

Channel-Dependent Scheduling of Real-Time Traffic in Wireless IP Networks

Kenneth S. Lee¹ and Magda El Zarki²

1 Department of Electrical Engineering
University of Pennsylvania
Philadelphia, PA 19104

2 Information and Computer Science
University of California, Irvine
Irvine, CA 92697

Abstract

This paper describes a scheduling algorithm for real-time traffic on a wireless link. The link is assumed to be the last hop of the connection and the transmission protocol used is RTP over UDP/IP. By exploiting information available from different layers of the protocol stack, it is possible to differentiate the priorities of different streams among the same class of real-time applications. Utilizing the fast power control feature that is available in CDMA-based wireless networks, a data link layer/network layer scheduling algorithm was designed and evaluated. We showed that scheduling is not necessary if only voice traffic is being carried over the link, but when video traffic is added scheduling becomes a must.

I. Introduction

Channel characteristics play a large role in determining the types and the amounts of data traffic that can be carried with sufficient quality of service (QoS) over the channel. This is especially true for the wireless channel where the bandwidth is limited and transmission errors are both location-dependent and time-dependent due to signal propagation and fading. Such dependence leads to a channel capacity (i.e. the amount of information that can be delivered without error) that varies in time and in accordance to

the spatial distribution of mobile stations (MS's). These features can have a large impact on the performance of applications that operates over the wireless channel.

Because of the present ubiquity of IP, most future networks, wired and wireless alike, are expected to be IP-centric [1]. Thus wireless devices will need to support IP and applications that run on top of IP. The poor performance of TCP over the wireless channel and possible remedies have been well studied. The emphasis of these studies has been on improving the overall throughput of TCP, and little attention was paid to delay and delay jitter of each packet. However, there is an increasing use of real-time applications over the IP network, such as voice over IP (VOIP) and video conferencing. TCP is unlikely to be used for real-time applications because the large delay that is required for retransmissions (more than the round-trip delay) is incompatible with real-time applications. In this paper we study the transport of real-time traffic over an end-to-end wireless IP network, concentrating on the scheduling algorithm in the last hop of the connection which we assume to be a wireless link (the downlink).

II. System Description

A cellular system architecture with an end-to-end IP transport is assumed in this

paper. Data packets generated at one end (wireless or otherwise) traverse the wired IP network until they reach the base station (BS) serving the wireless MS. The data then traverses over the wireless channel to the destination. While our study concerns only the final hop, the wireless link, to achieve adequate end-to-end performance, the wired IP network connected to the BS must provide the quality of service guarantee needs of real-time applications. The current Internet mostly supports only best effort service where the delivery of data packets is not guaranteed in any way. To achieve adequate performance of real-time applications, QoS must be guaranteed (either deterministically or probabilistically) over the Internet. The QoS architecture most likely to be deployed on the Internet is the Differentiated Services (DiffServ) model [2]. It uses the Type of Service field of the IP header for packet classification into different aggregates. These 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 (EF) seems suitable for real-time applications [3]. In short, EF PHB requires that the departure rate of the aggregate must equal or exceed some configurable rate.

MS's in a cell share a common RF channel for downlink communication using a slotted direct-sequence code division multiple access protocol (hybrid TDMA/ CDMA). The advantages of CDMA for cellular voice communication have become well known [5]. These advantages, including multipath resistance, etc., have been exploited to support multimedia communications over the wireless network [6][7]. Since the (downlink) capacity of CDMA systems are limited by interference, the total transmit power emitted by the BS must be carefully controlled. Otherwise, signal-to-interference ratios (SIR) of on-going connections will degrade and result in higher error probability. Thus, data will be sent only to a subset of MS's with on-going connections

at any given time (i.e. in each slot) while others maintain synchronization contact with the BS. Acquisition delays typically associated with discontinuous transmission can be decreased by this approach. This transmission scheme is similar to hybrid CDMA/TDMA proposed in [8][9]. Power control is an essential part of the CDMA operation since it is used to control interference. The power control can 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 [10][11].

One problem with an end-to-end 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 used to transport data packets, namely IP, UDP, and RTP. Our model uses IP/UDP/RTP to ensure delivery and playback of real-time data packets (as is done e.g. in H.323 and SIP). The required headers for a packet utilizing IP/UDP/RTP are 40 bytes long. In comparison, a G.729 vocoder using 20 millisecond frames generates a payload of 20 bytes during speech activity. This overhead problem has been addressed by a header compression scheme in [12]. It reduces the IP/UDP/RTP packet headers to 2 bytes most of the time when UDP checksums are not used and to 4 bytes when UDP checksums are used. However, the compression is done on a link-by-link basis and the state of each flow 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 network the BS (which also functions as a router), must keep track of each active mobile (i.e. maintain state) and header compression can easily be accommodated.

III. Packet Scheduler Design

The scheduling problem for conventional wireless networks has been studied extensively over the last few years (for a comprehensive survey, see [13]). 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 the model, the wireless channel state for each MS is divided into “good” and “bad” states, and transmission is allowed only to those MS’s that are experiencing the good channel. The main premise is that channel utilization can be improved by swapping the channel usage between error-prone (bad channel) and error-free (good channel) flows since any data that is transmitted over the bad channel is likely to be received in error. One limitation of this scheme is that performance depends largely on how the threshold between the good and the bad channel is set. The transmit power control available in CDMA systems can overcome some of the deficiencies of the 2-state channel model. Because more transmit power can be allocated to a MS experiencing poor channel condition (e.g. due to path loss and fading), the bad channel condition can be turned into a good channel with sufficiently low BER. However, the resource required (i.e. the transmit power) could become disproportionately large. Thus transmitting to MS’s experiencing bad channel conditions 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 would take place to those MS’s with the good channel condition. However, 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 though that decision may decrease the overall channel capacity. 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 metric is the packet loss rate, i.e. the percentage of packets that are not delivered to the receiver by the end of the packet’s usefulness, i.e. the deadline. Thus, 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 have not been transmitted by their respective deadlines. The deadline is an upper bounded by the tolerable end-to-end delay and must account for the coding/decoding algorithmic delay, the propagation delay and the queuing delay over the network leading up 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, will be addressed in further work. For the purpose of this study, the bandwidth available to real-time traffic is assumed to be fixed (premised on the IP Premium Service) although the goodput can vary depending on the channel state of the MS's.

The proposed scheduling policy is based on the earliest-deadline-first (EDF) policy. Because different MS's can have different transmission rates, the deadline for each packet is normalized by calculating a priority metric. The priority computation assumes that each mobile will try to transmit at the maximum transmission rate. The priority of packet i at 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 packet i (actually of the receiving 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 priority Φ corresponds to the fraction of remaining slots until the deadline by which the packet should be transmitted. Note that if Φ is greater than 1, then the packet cannot be delivered before its deadline and should be dropped at the BS. If Φ equals 1, then the packet should be scheduled for transmission in every slot until the packet is completely transmitted. For each slot, the packet scheduler schedules packets in decreasing order of Φ . Before a packet can be scheduled for transmission, the intra- and inter-cell interference requirement must be met. The maximum number of simultaneous transmissions in each slot varies as the channel conditions of different MS's vary. Equation (2) below is the condition that must be satisfied for each user to achieve the

desired level of SIR. W is the chip rate, R_i is the information rate of MS i , and P_i is the transmit power allocated for MS i . The limit on the BS's total transmit power controls the inter-cell interference in other cells. The maximum level, I_0 , is determined a priori by the network operator to satisfy a certain percentage (95 % in this case) of MS's in the cell. If condition (3) is violated, excessive inter-cell interference will be generated, and the performance of MS's in other cells will degrade.

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

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

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

If the order of Φ values directs that transmission should take place to a MS experiencing bad channel condition, the capacity 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 (i.e. has small Φ), then the scheduler delays the transmission of the packet until the next epoch at which time the channel condition for the mobile 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 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 a packet is `priority_threshold`. If the priority of the packet is above the `priority_threshold`, the packet is transmitted even though it may limit the total channel capacity.

IV. Performance Evaluation

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

Item	Value
Carrier frequency	2 GHz
Chip rate	1.25 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

Table 1 Parameters and their values used in the computer simulation.

Both voice and videoconferencing traffic were modeled as Markov-modulated stochastic processes. The voice traffic modeled the traffic generated from a G.729 vocoder with voice activity detection. The average bit rate produced by the vocoder was 6 kbps. The videoconferencing traffic was modeled after the 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 preemptive priority queue. IP/UDP/RTP header compression was used in the wireless link. The wireless channel model was subjected to distance attenuation, shadow fading, and Rayleigh fading. Rayleigh fading was simulated using Jake's model [14].

If real-time traffic consisted only of voice traffic, reasonable performance can be expected when DiffServ (EF PHB in this case) is implemented directly over the wireless link, provided that the fraction of bandwidth allocated to EF traffic is suitably limited. Since all real-time packets belong to the same aggregate, the packets are served

from the queue using the first-come-first-serve (FCFS) algorithm. Figure 1 shows that the performance of our priority-based scheduler is actually worse than the FCFS scheduler at high loads. This is because the priority-based scheduler tries to schedule all packets in the queue, and only fractions of packets end up being delivered before the deadline. The situation is different when more bursty data, such as variable bit rate video, is introduced. In this case, the FCFS scheduler severely under-performs the priority-based scheduler (Figure 2).

The performance of the scheduler was simulated under a moderately heavy traffic load. In Figure 3, late packet probabilities as a function of the priority_threshold under multiple MS speeds are plotted. If Φ_i of the packet i is higher than the value of priority_threshold, then packet i is transmitted even when the channel is in relatively bad condition. Regardless of the MS speed, the late packet probability is minimized around the priority_threshold value of 0.4. This may be due to the fact that the average fade duration (for all 4 speeds simulated) is much shorter than the typical amount of time remaining until the 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, late packet probability of around 2.8 % is observed.

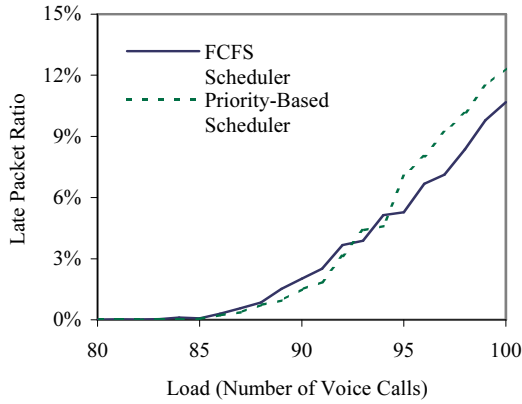


Figure 1 Late packet ratio when all real-time traffic consists of voice calls.

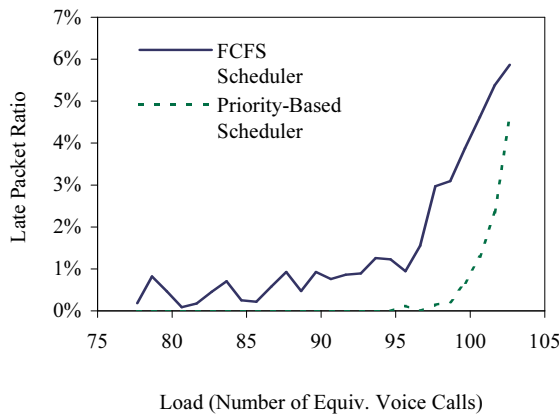


Figure 2 Late packet ratio with equal amount of voice and videoconferencing traffic.

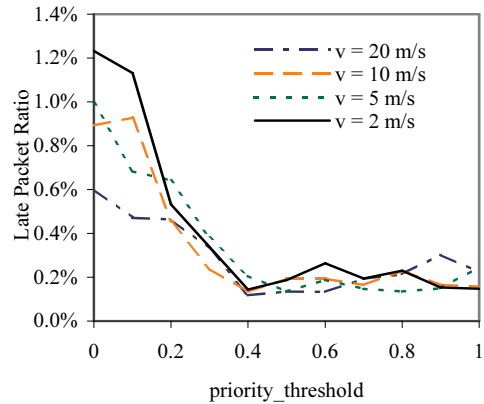


Figure 3 Late packet ratio as a function of priority_threshold and mobile station velocity.

Although not specifically considered in this work, another performance metric of interest is fairness of service among active MS's. Because propagation losses can vary several orders of magnitude for mobiles in a cell, a MS farther away from the BS is more likely to experience poor channel conditions, i.e. require more transmit power. Figure 4 (not included yet) shows the cumulative distribution function of late packet probability (at MS speed = 10 m/s). It shows that with priority_threshold value of 0.4, 90% of mobiles experience late packet probability less than 0.3% while in the FCFS case, the 90-percentile value increased to 10.9%. Thus, much fairer service (i.e. more MS's receive acceptable level of service) is observed for the proposed scheduler.

V. Conclusion

The increasing amount of real-time applications over IP networks will require better use of available resources, especially in the wireless network. By exploiting information available from different layers of the protocol stack, it is possible to differentiate the priorities of different streams among the same class of real-time applications. Utilizing the fast power control feature that is available in CDMA-based

wireless networks, a data link layer/network layer scheduling algorithm was designed and evaluated. As the demand for multimedia services over mobile wireless IP networks continues to grow, it will become important to exploit all available information, even from other layers of the protocol stack, to meet the demand on bandwidth and other QoS needs.

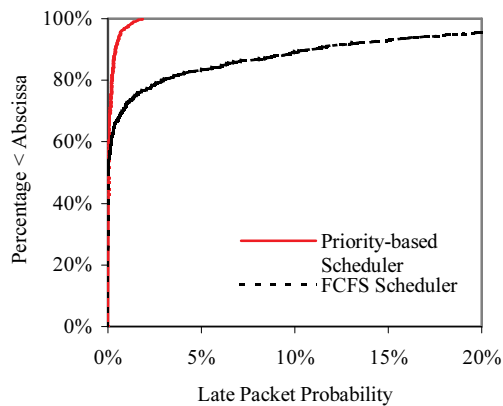


Figure 4 CDF of late packet ratio for priority-based scheduler and FCFS scheduler.

V. References

- [1] 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.
- [2] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An Architecture for Differentiated Service," IETF RFC 2475.
- [3] V. Jacobson, K. Nichols, and K. Poduri, "An Expedited Forwarding PHB," IETF RFC 2598, June 1999.
- [4] H. Naser, A. Leon-Garcia, and O. Aboul-Magd, "Voice over Differentiated Services", Internet Draft <draft-naser-voice-diffserv-eval-00.txt>, December 1998,
- [5] 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.
- [6] 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.
- [7] J. T. Wu and E. Geraniotis, "Power Control in Multimedia CDMA Networks," in *Proceedings of VTC'95*, Chicago, IL, July 1995, pp. 789-93.
- [8] 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.
- [9] T. Ojanperä, "FRAMES – Hybrid Multiple Access Technology," in *Proceedings of ISSSTA, '96*, Mainz, Germany, September 1996, pp. 320-4.
- [10] 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.
- [11] 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.
- [12] S. Casner and V. Jacobson, "Compressing IP/UDP/RTP Headers for Low-Speed Serial Links," IETF RFC 2508, February 1999.
- [13] 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.
- [14] W. C. Jakes, *Microwave Mobile Communications*. New York, NY: Wiley, 1974.