QoS Provisioning for Real-time Traffic in Wireless Packet Networks

Tianmin Ren, Iordanis Koutsopoulos and Leandros Tassiulas
Department of Electrical & Computer Engineering
and Institute for Systems Research
University of Maryland, College Park, MD 20742, USA
e-mail: \{rtm, jordan, leandros\}@isr.umd.edu

Abstract—QoS provisioning to users in the presence of volatility of the wireless channel is the most challenging issue in wireless system design. In this paper, we consider the problem of scheduling constant bit rate (CBR) traffic packets over the wireless channel, subject to packet delivery deadline constraints. We cast the problem as a Markov decision process and derive the optimal scheduling policy, in the sense of minimizing long-term packet loss due to deadline expirations. Performance bounds and design guidelines for general scheduling algorithms are obtained through analysis and simulations.

I. INTRODUCTION

Recent rapid evolutions in the area of telecommunications demonstrate that the demand for enhanced services is anticipated to grow in the near future. Advanced services, such as telecommuting or home/office networking and support of real-time traffic (live audio/video streams, video conferencing, multimedia) or data traffic (internet access, file transfer) are only the beginning of the projected demand for access to information sources of every kind. Furthermore, wireless access is necessitated by the need for ubiquitous coverage and connectivity in local loop, local area or wide area (WAN) networks, as well as the demand for mobility, flexibility and easiness of system deployment.

Given the inherent volatility of the wireless medium and the scarcity of resources, a major challenge in the design of wireless systems is quality of service (QoS) provisioning to users. Different notions of QoS are available in different communication layers. QoS in physical layer is synonymous to an acceptable signal-to-interference and noise ratio (SINR) or bit error rate (BER) at the receiver. In the MAC layer, QoS is usually expressed in terms of achievable bit rate or packet error rate (PER), while at higher layers QoS can be perceived as a minimum throughput or maximum delay requirement. The ability of the network infrastructure to fulfill QoS requirements and ultimately enhance system capacity depends on procedures that span several layers. In the physical layer transmission power, modulation level, or forward error correction coding (FEC) rate can be adapted, based on channel quality. At the MAC or network layer, QoS guarantees are provided by efficient resource management or scheduling strategies [1].

Scheduling in wire-line or wireless systems concerns allocation of the shared resource to users on a per packet basis. Scheduling is challenging in wireless systems, since the volatile link results in error bursts, during which packets cannot be reliably transmitted. Furthermore, channel errors and capacity are location-dependent, due to different fading characteristics of users, while channel quality varies randomly and asynchronously for users. Hence, the scheduling decision relies on channel states and packet flows of all users.

Two general approaches for scheduling can be identified in the literature. The first one focuses on fair resource allocation to users over a link. For wire-line networks, Weighted Fair Queuing (WFQ) was proposed in [2] as a packet-by-packet approximation to Generalized Processor Sharing (GPS) [3] for worst-case performance guarantees on throughput and delay. A modified version of WFQ for wireless links is presented in [4], where the impact of the wireless link is reflected in lagging and leading flows. A flow is said to be lagging (leading) if its queue length is greater (smaller) than the length of a virtual queue that corresponds to error-free channel. The idea is to allow lagging flows to make up their lag by causing leading flows to give up their lead.

The second approach deals with optimization of scheduling policies in a wider sense. In [5], the authors investigate the tradeoff between scheduling policies that are optimal in the sense of minimizing buffer and delay requirements. In [6], optimal scheduling without deadline constraints is studied for a wireless system with \(N\) queues and a single server, where packet arrivals and user channels are both modeled as i.i.d Bernoulli processes. It is shown that the policy that minimizes the total number of packets and delay in the system in a stochastic ordering sense is the one which serves the longest queue. For real-time traffic, each packet has a deadline, beyond which the packet is not useful to the user. The objective of a scheduling policy is to transmit maximum number of packets before their deadlines, or equivalently minimize packet loss due to deadline expirations. In [7], the authors prove that earliest deadline first (EDF) policy is optimal for wire-line networks and in [8], a modified version of EDF, the feasible earliest due date (FEDD) policy is proposed for scheduling in wireless systems with deadlines. FEDD policy schedules packets based on EDF over channels that are perceived to be in good state. The authors showed that FEDD
is optimal for symmetric systems and a class of deterministic arrival processes, but it is not optimal in general. Thus, to the best of our knowledge, the issue of optimal scheduling for real-time traffic with deadlines over wireless links has not been hitherto addressed in literature. Furthermore, the relative impact of user link qualities and packet deadline constraints on the performance of the scheduling strategy has not been precisely defined.

In this paper, we cast the scheduling problem over wireless links as a Markov decision process and identify the optimal scheduling policy, so that the average long-term packet loss due to deadline expirations is minimized. We investigate the case of constant bit rate (CBR) traffic. The main goal of our study is to quantify the performance bounds in terms of packet loss, identify the tradeoff between scheduling packets of users with better link quality and scheduling packets with smallest deadlines and motivate further research on practical scheduling algorithms that consider both channel qualities and packet deadlines.

The paper is organized as follows. In section II we provide the model and assumptions and in section III we formulate the problem and describe our solution. Numerical results are shown in section IV. Finally, section V concludes our study.

II. SYSTEM MODEL

We consider down-link transmission from the base station to $K$ users. The base station scheduler consists of $K$ infinite-length buffers, one for each user. Equal-length packets arrive at queues from higher layer applications and need to be transmitted over the wireless channel to users. Packet arrivals for each queue correspond to constant bit rate (CBR) traffic. The arrival process for queue $i$ is thus a deterministic periodic process and packets arrive every $D_i$ time units. The deadline for a packet at queue $i$ is $d_i + t_o$, where $t_o$ is the arrival epoch of the packet, which depends on $D_i$. If $t_o = 0$ for the first (head-of-line, HOL) packet of queue $i$, then the deadline of that HOL packet is $d_i$, while successive packet deadlines in queue $i$ are $d_i + D_i, d_i + 2D_i, \ldots$. The HOL packet deadlines of all queues at the beginning of slot $t$ are described by vector $d_t = (d_t^i : i = 1, \ldots, K)$. A packet deadline specifies the time until which the packet is useful for the receiver. An underlying slotted scheme is assumed. A first-in-first-out service order is applied, so that the scheduler allocates the HOL packet from a queue to each slot for transmission.

Wireless link quality is captured by channel PER and varies for each user and slot, as a result of location-dependency and time-variance of errors. PER takes values in the $L$-element set $\mathcal{P} = \{p_1, \ldots, p_L\}$. At time slot $t$, user $i$ has channel state $s_t^i = \ell$, if $p_t^i = p_\ell$. Channel conditions at slot $t$ are independent for each user and are known to the scheduler. They are described by vector $s_t = (s_t^i : i = 1, \ldots, K)$. For each user $i$, the time-varying channel condition is described by an $L$-state Markov chain, with transition probabilities $P(s_{t+1}^i = m | s_t^i = \ell) = p_{\ell m}$.

Feedback for a transmitted HOL packet is assumed to be available at the end of the corresponding time slot. If a packet is correctly received, it is removed from the queue. If the packet is not correctly received, it stays in the queue as HOL packet and can be retransmitted at a future time, provided that the deadline of the packet is not exceeded. In the event of deadline expiration, the packet is discarded from the queue and is considered to be lost.

III. SCHEDULING WITH DEADLINE CONSTRAINTS IN WIRELESS NETWORKS

A. Problem Statement

At the beginning of a time slot $t$, the scheduler knows the channel condition vector $s_t$ and HOL packet deadline vector $d_t$. In each time slot, the scheduler must transmit the HOL packet from a selected queue, so that the long-term number of lost packets due to deadline expirations is minimized. In the ideal situation of no channel errors for users, the intuitive optimal solution is to schedule the HOL packet with the smallest deadline. This EDF policy would incur packet losses due to deadline expirations only in the case where two or more HOL packets have equal deadlines, which expire simultaneously after one slot. Since one slot can accommodate only one packet, the remaining unscheduled packets with the same deadline would be discarded.

In the presence of channel errors, EDF is not optimal, since the scheduling policy should be based both on HOL packet deadlines and channel conditions. Some users experience more favorable channel conditions and are more likely to transmit packets successfully, whereas others experience poor channel conditions due to channel errors. Hence, a subset of users is more preferable for scheduling at a time slot and this subset changes with time. When good and bad channel conditions are expressed by PER=0 and 1 respectively and users are co-located, so that channel conditions are the same for users in a slot (but they differ from slot to slot), then selecting the HOL packet with the smallest deadline from a queue with PER=0 should reduce long-term packet loss due to deadline expiration. The same should hold when HOL packet deadlines for all queues are equal and the HOL packet from the queue with PER=0 is scheduled.

When HOL packet deadlines and channel conditions differ for different users, and channel conditions are not expressed with a 0-1 PER model, the problem becomes more challenging. Then, EDF does not necessarily provide reduced packet loss due to deadline expirations, since the user with the smallest HOL packet deadline may experience unfavorable channel conditions for some time, which will lead to retransmissions and possible deadline expiration. When HOL packet deadline does not dictate immediate transmission, another packet with greater deadline and better channel conditions could be transmitted. Transmission of the packet with the small deadline could be deferred until channel conditions improve. Even if the queue with the smallest HOL packet dead-
line and the least PER is selected, successful packet transmission is not guaranteed, since PER does not obtain values in \( \{0, 1\} \). Another factor that influences scheduling decisions is packet inter-arrival time \( D \) in queues. Packets that arrive in a queue with relatively small \( D \), should be granted some priority and transmitted more often, even when channel conditions are not the best, since deadlines of these packets are more likely to expire. The scheduling decision should take into consideration all aforementioned parameters.

### B. The Markov Decision Process (MDP) approach

The system state is described by a discrete-time Markov chain \( \{Y_t\}_{t=0}^{\infty} \) where \( Y_t = (d_t, s_t) \) is the system state at the beginning of slot \( t \). Thus, the state space is \( Y = Z^K \times P^K \). The scheduler is informed about HOL packet deadline and channel condition for each user at the beginning of each slot and makes the scheduling decision. Let \( u_t \in \{1, \ldots, K\} \) denote the control (decision) variable, indicating the served queue at slot \( t \) and assume that the scheduler never idles. A scheduling policy \( \pi \) is a process \( U^\pi = u_1^\pi, u_2^\pi, \ldots \), that includes the decision variables at consecutive slots. In this work, we focus on the class of stationary scheduling policies \( \Pi \), for which scheduling decisions are independent of \( t \) and depend only on \( d_t \) and \( s_t \).

The current slot is used as a reference. After each slot, the deadline of each packet in the queue is decremented by one and denotes the residual time until deadline expiration. If the HOL packet with deadline \( d_i \) from the selected queue \( i \) is received correctly, the second packet in the queue becomes HOL packet and HOL packet deadline becomes \( d_i + D_i - 1 \). If the HOL packet is not received correctly, it remains as HOL packet in the queue and its deadline is simply decremented by one. Clearly, deadlines of HOL packets of unselected queues are also decremented by one at each slot. When the deadline of a HOL packet reaches zero, the packet leaves the queue, regardless of the scheduling decision or successful transmission. It is counted as lost when it is not transmitted or when it is not received correctly. Then, the second packet in queue becomes HOL packet with deadline \( d_i \). Let \( x_t^{(i)} \) be the \( K \times 1 \) vector with \( k \)-th component equal to \( d_i \) and all other components zero and let \( 1 \) denote the \( K \times 1 \) vector of all ones. Furthermore, let \( Z_t = \{k : d_t^{(k)} = 1\} \) be the subset of queues with HOL packet deadlines equal to 1 at time \( t \) and let its cardinality be \( |Z_t| \). The state transitions \( Y_t \rightarrow Y_{t+1} \) depend on current state \( Y_t = (d_t, s_t) \) and the decision rule \( u_t \). The channel state transitions \( s_t \rightarrow s_{t+1} \) are determined by the Markov model for the channel. The HOL packet deadline transitions \( d_t \rightarrow d_{t+1} \) when \( u_t = k \) and \( k \notin Z_t \) can be succinctly given as follows:

\[
d_{t+1} = \begin{cases} 
  d_t - 1 + \sum_{i \in Z_t} x_t^{(i)} + x_t^{(k)}, & \text{w.p. } 1 - p_t^{(k)} \\
  d_t - 1 + \sum_{i \in Z_t} x_t^{(i)} + x_t^{(k)}, & \text{w.p. } p_t^{(k)},
\end{cases}
\]

(1)

where \( p_t^{(k)} \) depends on \( s_t \). When \( k \in Z_t \), we have

\[
d_{t+1} = d_t - 1 + \sum_{i \in Z_t} x_t^{(i)}, \text{ w.p. } 1 - p_t^{(k)}.
\]

The instantaneous cost \( C_t \) at time slot \( t \) is determined by the number of discarded packets due to deadline expirations and can be expressed as

\[
C_t = \begin{cases} 
  |Z_t|, & \text{if } u_t \in Z_t, \text{ w.p. } p_t^{(u_t)} \\
  |Z_t| - 1, & \text{if } u_t \in Z_t, \text{ w.p. } 1 - p_t^{(u_t)} \\
  |Z_t|, & \text{if } u_t \notin Z_t,
\end{cases}
\]

(3)

The long-term average cost per slot due to deadline expirations for policy \( \pi \in \Pi \) is,

\[
C^\pi(y) = \lim_{t \to \infty} \frac{1}{t} E_y^\pi \left[ \sum_{r=0}^{t-1} C_r \right], \text{ for } y \in Y,
\]

(4)

where \( E_y^\pi[\cdot] \) denotes expectation with respect to policy \( \pi \) on the process starting at an arbitrary state \( y \). Therefore, our problem can be rigorously stated as follows:

\[
\text{minimize } C^\pi(y) \text{ over all stationary scheduling policies } \pi \in \Pi.
\]

A policy \( \pi^* \in \Pi \) is optimal in the sense of minimizing long-term average cost, if \( C^\pi(y) \leq C^\pi^*(y) \) for any \( \pi \in \Pi \).

### C. State space reduction and solution

The major limitation in solving this MDP problem is the high dimensionality of the state space \( Y \), due to the large range of packet deadline values. The state space can be reduced by introducing an upper bound \( \bar{d} \) to HOL packet deadlines \( d_i \), so that a HOL packet can be transmitted only if \( d_i \leq \bar{d} \), for \( i = 1, \ldots, K \). After HOL packet transmission from queue \( i \), the HOL packet has deadline in range \( [1, \bar{d} + D_i - 1] \). Thus, the number of states is \( L^K \prod_{k=1}^{K} (\bar{d} + D_k - 1) \). Since the deadline of a packet denotes the time when the packet is used at the receiver, a larger deadline \( d_i \) for the HOL packet of queue \( i \) implies a larger number of \( [d_i / D_i] \) unused, queued packets at the buffer of receiver \( i \). Hence, the condition \( d_i \leq \bar{d} \) for transmission of a HOL packet could also be interpreted as a transmitter action to prevent potential buffer overflow in the receiver. The infinite-horizon MDP problem is solved by using the policy iteration algorithm [9].

### IV. SIMULATION RESULTS

We consider the scheduling problem for \( K = 2 \) queues, so as to keep complexity at a reasonable level and demonstrate our arguments. Packet inter-arrival time at queue \( i \) is \( D_i \) time units, for \( i = 1, 2 \). An upper bound \( \bar{d} \) is assumed for HOL packet deadlines. The classical 2-state Gilbert model with a good (G) and a bad (B) state and transition probabilities \( P_{bg} \) and \( P_{gb} \) is adopted for the wireless channel. The good and bad states are characterized by packet error rate \( p_L \) and \( p_H \) respectively, with \( p_H > p_L \). Unless otherwise stated,
of BCF becomes better than that of EDF when $p_b$ increases and $p_g$ decreases. This is because a higher $p_H$ does not affect user selection for BCF, but increases loss rate for EDF, since selected packets with small deadlines will most likely be lost. Lower $p_L$ results in fewer lost packets for BCF policy but does not improve loss rate in EDF so much.

Figure 4 shows performance of all methods for $p_H = 0.5$, $p_L = 0.05$, $D = (3, 5)$ and $d = 12$. Similar conclusions with those in Figure 2 can be drawn. However, for $d = 20$, PLRs for all policies are smaller than those for $d = 12$, by virtue of larger state space dimensionality. In Figure 5, we consider $p_b/g = 3$, so that the channel is in bad state for 25% of the time and we study the impact of channel state switching rate on performance. The lower PLR bound is again provided by the MDP policy. However, the relative performance of BCF and EDF policies changes for different ranges of $p_b/g$. For $p_b < 0.022$, i.e., for low channel switching rates, EDF policy yields lower PLR. A possible explanation is that deadline expirations are more likely in BCF due to longer periods when the channel is in bad state. On the other hand, BCF yields significantly lower PLR for $p_b > 0.022$. Indeed, when channel switching rate is higher, a queue is more likely to experience good channel state before its HOL packet deadline expires, so that packet will be successfully transmitted, if that queue is selected.

Significant conclusions can be drawn from these graphs. The MDP policy establishes the lower bound on PLR, since it stems from the solution to problem (5). The relative performance of practical EDF and BCF policies depends on traffic load, channel model and channel switching rate. EDF policy performs better for light traffic load and low channel switching rates, whereas BCF is better when traffic load increases and channel state changes rapidly. Furthermore, BCF performance becomes better than that of EDF as $p_H$ increases and $p_L$ decreases for $(p_H, p_L) = (1, 0)$ it almost reaches the lower bound of MDP.

V. DISCUSSION

In this paper, we addressed the problem of scheduling CBR traffic subject to deadline constraints, with the objective to reduce packet loss rate due to packet deadline expirations. The problem was studied in the context of MDP, where the main limitation is high state space dimensionality that stems from consideration of deadlines. Our primary goal is to quantify the relative impact of deadline constraints and channel conditions on scheduling policy, obtain performance bounds and draw the guidelines for the design of practical scheduling algorithms. There exist several directions for future study. For traffic patterns other than CBR, the policy that minimizes packet loss due to deadline expiration would require joint consideration of arrival processes, deadlines and channel conditions. For finite length buffers, losses due to buffer overflows need to be considered as well. Finally, devising practical scheduling policies with near-optimal performance is another issue that warrants further investigation.

REFERENCES


\[ P_{gh} = 0.01 \] and \( P_{bg} \) is a variable quantity. By computing the stationary distribution for the Markov chain, we find that the channel is in good and bad state for \( P_{bg}/(P_{bg} + P_{gh}) \) and \( P_{gh}/(P_{bg} + P_{gh}) \) of time on average. We evaluate and compare the performance of the following scheduling policies:

- **Markov Decision Process (MDP).** Results for this policy are generated by solving the MDP problem (5) with the policy iteration algorithm.
- **Earliest Deadline First (EDF).** This policy selects the queue with the smallest HOL packet deadline at each slot. If HOL packets of both queues have the same deadlines, the user with the best channel (lowest PER) is selected.
- **Best Channel First (BCF).** This policy schedules the user with the best channel (lowest PER) at each slot. If users have the same PER, the queue with the smallest HOL packet deadline is selected. The BCF policy thus resembles the FEDD policy, which is studied in [8].

The performance metric is the average long-term packet loss ratio (PLR) due to deadline expiration. Results were averaged over 1000 experiments and each experiment included measurements for \( n = 10^n \) time slots. The policy iteration algorithm for MDP converged in 5-6 iterations. For long-term average cost \( C \) as in (4), \( PLR = C/D_1/D_2/(D_1 + D_2) \), since \( n/D_1 \) packets arrive at queue \( i \) for transmission. First, we consider a system with \( p_H = 0.5 \), \( p_L = 0.05 \) and \( d = 20 \). In Figures 1 and 2, PLR is shown as a function of transition probability \( P_{bg}/g \), for inter-arrival times denoted by vectors \( D = (2, 3) \) and \( (3, 5) \) respectively. MDP approach always provides the lower bound in PLR. The BCF policy performs better than EDF for \( D = (2, 3) \), which corresponds to a scenario of small packet inter-arrival times in each queue and “dense” arrival events between the two queues. According to BCF policy, priority should be given to good channel conditions, rather than deadlines. On the other hand, EDF policy performs better for \( D = (3, 5) \), i.e., for larger inter-arrival times and sparser arrivals between queues. In that case, the scheduler can handle better HOL packet deadlines. It can be seen from Figure 2 that EDF performs gradually better than BCF as \( p_g \) increases, which implies a channel in good state for more time.

In Figure 3, PLR for \( p_H = 1 \), \( p_L = 0 \), \( D = (3, 5) \) and \( d = 20 \) is depicted. The BCF policy then almost reaches the lower bound of MDP policy. The explanation resides in the nature of BCF policy and the channel model: PLR is minimized, since users with PER=0 are always served. By comparing Figures 2 and 3, we note that BCF and EDF PLRs for \( p_H = 1 \) and \( p_L = 0 \) are higher than those for \( p_H = 0.5 \) and \( p_L = 0.05 \). However, we observe that the performance of BCF becomes better than that of EDF when \( p_H \) increases and \( p_L \) decreases. This is because a higher \( p_H \) does not affect user selection for BCF, but increases loss rate for EDF, since selected packets with small deadlines will most likely be lost. Lower \( p_L \) results in fewer lost packets for BCF policy but does not improve loss rate in EDF so much.


Fig. 1. PLR vs. $P_{bg}$ for $(p_H, p_L) = (0.5, 0.05), (D_1, D_2) = (2, 3)$ and $d = 20$.

Fig. 2. PLR vs. $P_{bg}$ for $(p_H, p_L) = (0.5, 0.05), (D_1, D_2) = (3, 5)$ and $d = 20$.

Fig. 3. PLR vs. $P_{bg}$ for $(p_H, p_L) = (1, 0), (D_1, D_2) = (3, 5)$ and $d = 20$.

Fig. 4. PLR vs. $P_{bg}$ for $(p_H, p_L) = (0.5, 0.05), (D_1, D_2) = (3, 5)$ and $d = 12$.

Fig. 5. PLR vs. $P_{bg}$ for $(p_H, p_L) = (0.5, 0.05), (D_1, D_2) = (2, 3)$ and $d = 12$. The ratio $P_{bg}/P_{gb} = 3$ is constant.