Queueing Analysis and Admission Control for Multi-Rate Wireless Networks with Opportunistic Scheduling and ARQ-Based Error Control

Long Bao Le, Ekram Hossain, and Attahiru Sule Alfa
Dept. of Elect. and Comp. Eng., University of Manitoba
Winnipeg, MB, Canada R3T 5V6
E-mail: {long, ekram, alfa}@ee.umanitoba.ca.

Abstract—
We analyze the radio link level queueing performance for a multi-rate wireless network using adaptive modulation and coding (AMC), scheduling, and automatic repeat request (ARQ)-based error control. The analytical framework, which is developed based on a vacation queueing model, can be applied to any scheduling scheme as long as the evolution of the joint service/vacation and channel processes can be determined. The exact statistics of queue length and delay are obtained. As an example of using the general analytical model, we analyze the performance of max-rate (MR) scheduling scheme which exploits multiuser diversity. Based on the queueing analysis, the impacts of channel and system parameters on the radio link level performance can be determined and hence cross-layer design and engineering can be performed. Also, efficient admission control schemes can be designed for delay-constrained applications.

Index Terms—Adaptive modulation and coding (AMC), finite state Markov channel (FSMC), quasi-birth and death (QBD) process, wireless scheduling, cross-layer design, admission control.

I. INTRODUCTION

High speed transmission of data traffic will be a key requirement for the next-generation wireless networks. Adaptive modulation and coding (AMC) technique is being used in most 2.5/3G wireless networks to increase the transmission rate by exploiting the wireless link level performance [1]. In the radio link layer, automatic repeat request (ARQ) is a useful error control technique to mitigate the residual error from the physical layer [2]-[3]. To further improve the transmission rate by exploiting the multiuser diversity, opportunistic scheduling can be employed [4].

In this paper, we analyze the radio link level queueing performance in multi-rate wireless networks with AMC, wireless scheduling, and ARQ. The results obtained in this paper enable us to quantify the impacts of physical/radio link and channel parameters on the system performance. Based on the obtained results, the cross-layer design and admission control can be carried out to optimally design the system under QoS constraints.

This work was supported by a grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

II. SYSTEM MODEL AND ASSUMPTIONS

A. System Description
Suppose that there are $L$ separate radio link level buffers which correspond to $L$ different mobile users. These buffers can be located either at the base station (BS) in case of downlink transmission or at the mobiles in case of uplink transmission. A wireless scheduler is deployed at the BS to schedule the transmissions corresponding to the different users in a time-division multiplexing (TDM) fashion. Transmissions occur within the fixed-sized time slots and during each time slot, the scheduler grants transmission for only one user.

Adaptive modulation and coding is employed at the physical layer with $K$ transmission modes corresponding to different channel states of a finite-state Markov channel. The channel is assumed to have $K+1$ states ($0, 1, \cdots, K$) as will be described in Section II.B. Now, we assume that when the channel is in state $k$ ($0, 1, 2, \cdots, K$), the transmitter transmits $c_k$ packets in one time slot ($c_k$ is an integer number). We further assume that $c_0 = 0$ (i.e., the transmitter does not transmit in channel state 0 to avoid high probability of transmission error). For convenience, let $c_K = N$.

The receiver decodes the received packets and negative acknowledgments (NACKs) are transmitted to the transmitter asking for retransmission of the erroneous packets (if any). In this paper, an error-free and instantaneous feedback channel is assumed, so that the transmitter knows exactly if there are any transmission errors at the end of the time slot. An infinite persistent ARQ protocol is assumed where the maximum number of retransmissions allowed for a packet is unbounded. The feedback channel carries both the channel state information (CSI) or the selected transmission mode and the NACKs corresponding to the ARQ protocol.

B. Channel Modeling for Adaptive Modulation and Coding
The channel model used in this paper is captured by a finite state Markov channel (FSMC) representing the multiple states of a slow Nakagami-$m$ fading channel. In this channel model, the signal-to-noise ratio (SNR) $X$ at the receiver is partitioned into finite number of intervals. Let $X_0 (= 0) < X_1 < X_2 < \cdots < X_{K+1} (= \infty)$ be the thresholds of the received SNR for different states. The channel is said to be in state $k$ if $X_k \leq X < X_{k+1}$ ($k = 0, 1, 2, \cdots, K$). As the thresholds of the received
SNR are determined, the transition matrix \( T \) for the channel states can be obtained as in [1].

In the physical layer of the considered systems, adaptive modulation (with or without coding) using QAM modulation is employed. For example, in channel state \( k \), the modulation scheme \( 2^k \)-QAM, \( k = 1, 2, \cdots, K \) is chosen. Note that, in state 0, no transmission is allowed. The average packet error rate for mode \( k \) (PER\( k \)) can be calculated as in [1].

III. FORMULATION OF QUEUEING MODEL

A. Evolution of the Joint Service/Vacation and the Channel Processes

The queueing analysis for a target queue can be performed by using a vacation queueing model. While a particular target user is served in a particular time slot, the queue is assumed to be in service; otherwise, it is said to be on vacation. We present the analysis for the case where the service/vacation process has one-step memory, the generalization to the case with any finite memory can be done easily.

Now, let \( s_n = 0 \) represent the service state and \( s_n = 1 \) represent the vacation state; \( t_n \) is the channel state in time slot \( n \). The queueing framework developed later in this paper can be applied as long as the evolution of a two-dimensional variable \( (s_n, t_n) \) can be determined. Thus, the queueing analysis for any scheduling policy of interest reduces to the determination of the conditional probability \( Pr \{ (s_{n+1}, t_{n+1}) | (s_n, t_n) \} \). To facilitate the queueing analysis, we represent this joint transition probability in the following matrix form:

\[
S = \begin{bmatrix}
0 & S_{0,1} \\
S_{1,0} & S_{1,1}
\end{bmatrix}
\]

(1)

where \( S_{i,j} \) is a \((K + 1) \times (K + 1)\) matrix whose elements \( S_{i,j}(k, l) \) are defined as follows:

\[
S_{i,j}(k, l) = \Pr \{ s_{n+1} = j, t_{n+1} = l | s_n = i, t_n = k \}. \]

(2)

B. Queueing Model and Analysis

The queueing problem is modeled in discrete time with one time interval equal to one time slot. The buffer size is assumed to be infinite. The packet arrival process is assumed to be Bernoulli with arrival probability \( \lambda \). The discrete-time Markov chain describing the system has state space \( \{ x_n, s_n, t_n \} \), \( x_n \geq 0, 0 \leq s_n \leq 1, 0 \leq t_n \leq K \}, \) where \( x_n \) is the number of packets in the target queue, \( s_n \) represents service/vacation state, and \( t_n \) is the channel state at the beginning of time slot \( n \). We assume that a packet arriving during time interval \( n - 1 \) cannot be transmitted until time interval \( n \), and the number of packets transmitted during time interval \( n \) is \( \min (x_n, c_n) \).

The resulting transition matrix describing the number of packets in the queue is written in (3) where the derivation of this matrix blocks is given in Appendix B. In this transition matrix, level \( i \) represents the fact that there are \( i - 1 \) packets in the queue. Inside each level, we capture the service/vacation state and the channel state for the target queue (chosen to be queue one in this analysis). Thus, the dimension of matrix blocks \( A_i, (i = 0, \ldots, N + 1) \) is \( 2(K + 1) \).

We re-block the transition matrix in (3) to obtain a quasi-death and death (QBD) process, which can be written as:

\[
P = \begin{bmatrix}
B & A \\
G & F \\
D_2 & D_1 & D_0
\end{bmatrix}
\]

(4)

The re-blocked transition matrix in (4) describes a QBD process, where the solution can be found by the well-established method proposed by Neuts [5]. In fact, the steady-state probability \( \pi = [\pi_0, \pi_1, \pi_2, \cdots] \) satisfies \( \pi P = \pi \) and \( \sum_{i=0}^{\infty} \pi_i = 1 \), where \( \pi_0 \) is a column vector of all ones with appropriate dimension. We can find \( \pi_0, \pi_1, \) and \( \pi_2 \) using the boundary and the normalization conditions. Other values of \( \pi_i \) (\( i > 2 \)) can be calculated from \( \pi_2 \) through a non-negative matrix \( R \) as follows:

\[
\pi_i = \pi_2 R^{i-2} \]

Now, we need to partition \( \pi_i \) to obtain the steady-state probability vector of the original transition matrix in (3). Let \( N_1 = 2(K + 1) \), \( N_2 = 2N(K + 1) \), and \( f_i = \left[ 0 \cdots I_{N_1} \cdots 0 \right]^T \) \((i = 1, 2, \ldots, N)\) be a matrix of size \( N_2 \times N_1 \), where \( I_{N_1} \) is an identity matrix in the \( i \)th position. These matrices are used to obtain the partitions of \( \pi_i \). The steady-state probability vector representing the case of having \( k \) packets in the buffer can be written as:

\[
x_0 = \pi_0, \quad x_k = \pi_i f_k, \quad k \geq 1
\]

(5)

where \((i - 1)N < k \leq iN \) and \( h = k - (i - 1)N \).

C. Delay Distribution

In this section, we derive the distribution of the total delay for an arriving packet to the target queue. The delay is the time for all packets ahead of the target packet (if any) and itself successfully leaving the queue. Let the arrival slot be numbered as slot zero and it is not included in the delay calculation.

Let us define the following matrices:

\[
\Omega(n,d) = \begin{bmatrix}
0 & \cdots & 0 \\
\cdots & \cdots & \cdots \\
0 & \cdots & 0
\end{bmatrix}
\]

(\(0 \leq i_1, i_2 \leq 1, 0 \leq j_1, j_2 \leq K\)) represent the probability that \( n \) packets are successfully transmitted in \( d \) slots, starting in (service/vacation state, channel state) corresponding to \((i_1, j_1)\) at the beginning of slot one and ending in (service/vacation state, channel state) corresponding to \((i_2, j_2)\) at the end of slot \( d \).

\[
W_l = \begin{bmatrix}
0 & \cdots & 0 \\
\cdots & \cdots & \cdots \\
0 & \cdots & 0
\end{bmatrix}
\]

(l = 0, \cdots, K) is a \((K + 1) \times 2(K + 1)\) matrix which is constructed by keeping only the \( l + 1 \)st row of \( S \) and setting all other rows to 0. This matrix represents the fact that the queue is in service and the channel is in state \( l \) at the beginning of the transmission slot.

\[
V = \begin{bmatrix}
0 & 0 \\
S_{1,0} & S_{1,1}
\end{bmatrix}
\]

This matrix represents evolution of the joint service/vacation and channel state processes given the queue is in vacation at the beginning of the transmission slot.

\[
H_{i,k,n} = \begin{bmatrix}
0 & \cdots & 0 \\
\cdots & \cdots & \cdots \\
0 & \cdots & 0
\end{bmatrix}
\]

(\(0 \leq i_1, i_2 \leq 1, 0 \leq j_1, j_2 \leq K\)) represent the probability that \( k \) packets are successfully transmitted given that there are \( n \) packets in the queue, service/vacation state changes from state \( i_1 \) to state \( i_2 \), the transmission starts in channel state \( j_1 \) and finishes in channel state \( j_2 \).
We have the following recursive relations:

\[
\Omega(n, d) = \sum_{k=0}^{N} H_{k,n} \Omega(n - k, d - 1)
\]

(6)

\[
\Omega(0, 0) = I_{2(K+1)}
\]

(7)

where \(H_{k,n}\) is calculated in terms of \(W_i\) and \(V\) in Appendix A. We can explain the above recursive relations as follows: if there are \(n\) packets which need to be transmitted in \(d\) slots, and \(k\) packets are successfully transmitted in the current slot, there remains \(n - k\) packets to be transmitted in \(d - 1\) slots. \(\Omega(0, 0)\) simply captures the ending point where the target packet leaves the queue.

Because the arrival process is memoryless, the steady-state probability vector seen by an arriving packet is \(x_i\). Thus, the probability that the delay is \(D\) slots (not including the arrival slot) can be written as follows:

\[
P_d(D) = \sum_{k=0}^{DN-1} x_k \Omega(k + 1, D) I_{2(K+1)}.
\]

(8)

The above summation is limited to \(DN - 1\) since at most \(N\) packets can be successfully transmitted in one time slot.

IV. ANALYSIS FOR MAX-RATE SCHEDULING SCHEME

The max-rate scheduling scheme works as follows. At any time slot, the channel states (one of the \(K + 1\) states of the FSMC) of all active users are assumed to be available at the scheduler without delay. We further assume that the channel processes of all users are independent. This assumption often holds in practice because of the location-dependent characteristics of the wireless channel. The max-rate scheduler grants the transmission to the user with the highest rate. If there are more than one user being in the highest channel state, the scheduler chooses one of them randomly.

Let \(s_n^{(1)}\) and \(t_n^{(1)}\) (\(t = 1, \ldots, L\)) be the service/vacation state and channel state of user \(i\) in time slot \(n\). We consider queue one, where data packets corresponding to user one are buffered. For notational simplicity, let \(s_n^{(1)} = s_n\) and \(t_n^{(1)} = t_n\). Then, we have

\[
Pr \{s_{n+1}, t_{n+1} | s_n, t_n\} = Pr \{s_{n+1} | s_n, t_{n+1}, t_n\} \times Pr \{t_{n+1} | s_n, t_n\}
\]

(9)

where \(Pr \{t_{n+1} | s_n, t_n\} = Pr \{t_{n+1} | t_n\}\) because the channel process is independent with the service/vacation process, and

\[
Pr \{t_{n+1} = | t_n = k\} = T_{kl}
\]

is the channel state transition probability, which is available from the FSMC model. We have

\[
Pr \{s_{n+1} | s_n, t_{n+1}, t_n\} = \frac{Pr \{s_{n+1}, s_n | t_{n+1}, t_n\}}{Pr \{s_n | t_{n+1}, t_n\}}
\]

(10)

where \(Pr \{s_n | t_{n+1}, t_n\} = Pr \{s_n | t_n\}\) since the service state at time \(n\) depends only on the channel state at time \(n\).

Since \(s_{n+1}\) can be either 0 or 1, we have

\[
Pr \{s_{n+1} = 1 - j | s_n = i, t_{n+1} = l, t_n = k\} = 1 - Pr \{s_{n+1} = j | s_n = i, t_{n+1} = l, t_n = k\}.
\]

(11)

Therefore, we need to consider only the case when \(s_{n+1} = 0\). Let us calculate the denominator and the numerator of (10) in the following sections.

A. Calculation of \(Pr \{s_n | t_n\}\)

We can write the denominator of (10) with \(s_n = 0\) as follows:

\[
Pr \{s_n = 0 | t_n = k\} = \frac{\sum_{t_n^{(2)}=0}^{K} \sum_{t_n^{(3)}=0}^{K} \ldots \sum_{t_n^{(L)}=0}^{K} Pr \{s_n = 0, t_n^{(2)}, \ldots, t_n^{(L)} | t_n = k\}}{\frac{1}{\alpha} \prod_{i=2}^{L} Pr \{t_n^{(i)}\}, \text{ otherwise}}
\]

(12)

where \(Pr \{s_n = 0, t_n^{(2)}, \ldots, t_n^{(L)} | t_n = k\}\) can be calculated as:

\[
Pr \{s_n = 0, t_n^{(2)}, \ldots, t_n^{(L)} | t_n = k\} = \left\{ \begin{array}{ll} 0, & \text{if } \exists i, (2 \leq i \leq L) \ s.t. t_n^{(i)} > k \\ \frac{1}{\alpha} \prod_{i=2}^{L} Pr \{t_n^{(i)}\}, & \text{otherwise} \end{array} \right.
\]

(13)

where \(\alpha\) is the number of users such that \(t_n^{(i)} = k\), \(Pr \{t_n^{(i)}\}\) is the state probability of user \(i\), which is given in the FSMC model.

Eq. (13) can be interpreted as follows. At time slot \(n\), user one is in service, given that its channel is in state \(k\) when all other \(L - 1\) users have channel states lower than or the same channel state as user one (i.e., state \(k\)). If there are \(a\) users with the same channel state \(k\), user one is chosen for transmission with probability \(1/a\). Note that, this equation holds because we assume that channel processes of all users are independent.

For \(s_n = 1\), \(Pr \{s_n = 1 | t_n = k\}\) can be written as:

\[
Pr \{s_n = 1 | t_n = k\} = 1 - Pr \{s_n = 0 | t_n = k\}.
\]

(14)
\[ Pr \{ s_{n+1} = s, s_n = v | t_{n+1} = l, t_n = k \} = \]
\[ = \sum_{i_{n+1}^{(2)}} \sum_{t_{n+1}^{(L)}} \sum_{i_{n}^{(L)}} \sum_{t_{n}^{(L)}} K \]
\[ Pr \{ s_{n+1} = s, s_n = v, t_{n+1}^{(2)}, \ldots, t_{n+1}^{(L)}, t_n^{(2)}, \ldots, t_n^{(L)} | t_{n+1} = l, t_n = k \} . \]

\[ (15) \]

B. Calculation of \( Pr \{ s_{n+1}, s_n | t_{n+1}, t_n \} \)

The numerator of (10) can be written as in (15). The cases \( s_{n+1} = 0, s_n = 0 \) and \( s_{n+1} = 0, s_n = 1 \) are calculated as in (16) and (17), respectively.

Eq. (16) can be interpreted as follows. User one is in service in two consecutive time slots \( n \) and \( n+1 \) given that the channel state for user one is \( k \) and \( l \) in slots \( n \) and \( n+1 \), respectively, when the channel states of all other \( L-1 \) users are smaller than or equal to \( k \) and \( l \) in slots \( n \) and \( n+1 \), respectively. If there are \( a \) and \( b \) users including the target user which have channel states satisfying \( i_n^{(i)} = k \) and \( t_n^{(i)} = l \), respectively, user one is granted transmission in both slots with probability \( 1/ab \).

The interpretation of (17) is similar to that of (16) but we have more cases.

V. Numerical and Simulation Results

In this section, we present typical numerical results considering an uncoded wireless system with five transmission modes. The fitting parameters of packet error rate curves are from [1]. We assume that \( c_k = k \), time slot interval \( T_s = 0.5 \) ms and the SNR thresholds of the FSMC model are found such that the average packet error rate \( P_{BER_k} = P_0, (k = 1, \ldots, 5) \) as in [1]. To save the simulation time in validating the analytical results, only two transmission modes are used, i.e., \( K = 2 \) (in Fig. 1, 2).

\[ \text{Fig. 2. Delay distribution (for } \lambda = 0.1, L = 3, \text{average SNR} = 15 \text{ dB, } P_0 = 0.1, m = 1, \text{and } f_d = 40 \text{ Hz).} \]

The complete delay statistics obtained from the analytical model above enables us to perform admission control under statistical delay constraint of the form: \( Pr(\text{delay} > D_{\text{max}}) = 1 - \sum_{k=1}^{D_{\text{max}}} P_d(k) \leq P_t \). For a certain setting of channel and system parameters, the admission control parameter \( \gamma \), which can be, for example, the admissible number of users or the packet arrival rate, can be found by a simple search such that the delay constraint is satisfied. As an example, in Fig. 3, we show typical variations in the maximum arrival rate with average SNR, when \( D_{\text{max}} = 70, P_t = 0.05, P_0 = 0.1, \) number of users \( L = 3 \), Doppler shift \( f_d = 60, 70, 80 \text{ Hz}, \) and Nakagami parameter \( m = 1 \). As is evident, higher Doppler shift and/or larger average SNR reduce delay thus allowing traffic to be admitted into the queue at a higher rate.

VI. Conclusions

A novel analytical framework for queueing performance evaluation of multi-rate wireless networks with ARQ and scheduling was presented in this paper. The exact queue length and delay distributions were derived analytically, which enables us to determine the impacts of system and wireless channel parameters on the system performance and optimally design the system.
under the statistical delay constraints. The analysis was validated by comparing it with simulation for the max-rate scheduling scheme. An example of packet-level admission control was also illustrated to highlight the usefulness the analytical model.

APPENDIX A
Derivation of $H_{k,n}$

Let $\theta_k$ be the probability of transmission error when the channel is in state $k$. Assume that the transmission errors of different packets are independent, the probability that $i$ packets are correctly received given that $j$ packets were transmitted when the channel is in state $k$ can be written as follows:

$$
 p_{ij}^{(k)} = \binom{j}{i} \theta_k^{j-i} (1-\theta_k)^i. 
$$

For $n < N$, $H_{k,n}$ can be calculated as follows:

$$
 H_{k,n} = \sum_{i=1}^{l-1} p_{i,j}^{(k)} W_i + \sum_{i=1}^{K} p_{i,k}^{(k)} W_i, \quad k > 0
$$

$$
 H_{0,n} = V + W_0 + \sum_{i=1}^{l-1} p_{0,i}^{(k)} W_i + \sum_{i=1}^{K} p_{0,i}^{(k)} W_i
$$

where $l$ satisfies $c_{l-1} \leq n < c_l$. In (19) and (20), the number of packets in the queue (equal to $n$) is less than the maximum number of packets that can be transmitted (equal to $N$ in channel state $K$). In (19), the first sum is for the case when the number of available packets in the queue is greater than the number of packets that can be transmitted and the second sum captures the case when the number of available packets in the queue is smaller than the number of packets that can be transmitted. In (20), the first term represents the vacation state and the second term is for channel state zero where no transmission is allowed, and the first and the second sums in (20) represent the same cases as in (19). For $n \geq N$,

$$
 H_{k,n} = \sum_{i=1}^{K} p_{k,ci}^{(k)} W_i, \quad k > 0
$$

$$
 H_{0,n} = V + W_0 + \sum_{i=1}^{K} p_{0,i}^{(k)} W_i
$$

where $H_{k,n} = 0$ if $k < 0$ or $k > n$, and $p_{i,j}^{(k)} = 0$ if $i > j$ or $i < 0$. The interpretation for (21) and (22) are similar to those for (19) and (20).

APPENDIX B
Derivation of Matrix Blocks in (3)

The transition matrix blocks in (3) can be written as follows:

$$
 B = (1-\lambda)S, \quad A_{0,0} = \lambda S, \quad G_n = (1-\lambda)H_{k,n}, 1 \leq n \leq N, \quad (23)
$$

The derivation of these three matrix blocks is quite straightforward. Note that, the matrix $G_n$ represents the fact that the number of packets decreases from $n (> 0)$ to 0, and therefore, all available packets must leave the queue and there is no new arrival. $A_{n,k}$ and $A_k$ can be written as follows:

$$
 A_{n,k} = (1-\lambda)H_{k-1,n} + \lambda H_{k,n}
$$

where $1 \leq n < N$, and $0 \leq k \leq n$ and

$$
 A_k = (1-\lambda)H_{k-1,N} + \lambda H_{k,N}
$$

where $0 \leq k \leq N + 1$.

REFERENCES


