
- [3] L. Zhang, S. Deering, D. Estrin, S. Shenker, and D. Zappala, "RSVP: A New Resource ReSerVation Protocol," *IEEE Network*, vol. 7, September 1993, pp. 8-18.
- [4] J. Wroclawski, "The Use of RSVP with IETF Integrated Services," IETF RFC 2210, September 1997.
- [5] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An Architecture for Differentiated Service," IETF RFC 2475.
- [6] J. Sanchez, R. Martinez, and M. W. Marcellin, "A Survey of MAC Protocols Proposed for Wireless ATM," *IEEE Network*, vol. 11, no. 6, November/December 1997, pp. 52-62.

- [7] 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, no. 4, July/August 1999, pp. 39-47.
- [8] T. Ojanperä and R. Prasad, "An Overview of Air interface Multiple Access for UMTS/IMT-2000," *IEEE Communication Magazine*, vol. 36, no. 9, 1998, pp. 82-95.
- [9] D. N. Knisely, S. Kumar, S. Laha, and S. Nanda, "Evolution of Wireless Data Services: IS-95 to CDMA2000," *IEEE Communication Magazine*, vol. 36, no. 10, 1998.
- [10] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP," IETF RFC 1889, January 1996.
- [11] V. Jacobson, K. Nichols, and K. Poduri, "An Expedited Forwarding PHB," IETF RFC 2598, June 1999.
- [12] 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 402, May 1991, pp. 303-31.
- [13] L. C. Yun and D. G. Messerschmitt, "Power Control for Variable QOS on a CDMA Channel," in *Proceedings of MILCOM'94*, Fort Monmouth, NJ, October 1994, pp. 178-82.
- [14] J. T. Wu and E. Geraniotis, "Power Control in Multimedia CDMA Networks," in *Proceedings of VTC'95*, Chicago, IL, July 1995, pp. 789-93.
- [15] 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, no. 4, November 1998, pp. 1148-61.
- [16] T. Ojanperä, "FRAMES – Hybrid Multiple Access Technology," in *Proceedings of ISSSTA, '96*, Mainz, Germany, September 1996, pp. 320-4.
- [17] C.-L. I and K. K. Sabnani, "Variable Spreading Gain CDMA with Adaptive Control for Integrated Traffic in Wireless Networks," in *Proceedings of VTC'95*, Chicago, IL, July 1995, pp. 794-8.
- [18] C.-L. I and R. D. Gitlin, "Multi-Code CDMA Wireless Personal Communications Networks," in *Proceedings of ICC'95*, Seattle, WA, June 1995, pp. 1060-4.
- [19] S. Casner and V. Jacobson, "Compressing IP/UDP/RTP Headers for Low-Speed Serial Links," IETF RFC 2508, February 1999.
- [20] H. Zhang, "Service Disciplines for Guaranteed Performance Service in Packet-Switching Networks," *Proceedings of the IEEE*, vol. 83, no. 10, October 1995, pp.1374-96.

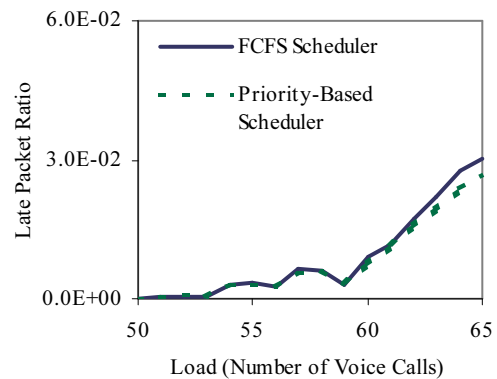


Figure 1 Late packet ratio when all real-time traffic consists of voice call. MS speed = 5 m/s, $W = 1.25$ MHz.

Figure 3

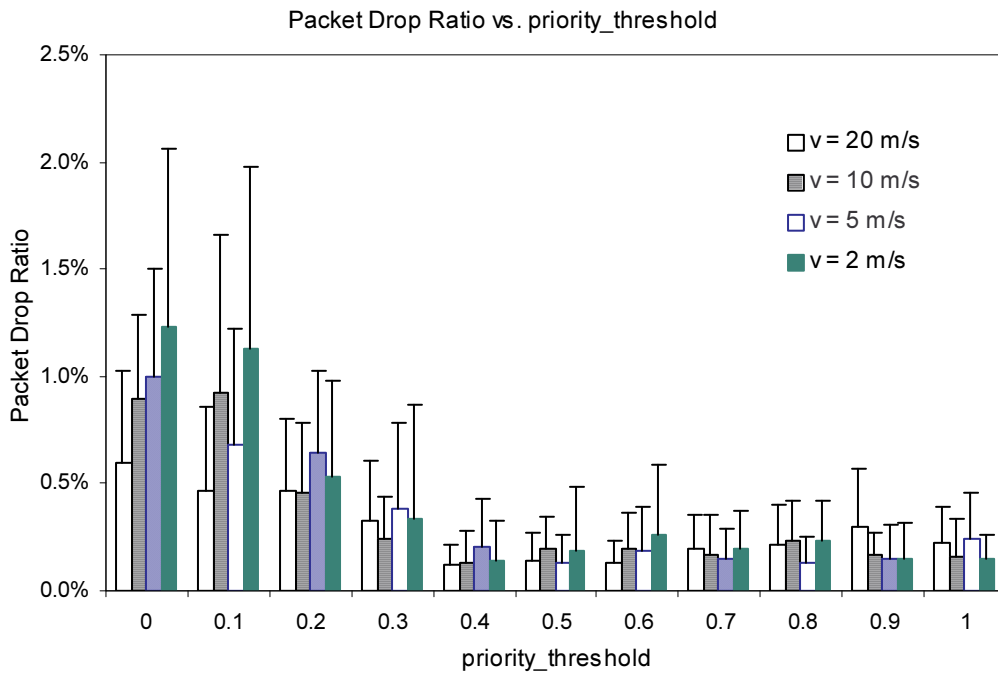


Figure 4

