Tandem Queue Models with Applications to QoS Routing in Multihop Wireless Networks

Long Le, *Member*, *IEEE*, and Ekram Hossain, *Senior Member*, *IEEE*

Abstract—We consider the problem of quality-of-service (QoS) routing in multihop wireless networks, where data is transmitted from a source node to a destination node via multiple hops. The key component of any QoS routing algorithm is the route discovery task, where a good route with sufficient radio resources needs to be found, and resource reservation needs to be performed in such a way that the end-to-end QoS requirements are satisfied. The route discovery essentially involves the link and path metric calculation, which depends on many factors such as the physical and link layer designs of the underlying wireless network and transmission errors due to channel fading and interference. The task of link metric calculation basically requires us to solve a tandem queuing problem, which is the focus of this paper. We present a unified tandem queue framework, which is applicable for many different physical layer designs. It considers the multirate transmission feature in the physical layer and automatic repeat request (ARQ) protocol for error recovery in the link layer. We present both exact and approximated decomposition approaches. Using the queuing framework, we can derive different performance measures, namely, end-to-end loss rate, end-to-end average delay, and end-to-end delay distribution. The proposed decomposition approach is validated, and some interesting insights into the system performance are highlighted. We then present how we can use the decomposition queuing approach to calculate the link metric and incorporate this into the route discovery process of the QoS routing algorithm. The numerical results for the proposed QoS routing algorithm using the queuing framework are presented, and the impacts of different system and QoS parameters on network performance are investigated. The extension of the queuing and QoS routing framework to wireless networks with class-based queuing for QoS differentiation is also presented.

Index Terms—Tandem queue, end-to-end quality of service (QoS), QoS routing, cross-layer design, multirate transmission, automatic repeat request (ARQ), multihop wireless networks.

1 Introduction

Multihop wireless networks are emerging as important components for future generation wireless systems [1], [2], [3], [4]. Although the problem of cross-layer performance modeling and analysis of single-hop wireless networks such as conventional cellular networks has been addressed in the literature [5], [6], [7], the analysis and optimization of multihop wireless networks with quality-of-service (QoS) constraints is an open research problem.

In multihop wireless networks, routing is one of the key research problems. The QoS constraints of many wireless applications make the routing problem even more challenging [8], [9]. One of the most important components of any QoS routing algorithm is the route discovery task, where link/path metric calculation and resource reservation should be performed such that the required QoS requirements are satisfied. Typical end-to-end QoS metrics such as delay and loss rate depend on the queuing dynamics at each node along the route, which again depends on factors such as traffic arrival pattern, physical and link layer designs, and transmission errors on the wireless channel. A complete tandem queuing model, considering the traffic arrival process along with the wireless channel and the physical/link layer parameters, would enable us to design and analyze the performance of QoS routing algorithms in a multihop wireless network.

For QoS routing, the performance metrics should be measured or calculated in a timely manner so that the routing algorithm can adapt to the system dynamics. In fact, QoS routing has been an active research topic over the last several years. Typical routing metrics adopted in the literature are bandwidth and delay [8], [10], [12]. Although bandwidth can usually be quantified, existing work in the literature usually assumes that routing algorithms have the capability of estimating the link delay. In addition, the end-to-end loss rate is usually ignored.

In this paper, we present an exact queuing model to analyze a tandem of queues, where batch traffic arrival process, multirate transmission in the physical layer, and automatic repeat request (ARQ)-based error recovery in the link layer are taken into account. This multirate transmission leads to a queuing model with state-dependent service rate. We assume per-flow queuing along the routing path, for which a separate queue is maintained for each flow. Since the computational complexity of the exact model is very high, we propose a decomposition approach for the tandem queue. Using the decomposition approach, we can calculate the performance measures such as end-to-end loss rate, average delay, and delay distribution with much lower computational complexity. We then show how the

- *E. Hossain is with the Department of Electrical and Computer Engineering, University of Manitoba, 75A Chancellor's Circle, Winnipeg, MB R3T 5V6 Canada. E-mail: ekram@ee.umanitoba.ca.*
- *L.B. Le is with the Department of Electrical and Computer Engineering, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada N2L 3G1. E-mail: longble@engmail.uwaterloo.ca.*

Manuscript received 5 Nov. 2006; revised 4 July 2007; accepted 17 Oct. 2007; published online 24 Oct. 2007.
For information on obtaining reprints of this article, please send e-mail to: tmc@computer.org, and reference IEEECS Log Number TMC-0304-1106.
Digital Object Identifier no. 10.1109/TMC.2007.70777.

decomposition approach can be incorporated into a QoS routing protocol so that the end-to-end QoS requirements are satisfied. Then, we extend the per-flow queuing-based QoS routing framework to a class-based queuing and QoS routing framework, which supports a finite number of service classes with differentiated QoS requirements.

The rest of this paper is organized as follows: Section 2 presents the background and related work in the literature. Section 3 describes the system model. An exact tandem queuing model is described in Section 4. Section 5 presents the decomposition approach for the tandem queue. The application of the proposed decomposition approach for QoS routing is presented in Section 6. Numerical results are presented in Section 7. Extension of the per-flow queuing-based QoS routing framework to class-based queuing implementation is presented in Section 8. Some further discussions and extensions are presented in Section 9. Section 10 states the conclusions.

2 Background and Related Work

Routing in wireless ad hoc networks has been an active research topic for the last several years [9]. Most of the routing algorithms in the literature find the routes for incoming connections based on the minimum hop count without any QoS guarantee. QoS routing, on the other hand, is an important subclass of routing algorithms where some specific end-to-end QoS requirements must be satisfied [8], [10], [11], [12]. Routing algorithms can also be classified as being of either single-path [16], [17], [18], [19], [20], [21], [22], [23] or multipath type [24], [25], [26]. Although multipath routing usually offers better load balancing, it incurs more overhead. In addition, compared to single-path routing, it is more difficult to provide QoS assurance for multipath routing.

One important component for any routing algorithm is the route discovery task, where good routes from the source node to the destination node are to be found for data delivery. For route discovery, each node maintains a routing table, which contains the routes to all other nodes in the network. The route update is performed periodically to keep track of the changes in the network topology and traffic load in the network [20]. To reduce control overhead and memory requirements, several hierarchical routing schemes were proposed in the literature [17]. Hierarchical routing schemes basically group wireless nodes into clusters, where intracluster and intercluster routes are found separately in different hierarchical levels of the routing architecture. For on-demand routing algorithms, route discovery is only performed when there is a demand to establish a route for an incoming connection [21], [22], [23]. Since on-demand routing scales well with the network size, most of the routing algorithms adopted by the IETF MANET Working Group belong to this routing category.

For on-demand QoS routing algorithms, link and path quality metrics are required in the route discovery phase to find good routes for an incoming connection. The two most popular QoS metrics for existing routing algorithms are bandwidth and delay [8]. In [10] and [12], the authors assumed that link delay can be estimated/measured with certain uncertainty. In practice, QoS metrics such as delay and loss rate can be calculated under realistic physical layer and link layer designs, considering the queuing dynamics at each of the nodes along the route to the destination. This leads to a tandem queuing problem, which is the focus of this paper. The queuing framework proposed in this paper captures the multirate transmission in the physical layer and the ARQ-based error recovery in the link layer. A recent work in the literature incorporated the retransmission effect due to an ARQ protocol into the link metric [18]. This work, however, did not consider any end-to-end QoS guarantee for incoming connections.

A tandem queuing model is also useful for evaluating the end-to-end performance for multihop wireless networks [27]. There are some tandem queuing models proposed in the literature. Tandem systems of two queues were modeled in discrete time in [28]. The end-to-end delay for time-division multiple access (TDMA) and ALOHA multiple access schemes was approximately derived in [29] for multihop networks, assuming constant bit rate traffic. The throughput of a tandem queuing system was investigated in [30]. In [31], the authors proved the concavity property for the throughput of tandem systems with buffer. However, the dynamics of arrival process and buffer overflow in the first queue in the tandem system were not considered. In [32], a decomposition approach for tandem queuing systems with blocking was proposed. A network calculus approach for statistical QoS provisioning of communication networks was proposed in [33]. Mainly proposed for wired networks, this approach, however, cannot be directly applied to wireless networks with sophisticated physical and link layer designs.

Since there are diverse techniques employed in the physical layer of different wireless standards, the notion of bandwidth depends on the underlying wireless technology. A physical channel can be provided by a spreading code in code-division multiple access (CDMA) systems [35], a subcarrier in orthogonal frequency-division multiplexing (OFDM) systems [15], or simply a frequency band in frequency-division multiple access (FDMA) systems [36]. Each of these physical channels may be divided into equal-sized time slots, and different time slots may be used for transmissions on different links in a common neighborhood (i.e., in CDMA-TDMA, OFDM-TDMA, and FDMA-TDMA systems).

The tandem queuing model presented in this paper can be applied to different wireless technologies, where different amounts of bandwidth can be allocated to each transmission link of the tandem system. The proposed decomposition queuing model is "distributed" in nature, and therefore, it can be incorporated into the route discovery component of a QoS routing algorithm. The proposed queuing and QoS routing framework provides a unified solution for the problem of QoS provisioning in multihop wireless networks, which has not been thoroughly treated in the literature so far.

3 System Model

3.1 Network Model

We consider a multihop wireless network with multiple ongoing connections, each of which spans several hops.

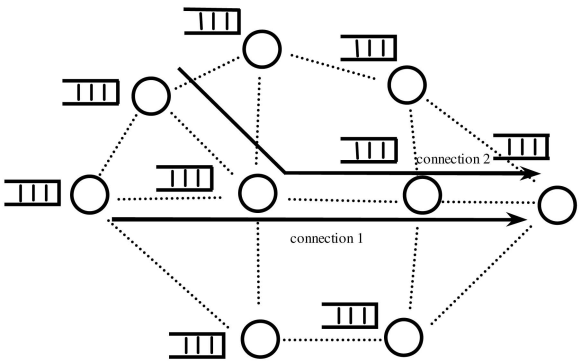

Fig. 1. A multihop wireless network with multiple ongoing connections.

Data traffic arriving at the source node is transmitted hop by hop to the destination node. We assume that each node in the network maintains a separate queue for each traffic flow traversing the link emanating from the node (i.e., per-flow queuing). A multihop network model with two ongoing connections is shown in Fig. 1, where, for convenience, we show only one queue at each node.

A particular amount of bandwidth is allocated for each hop along the routing path of the connection so that its end-to-end QoS requirements are satisfied. For a particular connection, the tandem system of queues along its routing path is illustrated in Fig. 2. This tandem queue has multiple concatenated queues, where the traffic coming out of each queue is fed into the next queue in the chain. The sequence of nodes that the traffic flow traverses is decided by a routing algorithm. The physical and link layer model for any hop along the routing path is described in the next section. Our objective is to find all end-to-end performance measures for a general tandem system of queues with an arbitrary number of hops. For notational convenience, we will occasionally construct a vector from the corresponding entries in the sequel. For example, vector d with elements d_i $(i = 0, 1, \cdots, M)$ will be denoted as $d = [d_0, d_1, \cdots, d_M]$.

3.2 Physical and Link Layer Model

We model the physical layer in a general way such that the tandem queuing model can be applied to many different physical layer technologies. Assume that there is a finite number of physical orthogonal channels separated in spreading code (e.g., for CDMA systems) or in the frequency domain (for OFDM or FDMA systems). The transmission time on each orthogonal channel is divided into fixed-sized time slots, which are occupied by only one link or are shared by different links in a common neighborhood, as in [10]. We will refer to the former case as a non-time-sharing system and the latter case as a time-sharing system.

For time-sharing systems, a number of consecutive time slots form a fixed-sized time frame, where the time slots in each time frame are periodically allocated for some transmission links in a common neighborhood. Each link may be allocated time slots in each time frame from different orthogonal channels. For non-time-sharing systems, each orthogonal channel is allocated to only one link. Therefore, a non-time-sharing system is a special case of a time-sharing system, where one time frame is equal to one

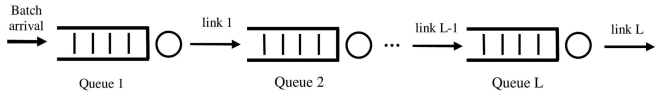

Fig. 2. A tandem queue.

time slot. In the queuing model, we observe the system states at the beginning of each time frame without explicitly stating the detailed resource allocation mechanism (i.e., either a time-sharing system or a non-time-sharing system).

In some practical scenarios, each wireless node is equipped with only one radio, which can either transmit or receive at any particular time. In addition, the number of orthogonal channels may not be large enough so that all the links in a common neighborhood can transmit/receive simultaneously. For these cases, a periodic transmission schedule may be constructed, which should resolve the collision of different simultaneous transmissions on the same channel and take care of the half-duplex constraint (i.e., one node has only one radio). The construction of such a schedule is an interesting and challenging research problem, but it is outside the scope of this paper. When the transmission schedule is known, the schedule length (i.e., the number of time slots in one period of the schedule) can be treated as the time frame mentioned before. Thus, the presented queuing model can still be applied. Note also that we allow spatial reuse exploitation, where wireless links weakly interfering with each other can transmit simultaneously. The interference of simultaneous transmissions is captured in the signal to interference plus noise ratio (SINR) at the receiving side of each link.

We assume that the packet length is fixed. The physical layer employs the AMC technique, where there are a finite number of transmission modes, each of which corresponds to a unique modulation and coding scheme [13], [14]. In addition, each transmission mode corresponds to one particular interval of the received SINR. Specifically, the SINR at the receiving side of a wireless link is partitioned into a finite number of intervals, with threshold values $X_0(=0) < X_1 < X_2 < \cdots < X_{K+1}(=\infty)$. If X is the SINR at the receiver, transmission mode k is employed if $X_k \le X < X_{k+1}$ $(k = 0, 1, 2, \cdots, K)$, which will be called the channel state k in the sequel.

The transmission rate at each transmission mode is proportional to its spectral efficiency. We assume that if the channel is in channel state k, c_k packets can be transmitted in one time slot. We also assume that $c_0 = 0$ (i.e., no packet is transmitted in channel state zero to avoid the high transmission error probability) and $c_K = H$. Assuming a Nakagami-m wireless channel model, the average packet error rate (PER) for transmission mode k can be obtained based on the Nakagami parameter m, the average SINR [14].

We will choose the SINR thresholds X_k such that the average PER for all transmission modes in allocated time slots of hop l is equal to a particular value denoted by $\beta^{(l)}$. In each hop, an infinite-persistent ARQ protocol is employed in the link layer, where an erroneous packet is retransmitted until it is received correctly at the receiving end of each hop. This is justifiable due to the fact that a large

number of retransmission attempts are usually recommended in the link layer to shield wireless errors from the higher layer [37]. Depending on the transmission outcome in each time frame, an acknowledgment (ACK) or a negative acknowledgment (NACK) is fed back from the receiver to the transmitter of each hop for each transmitted packet. We assume that the ACK/NACK packets are available at the end of the transmission time frame, and the feedback channel carrying ACK/NACK packets is a reliable one (e.g., due to the use of a strong error correction code and/or high transmission power).

The channel state is assumed to be stationary in each time frame, but it changes independently in consecutive time frames (i.e., block fading channel). We assume that the transmission link in hop l of the tandem system is allocated ω_l time slots from θ_l different orthogonal physical channels in each time frame. Note that the channel states in different time slots of one time frame on any allocated channel is the same. Let $\varphi_h^{(l)}(k)$ be the probability that the allocated channel h in hop l is in state k, which can be calculated using the channel parameters on the corresponding allocated channel.

Recall that the number of packets transmitted on any channel in one time slot varies depending on the corresponding channel state. Thus, the total number of packets transmitted on any link depends on the states of the channels allocated to that link. In addition, there may be several different channel state combinations for the allocated channels, which result in the same number of packets transmitted on one link. Now, assuming that the channel states of different allocated channels are independent, we can calculate the probability that i packets are transmitted during one time frame in hop l as

$$p_i^{(l)} = \sum_{k \in \Sigma_i} \prod_{h=1}^{h=\theta_l} \varphi_h^{(l)}(k), \tag{1}$$

where $0 \le i \le H\omega_l$ and Σ_i is the combination of all possible channel states on θ_l allocated channels for hop l such that the total number of packets transmitted in all allocated time slots is equal to i. For the tandem system, we will use the terms "link l" and "hop l" interchangeably in the sequel.

Example. A particular link l of the tandem system is allocated four time slots on two different orthogonal channels: time slots 1 and 2 on channel 1, time slots 3 and 4 on channel 2. The channel states in time slots 1 and 2 (also in time slots 3 and 4) in any time frame are the same because they belong to the same channel. If three and four packets can be transmitted in each time slot of channels 1 and 2, respectively, the total number of packets that can be transmitted in all allocated time slots for this link is $3 \times 2 + 4 \times 2 = 14$ packets.

4 An Exact Tandem Queue Model

In this section, we present an exact model for the tandem queue. We model the tandem queue in discrete time with a time unit being equal to a time frame. We observe the system state at the beginning of each time frame. We assume that traffic arrives at the source node buffer according to a batch Bernoulli arrival process, where i packets arrive in one time frame with probability a_i ($i = 0, 1, 2, \cdots, M$, where M can go to ∞). We assume that packets arriving in time frame $t-1$ can only be transmitted during time frame t at the earliest. When the time frame is large, it would be better to assume that packets can be served in the time frame that they arrive. This issue, however, can be adapted easily in the proposed queuing models. The queues of the tandem system are numbered using an increasing sequence of integers, where the source node maintains queue 1, and queue i has the buffer size of Q_i packets. Packets arriving at each buffer who could not find space will be dropped.

Packets successfully received at the receiving side of each link are buffered for either to be delivered to the application layer if it is the last link of the connection or transmitted to the next hop otherwise. The transmission rate on each link in each time frame depends on the channel states in the allocated time slots. We assume that all wireless links employ AMC with the same number of K transmission modes. The probability that i packets are transmitted on all allocated time slots of hop l is $p_i^{(l)}$, which can be calculated using (1).

4.1 Two Queue Case

We first consider a simple tandem system with two queues. The more general case with L queues ($L > 2$) will be considered in the next section. Let $q_i(t)$ be the number of packets in queue i in time frame t. The random process $X(t) = \{q_1(t), q_2(t)\}$ ($0 \le q_1(t) \le Q_1, 0 \le q_2(t) \le Q_2$) forms a discrete-time Markov chain (MC). For notational convenience, we omit the time index t in the related variables when it does not create confusion. Let (x, y) be the generic system state (i.e., $q_1 = x$ and $q_2 = y$) and $(x_1, y_1) \to (x_2, y_2)$ be the system transition from state (x_1, y_1) to state (x_2, y_2). The transition probabilities $\Pr\{(x_1, y_1) \to (x_2, y_2)\}$ for the underlying MC are derived in Appendix A.

Note that the number of packets transmitted on each link in any time frame is the minimum of the number of packets in the corresponding queue and the transmission capability in all allocated time slots. Recall that the maximum number of packets that can be transmitted in one time slot is H (i.e., $H = c_K$). Thus, the number of packets in queue one can be reduced at most by $N = H\omega_1$, where ω_1 is the total number of allocated time slots for link 1 in one time frame. Since there are at most M packets arriving at queue 1 (from the source node) and at most N packets enter queue 2 (due to successful transmissions from queue 1) in one time frame, the number of packets can increase at most by M for queue 1 and by N for queue 2.

Hence, if we write the transition probabilities $(x_1, *) \to (x_2, *)$ in a matrix block $\mathbf{A}_{x_1,x_2}$, the probability transition matrix of the MC $X(t)$ can be written as in (2), shown in Fig. 3. The order of matrix block $\mathbf{A}_{x_1,x_2}$ is $(Q_2+1) \times (Q_2+1)$, and its (y_1, y_2)th element is $\mathbf{A}_{x_1,x_2}(y_1, y_2) = \Pr\{(x_1, y_1) \to (x_2, y_2)\}$.

Now, we are ready to derive the steady state probabilities for MC $X(t)$. Let π be the steady state probability vector for $X(t)$. We have

$$\pi\mathbf{P} = \pi, \qquad \pi\mathbf{1} = 1, \tag{3}$$

$$\mathbf{P}=\begin{bmatrix}
\mathbf{A}_{0,0} & \mathbf{A}_{0,1} & \cdots & \mathbf{A}_{0,M} & & & & & & \\
\mathbf{A}_{1,0} & \mathbf{A}_{1,1} & \mathbf{A}_{1,2} & \cdots & \mathbf{A}_{1,M+1} & & & & & \\
\mathbf{A}_{2,0} & \mathbf{A}_{2,1} & \mathbf{A}_{2,2} & \mathbf{A}_{2,3} & \cdots & \mathbf{A}_{1,M+2} & & & & \\
\mathbf{A}_{N,0} & \mathbf{A}_{N,1} & \mathbf{A}_{N,2} & \mathbf{A}_{N,3} & \mathbf{A}_{N,4} & \mathbf{A}_{N,5} & \cdots & \mathbf{A}_{N,N+M} & & \\
 & & & & \ddots & \ddots & \ddots & \ddots & & \\
 & & & \mathbf{A}_{Q_1-1,Q_1-N-1} & \mathbf{A}_{Q_1-1,Q_1-N} & \mathbf{A}_{Q_1-1,Q_1-N+1} & \mathbf{A}_{Q_1-1,Q_1-N+2} & \cdots & \mathbf{A}_{Q_1-1,Q_1} & \\
 & & & & \mathbf{A}_{Q_1,Q_1-N} & \mathbf{A}_{Q_1,Q_1-N+1} & \mathbf{A}_{Q_1,Q_1-N+2} & \cdots & \mathbf{A}_{Q_1,Q_1} &
\end{bmatrix}. \quad (2)$$

Fig. 3. Transition probability matrix.

where $\mathbf{1}$ is a column vector of all 1s with the same dimension as π, which is $(Q_1+1)(Q_2+1)$. We can expand π as follows:

$$\pi=\left[\pi_0,\pi_1,\pi_2,\cdots,\pi_{Q_1}\right],$$

where π_i is a row vector of dimension Q_2+1, which can be further expanded as $\pi_i=\left[\pi_{i,0},\pi_{i,1},\pi_{i,2},\cdots,\pi_{i,Q_2}\right]$, where $\pi_{i,j}$ is the probability that the queuing system is in state (i,j). Given the steady state probability vector π, which is calculated using (3), we can derive the following end-to-end QoS measures.

4.1.1 End-to-End Loss Rate

Packets can be lost due to buffer overflow at one of the queues in the tandem. The buffer overflow probability for queue k can be calculated as a ratio between the average number of dropped packets due to overflow at queue k (denoted as $\overline{O}_k$) and the average number of packets arriving at queue k in one time frame (denoted as $\overline{A}_k$). Hence, the buffer overflow probability for queue k can be written as $P_l^{(k)}=\frac{\overline{O}_k}{\overline{A}_k}$.

Note that the average number of packets arriving at queue 1 in one time frame is $\overline{A}_1=\sum_{i=1}^{M} ia_i$. To calculate the average number of dropped packets due to overflow at queue 1, let us define z_i as the marginal probability that there are i packets in queue 1. We have $z_i=\pi_i\mathbf{1}_{Q_2+1}$, where $\mathbf{1}_{Q_2+1}$ is a column vector of all 1s with dimension Q_2+1. The average number of dropped packets due to overflow at queue 1 can be calculated as

$$\overline{O}_1=\sum_{i=1}^{M}\sum_{j=Q_1-M}^{Q_1} a_i z_j\times\max\{0,i+j-Q_1\},$$

where $\max\{0,i+j-Q_1\}$ is the number of dropped packets (if any), given that there are j packets in queue 1 and i arriving packets. Now, we calculate the buffer overflow probability at queue 2. We first determine the arrival probability for packets entering queue 2 due to successful transmissions from queue 1. In fact, the number of packets arriving at queue 2 is those successfully transmitted over link 1. The probability that i packets arrive at queue 2 can be approximated as

$$b_i\approx\sum_{k=0}^{Q_1}\sum_{l=0}^{N} z_k p_l^{(1)}\times\gamma^{(1)}(\min\{k,l\},i),$$

where $\gamma^{(l)}(j,i)$ is the probability that i packets are correctly received, given that j packets were transmitted over link l of the tandem system, which is given in Appendix A. Recall that z_k is the probability of having k packets in queue 1, and $p_l^{(1)}$ is the probability that l packets can be transmitted in all allocated time slots over link 1.

The average arrival rate to queue 2 can be calculated as $\overline{A}_2=\sum_{i=1}^{N} ib_i$. To calculate the average number of dropped packets due to overflow at queue 2, let us define w_i as the marginal probability that there are i packets in queue 2, which can be calculated as $w_i=\sum_{j=0}^{Q_1}\pi_{j,i}$. Similar to that for queue 1, the average number of dropped packets due to overflow at queue 2 can be calculated as

$$\overline{O}_2=\sum_{i=1}^{N}\sum_{j=Q_2-N}^{Q_2} b_i w_j\times\max\{0,i+j-Q_2\}.$$

Finally, the end-to-end loss rate can be approximated as

$$P_l\approx 1-(1-P_l^{(1)})(1-P_l^{(2)}), \quad (4)$$

where the loss due to overflow at both buffers are taken into account, and this approximation is tight when the loss rates at different queues are weakly dependent, as will be validated in Fig. 6.

4.1.2 End-to-End Average Delay

The end-to-end delay is the sum of delays that any packet experiences in all queues and links along its routing path. Assuming that the propagation delay over the wireless channel is negligible (i.e., only the waiting time in the buffers and delay due to retransmissions of the ARQ protocol is considered in the calculation), using Little's law, the end-to-end average delay can be written as

$$D=\frac{\sum_{i=1}^{Q_1} iz_i}{\overline{A}_1(1-P_l^{(1)})}+\frac{\sum_{i=1}^{Q_2} iw_i}{\overline{A}_2(1-P_l^{(2)})}, \quad (5)$$

where the numerator of each term is the average length of each queue, and the denominator is the average arrival rate considering packet loss due to overflow.

4.2 General Case $(L>2)$

We consider a general tandem system with L queues $(L>2)$, which are concatenated to each other as a chain (see Fig. 2). The buffer size of queue i is assumed to be Q_i packets. Similar to the previous section, let $q_i(t)$ be the number of packets in queue i in time frame t $(i=1,2,\cdots,L)$. The random process $Y(t)=\{q_1(t),q_2(t),\cdots,q_L(t)\}$ $(0\le q_1(t)\le Q_1, 0\le q_2(t)\le Q_2,\cdots,0\le q_L(t)\le Q_L)$ forms a discrete-time MC.

An approach similar to that in Section 4.1 can be pursued to obtain the transition probabilities for this MC. The number of state transitions for this MC, however, grows

exponentially with the number of queues in the tandem system. In fact, the order of the transition probability matrix $\mathbf{P}$ is $\prod_{i=1}^{L}(Q_i+1)\times\prod_{i=1}^{L}(Q_i+1)$. Therefore, the computational complexity is very high for the large number of queues and large buffer sizes. Theoretically, we can follow the procedure similar to that in Section 4.1 to derive the transition probabilities and obtain the steady state probability vector and, subsequently, the end-to-end performance measures.

5 Solution of the Tandem Queuing Model: Decomposition Approach

We present a novel decomposition approach to solve the general tandem queue, where the computational complexity grows only linearly with the number of queues in the system. For ease of reference, buffers (queues) along the routing path are numbered by an increasing sequence of integers with the buffer at the source node denoted as buffer (queue) 1.

5.1 Markov Chain and Steady State Probability

We consider the tandem system with L queues as in Fig. 2. For notational convenience, we assume that i packets arrive at queue k with probability $a_i^{(k)}$ ($a_i^{(2)}=b_i$ for the two queue case considered in Section 4.1). Note that the maximum batch size (the maximum number of packets arriving in one time frame) captured in $a_i^{(k+1)}$ for $k\geq 1$ is $N_k=H\omega_k$, whereas the maximum batch size captured in $a_i^{(1)}$ for the first queue is M. The buffer sizes for all queues are as in Section 4.2, and the queuing rules are as in Section 4.

We observe that the behavior of queue $i+1$ does not impact queue i in the chain. This is because the outcomes (i.e., successfully transmitted packets) from queue i are fed into queue $i+1$. Thus, instead of forming the MC, which captures the queue length dynamics of all queues, we could find the queue length dynamics for one queue at a time where its input is the output of the previous queue in the chain (except for queue 1).

Specifically, we pursue the following steps: First, form the MC for queue 1 and calculate the corresponding steady state probability vector. Based on the steady state probabilities, we calculate the packet arrival probabilities to the next queue. These arrival probabilities are used to derive the arrival probabilities for the next queue in the chain. This procedure is repeated until we obtain the solutions for the last queue of the tandem system.

Obviously, by using this decomposition approach, the joint steady state probability vector could not be found as in Section 4. However, the steady state probability vector for each queue in the chain is what we need to calculate the desired queuing performance measures. Essentially, the presented procedure requires us to solve L separate queues, each of which accepts batch arrival traffic and serves packets also in batches. Let us consider a particular queue k of the chain and form the MC $X_k(t)=\{q_k(t)\}$, $(0\leq q_k(t)\leq Q_k)$, where $q_k(t)$ denotes the number of packets in queue k with the arrival process described by $a_i^{(k)}$. The transition probabilities for this MC are derived in Appendix B.

Given the transition probabilities, we can easily calculate the steady state probability vector of this MC, which is denoted as $\pi^{(k)}=[\pi_0^{(k)},\pi_1^{(k)},\cdots,\pi_{Q_k}^{(k)}]$, where $\pi_i^{(k)}$ denotes the probability that there are i packets in queue k.

5.2 End-to-End Loss Rate and Average Delay

As in Section 4, the buffer overflow probability at queue k can be calculated as

$$P_l^{(k)}=\frac{\overline{O}_k}{\overline{A}_k}. \tag{6}$$

The average arrival rate at queue k can be written as $\overline{A}_k=\sum_{i=1}^{B^{(k)}} i a_i^{(k)}$, where $B^{(k)}$ is the maximum batch size of the arrival process to queue k. The probability that i packets are successfully transmitted from queue k and arrive at queue $k+1$ can be approximated as

$$a_i^{(k+1)}\approx\sum_{j=0}^{Q_k}\sum_{l=0}^{N_k}\pi_j^{(k)}p_l^{(k)}\times\gamma^{(k)}(\min\{j,l\},i). \tag{7}$$

These arrival probabilities are used to derive the queuing solution for queue $k+1$, as mentioned before, and the average number of dropped packets due to overflow at queue k can be calculated as

$$\overline{O}_k=\sum_{i=1}^{B^{(k)}}\sum_{j=Q_k-B^{(k)}}^{Q_k}a_i^{(k)}\pi_j^{(k)}\times\max\{0,i+j-Q_k\}.$$

The end-to-end loss rate can be approximated as

$$P_l\approx 1-\prod_{k=1}^{L}(1-P_l^{(k)}), \tag{8}$$

and the end-to-end average delay can be written as

$$D=\sum_{k=1}^{L}D_k, \tag{9}$$

where D_k is the average queuing delay at queue k, which is given by

$$D_k=\frac{\sum_{i=1}^{Q_k} i\pi_i^{(k)}}{\overline{A}_k(1-P_l^{(k)})}. \tag{10}$$

5.3 End-to-End Delay Distribution

The proposed decomposition approach for tandem queues enables us to derive the end-to-end delay distribution with reasonable accuracy, which is necessary for statistical delay provisioning in multihop wireless networks. Let $\Omega_{i,l}^{(k)}$ denote the probability that i packets are successfully transmitted from queue k in l time frames. Because the tagged packet can see at most Q_k-1 head of line (HOL) packets, the probability that the tagged packet sees i HOL packets is $\chi_i^{(k)}=\pi_i^{(k)}/(1-\pi_{Q_k}^{(k)})$. The probability that the tagged packet experiences delay of l time frames in queue k of the tandem system can be calculated as

$$P_l^{(k)}=\sum_{i=0}^{\Delta_l}\chi_i^{(k)}\Omega_{i+1,l}^{(k)}, \tag{11}$$

where $\Delta_l = \min\{lN_k - 1, Q_k - 1\}$ because at most N_k packets can be transmitted in one time frame and the tagged packet sees at most $Q_k - 1$ HOL packets. Let us put the delay distribution at each queue k of the tandem system into a vector $P_d^{(k)}$ and put the end-to-end delay distribution vector into vector P_d. Then, we have

$$P_d \approx \otimes_{k=1}^{L} P_d^{(k)}, \tag{12}$$

which is obtained by performing convolutions of L vectors $P_d^{(k)}$ $(k = 1, \cdots, L)$. Note that the first element of vector P_d represents the probability that the end-to-end delay is L time frames, which is the minimum end-to-end delay. The remaining task is to derive $\Omega_{i,l}^{(k)}$, which can be done by using the following recursive relations:

$$\Omega_{i,l}^{(k)} = \sum_{j=0}^{N_k} \Psi_{i,j}^{(k)} \Omega_{i-j,l-1}^{(k)}, \qquad \Omega_{0,0}^{(k)} = 1, \tag{13}$$

where $\Psi_{i,j}^{(k)}$ is the probability that j packets are successfully transmitted from queue k, given that there were i packets in this queue before transmissions. Equation (13) can be interpreted as follows: If there are i packets in queue k, which must be transmitted in l time frames, and j packets are transmitted in the first time frame, there remain $i - j$ packets to be transmitted in $l - 1$ time frames. Now, $\Psi_{i,j}^{(k)}$ can be calculated as

$$\Psi_{i,j}^{(k)} = \sum_{v} p_v^{(k)} \times \gamma^{(k)}(\min\{i, v\}, j), \tag{14}$$

where the sum includes only v such that $\min\{i, v\} \geq j$.

6 Application of the Tandem Queuing Model for Quality-of-Service Routing

We show how we can incorporate the proposed tandem queuing model into a QoS routing algorithm. The tandem queuing models proposed in Sections 4 and 5 are solved for a particular connection, given that its routing path is known. Now we want to tackle the reverse problem, where the tandem queuing model is used to discover a route for a connection from a source node to its desired destination node such that the QoS requirements for the connection are satisfied. One possible approach for QoS routing is to find all possible routes from the source to the destination. The source node, upon gathering all possible routes, will check the routes one by one to find the best feasible route by using the presented tandem queuing model. This approach, however, results in a very large amount of signaling/communication overhead in the route discovery phase and a large computational burden for the source node.

We observe that the decomposition approach for the tandem queue has a nice distributed nature such that the link QoS metrics (i.e., the average link delay D_k and loss rate $P_l^{(k)}$ for link k) can be calculated if the routing path up to a particular hop is known. The decomposition approach can be used as an efficient tool to search for feasible routes such that the QoS requirements of the connection are satisfied. In addition, the route discovery can be done on a hop by hop basis, and only the potentially feasible routes are explored further. This reduces the route searching overhead significantly and therefore avoids huge computation effort at the source node.

The unique feature of our queuing and routing framework is that the three most important QoS metrics, namely, end-to-end bandwidth, delay, and loss rate, can be integrated into the QoS routing algorithm compared with only delay and/or bandwidth, as done in traditional QoS routing algorithms in the literature [8], [10], [11], [12]. In addition, most of the existing work assumed that link delay can be measured and/or estimated in a timely manner [8], [12]. The dynamics of the traffic arrival process, wireless channel fading, and complex physical and link layer designs of wireless systems, however, would render this delay measurement/estimation a time-consuming task. Our queuing model provides an accurate and efficient tool for link metric calculation.

In the remainder of this section, we will describe an on-demand unicast QoS routing algorithm by using the tandem queue model based on the decomposition approach presented in Section 5. Similar to other QoS routing algorithms in the literature, two main components of our QoS algorithm are route discovery with bandwidth reservation and route maintenance. The QoS constraints for an incoming connection are end-to-end bandwidth, average delay, and loss rate. An additional statistical delay requirement can be also imposed by an incoming connection.

6.1 Route Discovery and Resource Reservation

To establish a connection, the source node broadcasts the route request packet (RRQ) into the network to search for good routes to the destination node, which satisfy the QoS requirements of the connection. The incoming connection submits the traffic profile (i.e., packet arrival probabilities to the source node buffer $a_i^{(1)}$) and its QoS requirements to the source node. The RRQ packet contains the addresses of the source node and the destination node, the request ID, and the end-to-end QoS requirements. Let the target QoS requirements for an incoming connection c be end-to-end bandwidth $B(c)$, end-to-end average delay $D(c)$, end-to-end loss rate $P_l(c)$, and an optional end-to-end statistical delay requirement of the form $\Pr\{\text{end-to-end delay} > D_t(c)\} \leq P_t(c)$.

The required amount of bandwidth $B(c)$ needs to be reserved for each link along the routing path. The bandwidth here refers to the time slots for transmissions using different channels. On any orthogonal channel, a particular time slot can be allocated to only one link in a common neighborhood. For ease of presentation, we assume that a static resource allocation scheme is adopted, where each link in the network is allocated a certain number of time slots for transmission using some orthogonal channels from the set of available channels (the allocation is assumed to be repeated in each frame time if time sharing is implemented). An incoming connection may take some of these preallocated time slots of the link (i.e., the bandwidth taken by the connection is equal to its bandwidth requirement) if its routing path traverses the corresponding link. This static predetermined allocation should be done such that two different links using the same channel in a common neighborhood are not allocated the same time slot. Each node is assumed to have enough radios to

communicate with its neighbors on the allocated channels, as in CDMA or multichannel networks [10], [36].

When a node receives the RRQ packet, it checks the available bandwidth on the outgoing links, and only outgoing links having enough bandwidth to accommodate the new connection are considered further. The outgoing links with enough bandwidth will be called the BW-feasible links. A more flexible resource allocation scheme would allow a link to borrow bandwidth from its neighboring links if this mechanism can potentially enhance the system performance. This scheme, however, requires local negotiation and reservation. We assume that the transmitter of each link knows the channel parameters of the link (i.e., the average SINR at the receiving end and the Nakagami parameter m) by using some estimation technique.

Now, we describe how each node calculates the link QoS metrics together with the resource reservation mechanism mentioned above. Initially, the source node of the incoming connection calculates the average link delay and loss rate for each of the outgoing links, which has enough bandwidth to accommodate the incoming connection. This is done by using the queuing model presented in Section 5.1 based on the submitted traffic profile and link channel parameters at the transmitting node. For outgoing links that satisfy the connection QoS requirements, the source node calculates the arrival probabilities to the next node along the corresponding outgoing link as in (7), records these arrival probabilities, average link delay, and loss rate into the RRQ packet header, and forwards the RRQ packet to the receiving node of the corresponding link. The receiving nodes of these feasible links join a set of nodes called the broadcast group (BG). Each node, upon receiving the RRQ packet, calculates the link delay, the loss rate using the arrival probabilities retrieved from the RRQ packet header, and the channel parameters for its BW-feasible outgoing links. The node then accumulates the QoS metrics and checks the feasibility of the QoS requirements for the connection. For each feasible outgoing link, the node records the arrival probabilities for the next node, route metrics into the RRQ packet header, and forwards the RRQ packet to the corresponding node.

In the above procedure, only routing paths along the links that have enough bandwidth required by the incoming connection with the feasible path metrics are explored further. Now, we describe how we can calculate the path metric and choose the best routing path. Let the average delay and loss rate over link (i,j) be $D(i,j)$ and $P_l(i,j)$, respectively. The end-to-end average delay and loss rate for routing path $R = i \rightarrow j \rightarrow \ldots k \rightarrow l$ can be adapted from (8) and (9) as

$$\begin{aligned} \text{delay}(R) &= D(i,j) + \cdots + D(k,l), \\ \text{loss}(R) &= 1 - (1 - P_l(i,j)) \ldots (1 - P_l(k,l)) \\ &\approx P_l(i,j) + \cdots + P_l(k,l), \end{aligned}$$

where the approximation is tight for small loss rate P_l. In addition, $P_l(i,j)$ and $D(i,j)$ can be calculated by using (6) and (10), respectively, for the corresponding link (hop). Since there are multiple QoS requirements, the definition of the best routing path is not unique. To resolve this issue, we define the weighted average QoS metric as follows:

$$\text{metric}(R) = \alpha \frac{\text{delay}(R)}{D(c)} + (1 - \alpha) \frac{\text{loss}(R)}{P_l(c)}, \qquad (15)$$

where α determines the importance of the delay requirement in comparison with the loss rate requirement.

A node may receive the RRQ packet with the same request ID more than once. If the QoS metric (i.e., $\text{metric}(R)$) retrieved from the RRQ packet header for the current reception is smaller than that due to the previous reception, it will rebroadcast the RRQ packet with the new QoS metrics. Finally, if a node finds out that it is the destination of the connection, it sends the route reply packet (RRP) back to the source node. If the destination node finds a route to the source node in its route cache, it can send the RRP packet along this route. If this reverse route is not available but each link along the newly discovered route works well in both directions, the RRP packet follows the reverse route to reach the source node. If the reverse route does not work well, the RRP packet can be piggybacked (as in [21]). In addition, we may use one of the following ways to record the feasible routes: The first way is to record the route into the RRQ and RRP packet headers, and the second way is to record the route at intermediate nodes in a hop by hop basis.

Before transmitting data on the newly discovered route, each link along this route updates its available bandwidth for possible future connections. The available bandwidth on these links will also be updated when the connection is released. One important issue here is to limit the overhead caused by the route discovery process. One way to reduce the overhead is to use ticket-based route discovery, as proposed in [12]. Another way is to record the number of hops that the RRQ packet has traversed and limit the number of hops that the RRQ can be broadcast. However, another approach is to use a timeout mechanism to discard an RRQ packet. The route discovery algorithm is summarized in Fig. 4.

6.2 Route Maintenance

Another important task for any routing algorithm is route maintenance to make sure that the route works well during the lifetime of the connection. Some links along the route may be broken due to factors such as node mobility and the degradation of wireless links. Because we consider wireless systems which employ the link level ARQ error recovery protocol, a broken link can be discovered if the transmitting node of a link does not receive ACK/NACK packets within a predetermined timeout period. The node that detects a link break sends a route error packet to the source node by using the same technique used for sending the RRP packet. The receiving node of the broken link can also detect the link break if it does not receive any data packet within a predetermined timeout period. This receiving node, upon detecting the link break, will also send the route error packet to release the route in the forward direction.

When the source node receives the route error packet, it may initiate a route discovery to find a new feasible route to the destination. Since the connection setup time may be very long, we may maintain multiple feasible routes. This can be done if the destination node sends the RRP packet containing several feasible routes to the source node. The

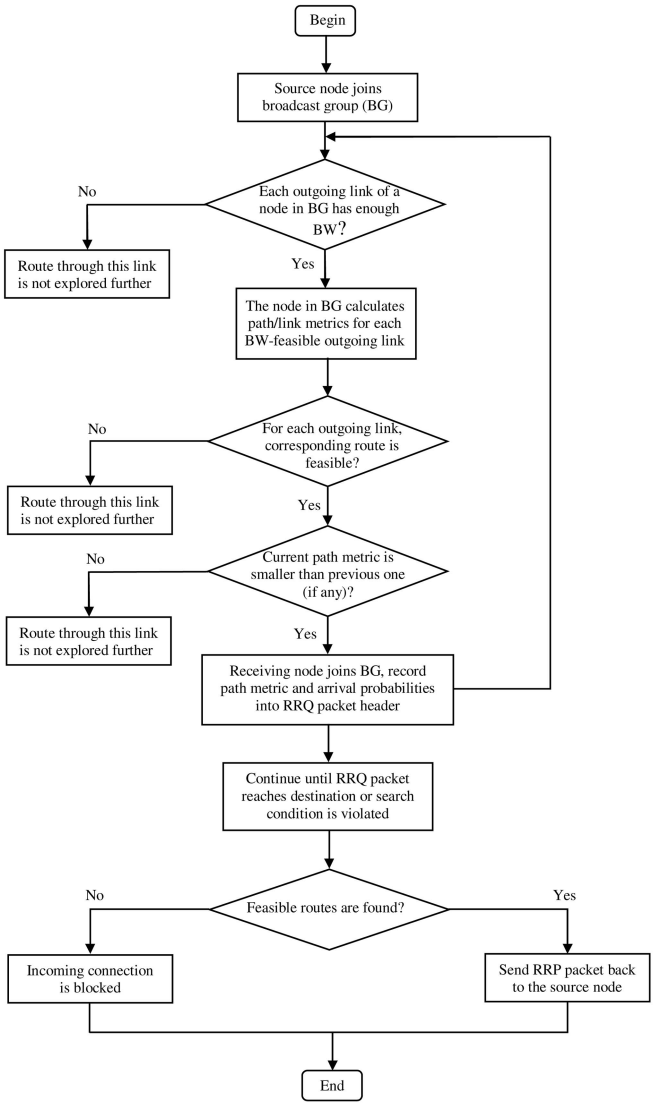

Fig. 4. Route discovery algorithm for QoS routing.

route with the smallest QoS metric will be used for data transmission. If this best route is broken, the source node may try using the next best route that it has received through the RRP packet. In addition, the QoS of the chosen route may degrade during the lifetime of the connection. Hence, the source node may periodically send the route maintenance packet along the route to check the QoS feasibility of the route. If the QoS feasibility condition of the current route is violated, the next best route may be used, or a route discovery may be initiated to find an alternative route to the destination.

7 Validation of Decomposition Approach and Typical Numerical Results

7.1 Parameter Setting

We consider wireless networks employing adaptive M-ary quadrature amplitude modulation (M-QAM) without coding by using five transmission modes for all transmission links. We assume that $c_k = k$ (k packets are transmitted in one time slot in channel state k) and the

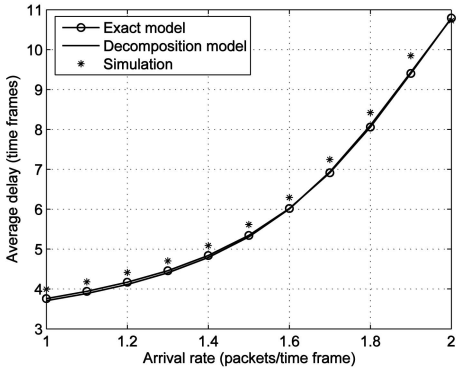

Fig. 5. End-to-end average delay versus packet arrival rate for a tandem system of two queues (for $Q = 20$, $L = 2$, the Nakagami parameter $m = 1$, and the average $\mathrm{SINR} = 15$ dB for both hops).

Nakagami parameter $m = 1$ (i.e., the Rayleigh fading channel). The SINR switching thresholds for the transmission modes are chosen such that the average PER satisfies $\overline{\mathrm{PER}}_k = 0.1$ for all transmission modes in all hops (i.e., $\beta^{(l)} = 0.1$ for $l = 1, 2, \cdots, L$). The fitting parameters to calculate the PER are available in [14, Table 1]. The arrival probability vector to the source node queue is chosen to be $a^{(1)} = [1 - 25A/48, A/4, A/8, A/12, A/16]$, where the average arrival rate is A. For all the results presented here, the buffer sizes of all queues in the tandem system are the same, which will be denoted as Q.

The numerical results for QoS routing are obtained for networks with 20 nodes, which are randomly generated in an area of 1,000 m $\times$ 1,000 m. Node mobility is not considered in deriving the results. We consider a non-time-sharing system with static resource allocation, where separate sets of orthogonal channels are allocated to different links in the network. Each node uses a fixed transmit power level on each allocated channel, and the average SINR at the receiving node of link (i, j) is modeled as $\mathrm{SINR}(i, j) = K \times d(i, j)^{-3}$, where $K = 5 \times 10^8$ captures transmission power, antenna gain, and other factors, $d(i, j)$ is the distance from node i to node j, and the path-loss exponent is assumed to be 3. Note that these assumed values are for presenting the illustrative results only, whereas the queuing and QoS routing framework can be applied to other network settings.

7.2 Numerical Results and Validation of Queuing Models

We validate the decomposition approach for the tandem queue and present some typical numerical results. Each link of the tandem system is allocated one time slot in each time frame. The average SINR for all links of the tandem system of queues is chosen to be 15 dB. Typical variations in the end-to-end average delay and end-to-end loss rate with packet arrival rate are shown in Figs. 5 and 6, respectively, for a tandem system of two queues. In these two figures, we show results obtained from the exact queuing model (presented in Section 4), the decomposition approach (presented in Section 5), and the simulations. As is evident, the decomposition approach provides accurate measures for average delay and loss rate. The analytical results also

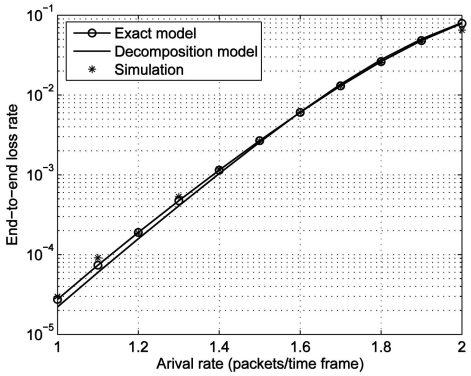

Fig. 6. End-to-end loss rate versus packet arrival rate for a tandem of two queues (for $Q = 20$, $L = 2$, the Nakagami parameter $m = 1$, and the average $\mathrm{SINR} = 15$ dB for both hops).

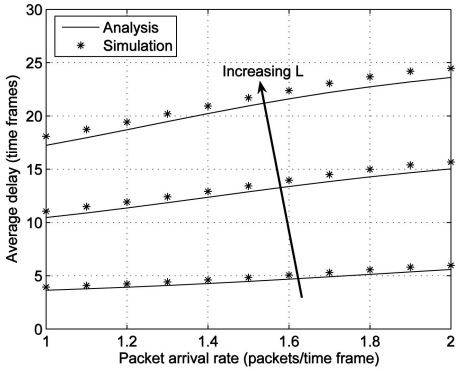

Fig. 8. End-to-end average delay versus packet arrival rate for tandem systems consisting of different numbers of queues (for $Q = 10$, $L = 2, 6, 10$, the Nakagami parameter $m = 1$, and the average $\mathrm{SINR} = 15$ dB).

follow the simulation very closely, which confirms the correctness of the proposed queuing models.

We illustrate the complementary cumulative delay distributions (obtained in Section 5.2) in Fig. 7 for tandem systems with different numbers of queues $(L = 2, 4, 6)$. The results obtained from the simulations are also presented. Note that the complementary cumulative delay distribution is represented by the probabilities $\Pr(\text{delay} > D) = 1 - \sum_{k=1}^{D} P_d(k)$, which can be calculated by using P_d in (12).

We observe that the analytical model slightly overestimates the end-to-end delay in the statistical sense (i.e., $\Pr(\text{delay} > D)$ obtained from the analytical model is greater than that due to simulation for a given value of D). In fact, the end-to-end delay of a target packet is the time that it spends in all queues of the tandem system. To calculate this delay, we can turn off the arrival traffic to the tandem system after the target packet enters the tandem system. This is because arriving packets following the target packet do not impact the delay experienced by the target packet.

Due to "turning off" the arrival and the batch transmission effects (because of multirate transmission), in a statistical sense, the target packet would see a smaller number of HOL packets in queue 2 onward, as compared to the marginal distribution derived in (10). Therefore, by calculating the delay distribution at each queue by using the queue length distribution in (10), we overestimate the traffic arrival probabilities to the queues in the chain, except for queue 1. A more accurate model can be developed by tracking the number of HOL packets in all queues until the target packet leaves the last queue of the chain. This procedure, however, has a very high computational complexity. The approximation method presented in Section 5.2 results in reasonably accurate results with low complexity.

Typical variations in end-to-end average delay and loss rate with packet arrival rate are presented in Figs. 8 and 9 for tandem systems consisting of different numbers of queues, respectively. As expected, the end-to-end average delay increases almost linearly with the number of queues in the tandem. In fact, the packet arrival rate to each queue in the chain is roughly the same because the loss rate at each of the queues is quite small. Thus, the average queuing delay at each queue is roughly the same, given the same service rate (i.e., all transmission links are assumed to be the same). The variations in end-to-end loss rate with the number of hops can be interpreted in a similar manner, considering the approximation obtained in (15) for small loss rates. These two figures show that the decomposition approach can accurately calculate these two important performance measures, namely, end-to-end average delay and loss rate.

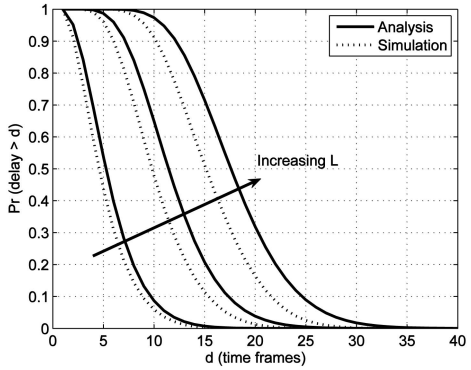

Fig. 7. End-to-end complementary cumulative delay distribution for tandem systems with different numbers of queues (for $Q = 20$, the Nakagami parameter $m = 1$, the average $\mathrm{SINR} = 15$ dB for all hops, the number of queues $L = 2, 4, 6$, and the packet arrival rate = 1.5 packets/time frame).

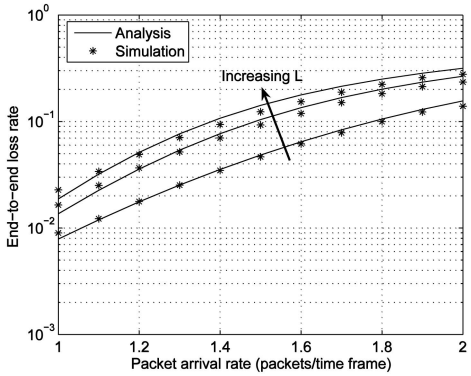

Fig. 9. End-to-end loss rate versus packet arrival rate for tandem systems with different numbers of queues (for $Q = 10$, $L = 2, 6, 10$, the Nakagami parameter $m = 1$, and the average $\mathrm{SINR} = 15$ dB).

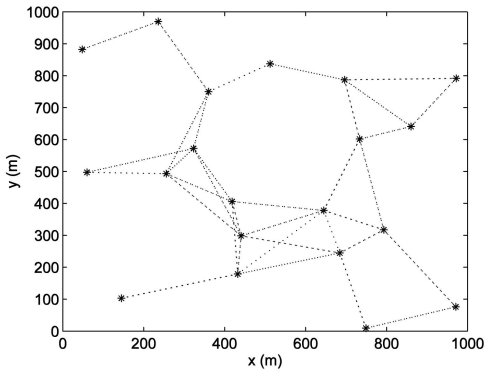

Fig. 10. Simulation topology with 20 nodes.

7.3 Numerical Results for the Proposed QoS Routing Algorithm

We present some illustrative numerical results for the proposed QoS routing algorithm, which is based on the decomposition approach for the tandem queue for a network with 20 nodes, as shown in Fig. 10. A link exists between any two nodes if the distance between them is less than 300 m. For the results presented in this section, each connection requires one channel for each link along its routing path. For all the presented results, we allow the RRQ packet to traverse at most eight hops from the source node in the network. In addition, all results in this section are obtained by running the routing algorithm over 10^4 time frames.

We assume that connection requests arrive at each node in each time frame with connection arrival probability λ_c. Each incoming connection submits the traffic profile to the source node who initiates the route discovery process to find a routing path to its desired destination. The destination node for each incoming connection is chosen randomly among the remaining nodes. If the QoS routing algorithm succeeds in finding a feasible routing path, the connection remains in the network for a random interval, which is exponentially distributed with mean value equal to $\mu_c = 200$ time frames.

We show typical variations in the connection blocking probability with network load, where each link in the network is statically allocated different numbers of channels. The network load is calculated as $\rho = (\text{number of nodes}) \times \mu_c \times \lambda_c$. We vary the network load by changing the connection arrival probability λ_c. When each link in the network is allocated more channels, the network capacity increases; therefore, the connection blocking probability decreases. In addition, the connection blocking probability increases with the network load. Fig. 11 shows that the connection blocking probability decreases significantly when the number of channels allocated to each link increases from one to three.

Fig. 12 shows variations in the connection blocking probability with packet arrival rate for different sets of QoS requirements. As expected, the more stringent the QoS requirements, the less the probability that the routing algorithm can find a feasible route to the destination. However, Fig. 12 shows that the performance degradation in terms of the connection blocking probability may be moderate, even when more stringent QoS requirements are imposed by the arriving connections. This implies that the proposed queuing and QoS routing framework performs load balancing well by finding the low-load routes (if any).

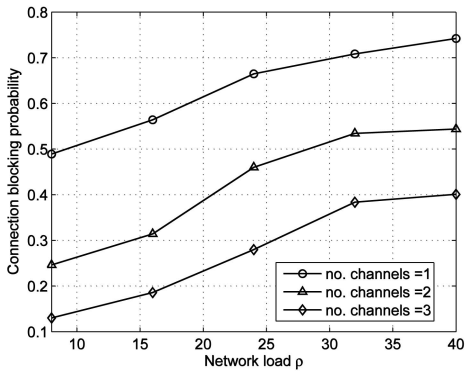

Fig. 11. Connection blocking probability versus network load for different numbers of channels (for $Q = 10$, $\alpha = 0.5$, the average connection holding time = 200 time frames, the packet arrival rate for each connection = 2 packets/time frame, and the QoS requirements are $B(c) = 1$ channel, $D(c) = 20$ time frames, and $P_l(c) = 0.2$).

8 Extension to Multihop Wireless Networks with Class-Based Queuing

In the previous sections, we have presented the tandem queuing models and their applications for QoS routing, assuming per-flow queuing at each node along a routing path. However, per-flow queuing may not be scalable. In contrast, a class-based queuing would provide a more scalable solution, where each node maintains a finite number of queues corresponding to a finite number of service classes with differentiated QoS requirements.

In this section, we extend the per-flow queuing-based QoS routing framework to a class-based QoS routing framework. With class-based queuing, the transmitting node of each link maintains a finite number of queues for each link, which corresponds to different service classes. Traffic from connections of the same service class traversing a particular link are buffered in the same queue. The bandwidth allocated to each queue depends on the

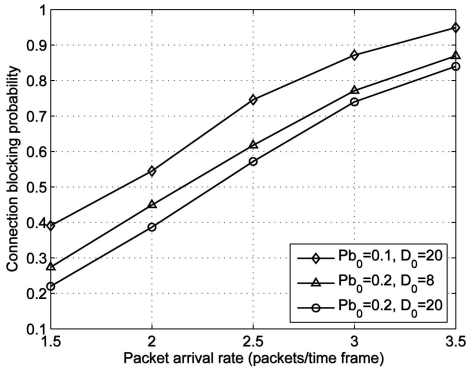

Fig. 12. Connection blocking probability versus packet arrival rate for different QoS requirements (for $Q = 10$, $\alpha = 0.5$, the number of channels per link = 2, the connection arrival probability = 0.005, and the average connection holding time = 200 time frames).

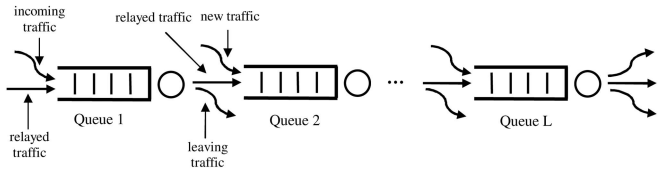

Fig. 13. A tandem queue for one connection of a particular service class.

bandwidth requirement of each connection and the number of connections being served by the queue.

8.1 Tandem Queuing Model

We show how we can extend the per-flow queuing model to this class-based queuing implementation. Specifically, we consider the tandem system of queues along a routing path for a connection of a particular service class, as shown in Fig. 13. Data traffic entering each queue may come from different connections. As shown in this figure, traffic from connections other than the considered connection may come and leave the tandem system at any queue.

Note that traffic of all connections entering a particular queue of the tandem system has the same queuing performance. Now, using the decomposition approach similar to that presented in Section 5, we need to solve L single queues, where the arrival traffic to each queue is from the previous queue of the tandem system and from other connections traversing the corresponding link (except for queue 1). Given the allocated bandwidth for each link along the tandem system, we can determine the service rate probabilities from (1). Thus, to calculate the queuing performance measures (i.e., the overflow probability and delay) for each queue in the tandem system, we need to determine the arrival probabilities for the aggregate traffic to the considered queue.

For each queue of the tandem system (except for queue 1), we consider new, relayed, and leaving traffic, as shown Fig. 13. Note that these traffic sources are aggregated from different connections. Since traffic flows from different connections may be transmitted over several links of the tandem system, we need to keep track of the connections, which contribute to the traffic on each link. Let the sets of connections that contribute to the relayed and the leaving traffic flows from queue k of the tandem be $\Phi_{\text{rel}}^{(k)}$ and $\Phi_{\text{lev}}^{(k)}$, respectively, and the set of connections that contribute new traffic flow to queue k be $\Phi_{\text{new}}^{(k)}$.

Now, denote the arrival probability vector to queue k due to connection c as $a^{(k,c)}$ and its average arrival rate as $\overline{A}^{(k,c)}$. In addition, let $a_{\text{lev}}^{(k)}$ denote the aggregate arrival probability vector of leaving traffic from queue k and let $a_{\text{new}}^{(k)}$ and $a_{\text{rel}}^{(k)}$ denote the aggregate arrival probability vectors of the new and relayed traffic to queue k, respectively. We have

$$a_{\text{new}}^{(k)} = \otimes_{c\in\Phi_{\text{new}}^{(k)}} a^{(k,c)}, \tag{16}$$

where $a_{\text{new}}^{(k)}$ is obtained by taking the convolutions of the arrival probability vectors $a^{(k,c)}$. As shown in Fig. 13, data packets successfully transmitted from queue k may enter queue $k+1$ (i.e., relayed traffic) or leave the considered tandem system (i.e., leaving traffic). These data packets that constitute the relayed traffic and the leaving traffic flows belong to connections in the sets $\Phi_{\text{rel}}^{(k)}$ and $\Phi_{\text{lev}}^{(k)}$, respectively. Given the allocated bandwidth for link l and the arrival probabilities to queue k, we can calculate the probabilities for data packets successfully transmitted from queue k as in (7). Let $a_{\text{suc}}^{(k)}$ be the probability vector for data packets successfully transmitted from queue k, where its element $a_{\text{suc},i}^{(k)}$ can be calculated by adopting (7) as follows:

$$a_{\text{suc},i}^{(k)} \approx \sum_{j=0}^{Q_k}\sum_{l=0}^{N_k} \pi_j^{(k)} p_l^{(k)} \times \gamma^{(k)}(\min\{j,l\},i), \tag{17}$$

where $\pi_j^{(k)}$, $p_l^{(k)}$, and $\gamma^{(k)}(\min\{j,l\},i)$ are defined as in Section 5, and $a_{\text{suc},i}^{(k)}$ is the probability that i packets are successfully transmitted from queue k in one time frame. Denoting the aggregate traffic arrival rate to queue k as $\overline{A}^{(k)}$, we can approximately calculate the probability vector for data packets of connection c, which is successfully transmitted from queue k by scaling $a_{\text{suc},i}^{(k)}$ with a number representing the ratio between the traffic arrival rate of connection c to queue k and the aggregate traffic arrival rate to queue k as follows:

$$\begin{aligned} a_i^{(k+1,c)} &\approx \frac{\overline{A}^{(k,c)}}{\overline{A}^{(k)}} \times a_{\text{suc},i}^{(k)}, \qquad 1 \le i \le N_k, \\ a_0^{(k+1,c)} &\approx 1 - \sum_{i=1}^{N_k} a_i^{(k+1,c)}, \end{aligned} \tag{18}$$

where N_k is the maximum number of packets transmitted in one time frame from queue k. This approximation can be justified by the fact that successfully transmitted packets from queue k may belong to different input connections. Thus, when successfully transmitted traffic is split into relayed and leaving traffic, the corresponding probability vectors can be calculated by scaling the probability vector of successfully transmitted traffic by using scaling factors proportional to their contributions to the aggregate traffic in terms of the average rate. Similarly, the arrival probabilities for the aggregate relayed traffic to queue $k+1$ can be approximated as

$$\begin{aligned} a_{\text{rel},i}^{(k+1)} &\approx \frac{\sum_{c\in\Phi_{\text{rel}}^{(k)}} \overline{A}^{(k,c)}}{\overline{A}^{(k)}} \times a_{\text{suc},i}^{(k)}, \qquad 1 \le i \le N_k, \\ a_{\text{rel},0}^{(k+1)} &\approx 1 - \sum_{i=1}^{N_k} a_{\text{rel},i}^{(k+1)}. \end{aligned} \tag{19}$$

The arrival probabilities for the aggregate relayed traffic leaving queue k can be approximated as

$$\begin{aligned} a_{\text{lev},i}^{(k)} &\approx \frac{\sum_{c\in\Phi_{\text{lev}}^{(k)}} \overline{A}^{(k,c)}}{\overline{A}^{(k)}} \times a_{\text{suc},i}^{(k)}, \qquad 1 \le i \le N_k, \\ a_{\text{lev},0}^{(k)} &\approx 1 - \sum_{i=1}^{N_k} a_{\text{lev},i}^{(k)}. \end{aligned} \tag{20}$$

The aggregate traffic arrival probability vector to queue $k+1$, therefore, can be calculated as

$$a^{(k+1)} = a_{\text{new}}^{(k+1)} \otimes a_{\text{rel}}^{(k+1)}, \tag{21}$$

where the arrival probability vectors for the new and relayed traffic flows are calculated by (16) and (19), respectively. Let us assume that traffic arriving to queue 1 is from the incoming connection with an arrival probability vector denoted by a_{in} and from the relayed traffic with an arrival probability vector denoted by $a_{\text{rel}}^{(1)}$. Therefore, the aggregate arrival probability vector to queue 1 is $a^{(1)} = a_{\text{in}} \otimes a_{\text{rel}}^{(1)}$.

In summary, the arrival probability vector to queue 1 can be calculated as $a^{(1)} = a_{\text{in}} \otimes a_{\text{rel}}^{(1)}$, and the steady state probability vector for queue 1 (i.e., $\pi^{(1)}$) can be calculated as in Section 5. With the steady state probability vector $\pi^{(1)}$, we can calculate the aggregate arrival probability to queue 2 by using (21). This procedure is repeated until the solution for the last queue of the tandem system is found.

8.2 QoS Routing Algorithm

With class-based queuing, we now show how we can integrate the tandem queuing model presented in the previous section into the QoS routing algorithm. We only discuss the route discovery phase. As before, the incoming connection submits its traffic profile and service class with QoS requirements to the source node. For ongoing connections, we require each node along the routing path to record the aggregate arrival probability vectors to all queues of different service classes in all outgoing links and the current link QoS metrics on these links (i.e., link delay and loss) in its route cache.

The source node initiates the route discovery to find feasible routes to its desired destination as follows: First, the source node checks each outgoing link to see whether it has enough bandwidth to accommodate the incoming connection. The outgoing link with enough bandwidth will be called the BW-feasible link as before. Then, the required amount of bandwidth is allocated to the BW-feasible link, which increases the average service rate of the queue corresponding to the service class of the incoming connection. Note that this queue may be buffering data flows of other ongoing connections. For each BW-feasible link, the source node calculates the updated aggregate arrival probability vector to the corresponding queue as follows:

$$a^{(1)} = a_{\text{in}} \otimes a_{\text{rel}}^{(1)}, \tag{22}$$

where a_{in} and $a_{\text{rel}}^{(1)}$ denote the arrival probability vector of the incoming connection and the aggregate arrival probability vector of ongoing connections being relayed on the considered link, respectively. If there is no ongoing connection on the explored link, we can simply set $a_{\text{rel}}^{(1)} = 1$.

The source node calculates link QoS metrics for each BW-feasible link by using the updated arrival probability vector and allocated bandwidth. Based on the calculated QoS metrics, the RRQ packet is forwarded to receiving nodes via links that satisfy the QoS requirements of the incoming connection and do not degrade the QoS performances of the ongoing connections traversing these links. This guarantees that the QoS requirements of ongoing connections are not violated after admitting the incoming connection into the network. In this case, the source node calculates the aggregate relayed arrival probabilities to the next queue by using (19) and records this arrival probability vector into the RRQ packet header.

Now, each node, upon receiving the RRQ packet, will check the available bandwidth on its outgoing links. For each BW-feasible outgoing link, the node calculates the aggregate arrival probability vector from both the incoming connection and the ongoing connections traversing the explored link by using (21). If the link QoS metrics for the incoming connection are satisfied and those for the ongoing connections traversing the link are not degraded due to the admission of the incoming connection, the RRQ packet will be forwarded to the receiving node of that link after the path QoS metrics and the aggregate relayed arrival probabilities have been recorded into the RRQ packet header. This procedure is repeated until either a feasible route to the destination is found or the incoming connection is blocked. The destination, upon receiving RRQ packet with satisfied path QoS metrics, will send an RRP packet to the source node, as in Section 6. Finally, we need to update the arrival probability vectors to the queues along the routing paths of affected ongoing connections and the QoS metrics of the affected links.

9 Further Discussions and Extensions

9.1 Tandem Queue with Blocking

In this section, we discuss the implementation and solution issues for tandem queues with blocking. In particular, queue $k+1$ $(k > 0)$ in the tandem system may block transmissions from queue k if its buffer is full. In addition, queue k should know how many packets queue $k+1$ can accommodate in each time interval. To avoid buffer overflow, the number of transmitted packets from queue k should be kept smaller than the room available in queue $k+1$.

Although it is possible to implement this blocking option, the communication overhead involved may not be desirable for most wireless applications. In addition, the blocking implementation will result in higher overflow probability in queue 1, which may not be able to block arrivals from the underlying applications. Blocking may also increase the end-to-end delay and even end-to-end loss rate (due to high buffer overflow in queue 1). Therefore, in the context of QoS routing, where only routing paths with satisfied QoS requirements are chosen for end-to-end data transmission, it may be more desirable to avoid the blocking implementation.

For tandem queue with blocking, we can use a decomposition method that is similar in spirit to the method in [32]. The decomposition method is iterative, and it works as follows: In each iteration, we solve all pairs of consecutive queues in the tandem and find steady state probability vectors (i.e., queues 1 and 2, queues 2 and 3, and so on). In addition, the steady state probability vectors of queues $k-1$ and $k+2$ in iteration t are used to solve a pair of queues k and $k+1$ in iteration $t+1$, with blocking being taken into account. The calculation is repeated until a predefined convergence criterion is met.

9.2 Tandem Queuing Models with Batch Markovian Arrival Process

In this section, we show how we can extend the presented queuing model with the Batch Markovian Arrival Process (BMAP). Due to space constraints, we only present the extension for the flow-based queuing model using the decomposition approach. BMAP is a very general arrival process, which can capture correlation (i.e., burstiness) in the arrival traffic [34]. Specifically, BMAP can be represented by $M+1$ matrices $C_v\ (v=0,1,2,\cdots,M)$ of order $S\times S$, whose elements $C_v(i,j)$ represent transition from phase i to phase j, with v arriving packets. Letting $C=\sum_{v=0}^{M}C_v$, then C is a stochastic matrix whose elements $C(i,j)$ denote transition from phase i to phase j. Now, let u be a row vector of dimension S, which satisfies $u=u.C$ and $u.\mathbf{1}=1$, where $\mathbf{1}$ is a column vector of all 1s with dimension S. Then, the arrival rate of this BMAP source is $\overline{A}=u\sum_{v=1}^{M}vC_v\mathbf{1}$ [34]. We refer the readers to [34] and the references therein for more details about BMAP. Note that the batch Bernoulli arrival process is a special case of BMAP, where C_v degenerates into a scalar, which is the probability of having v arrivals in one time frame.

Now, we show how we can extend the flow-based queuing model by using the decomposition approach with BMAP. As before, we first find the queue length distribution for queue 1, use it to calculate the arrival probabilities to queue 2, and so on. The queuing model for queue $k\ (k\geq 2)$ remains the same as in Section 5. We show how we can solve the queue length distribution for queue 1 with BMAP. Let $r(t)$ denote the arrival phase of the arrival process and $q_1(t)$ denote the number of packets in queue 1 in time frame t. Then, the random process $Y_1(t)=\{q_1(t),r(t)\}$, $(0\leq q_1(t)\leq Q_1$, $1\leq r(t)\leq S)$ forms a discrete-time MC. Let (x,r) denote the generic system state, and we write the probabilities corresponding to transitions $(x_1,*)\rightarrow(x_2,*)$ into the matrix block $\mathbf{E}_{x_1,x_2}$, whose element $\mathbf{E}_{x_1,x_2}(r_1,r_2)$ denotes the probability of transition $(x_1,r_1)\rightarrow(x_2,r_2)$.

Similar to the derivation in Appendix B, consider transitions $(x_1,*)\rightarrow(x_2,*)$, let s be the number of packets arriving at queue 1, and the transmission capacity of link 1 is l packets during the considered time frame. Let us assume that i packets among $\min\{x_1,l\}$ transmitted packets are correctly received. Then, we have $x_2=\min\{x_1+s,Q_k\}-i$. Thus, we can calculate $\mathbf{E}_{x_1,x_2}$ as follows:

$$\mathbf{E}_{x_1,x_2}=\sum_{s}\sum_{l}\sum_{i}C_s\times p_l^{(1)}\times\gamma^{(1)}(\min\{x_1,l\},i), \tag{23}$$

where all combinations of s, l, and i, for which $x_2=\min\{x_1+s,Q_1\}-i$, are included in the sum.

Given the transition probabilities, we can easily calculate the steady state probability vector of this MC, which is denoted as $\pi^{(1)}=[\pi_0^{(1)},\pi_1^{(1)},\cdots,\pi_{Q_1}^{(1)}]$, where $\pi_i^{(1)}$ can be further expanded as $\pi_i^{(1)}=[\pi_{i,1}^{(1)},\pi_{i,2}^{(1)},\cdots,\pi_{i,S}^{(1)}]$. Here, $\pi_{i,j}^{(1)}$ denotes the probability that there are i packets in queue 1 and the arrival process in phase j. The probability that i packets are successfully transmitted from queue 1 and arrive at queue 2 can be approximated as

$$a_i^{(2)}\approx\sum_{j=0}^{Q_1}\sum_{r=1}^{S}\sum_{l=0}^{N_1}\pi_{j,r}^{(1)}p_l^{(1)}\times\gamma^{(1)}(\min\{j,l\},i). \tag{24}$$

Here, compared to (7) for the case of the batch Bernoulli arrival process, we have to sum up the probabilities $\pi_{j,r}^{(1)}$ over all arrival phases. These arrival probabilities will be used to calculate the queuing solution for queue 2 in the decomposition approach.

9.3 Wireless Network with Random Access

In this section, we show how we can extend the proposed queuing models to wireless networks with random access MAC using the CSMA/CA protocol [38]. We note that, for the QoS routing application, static MAC using CDMA, TDMA, FDMA, or their combinations may be more appropriate, because bandwidth allocation can be performed such that bandwidth requirements for connection requests are satisfied [8], [10]. The presented extension here is useful for other application contexts such as end-to-end performance analysis.

Now, suppose that there is one channel that is shared by N_a links on a contention basis in a common neighborhood. As our first step, we show how we can modify the transmission probability in (1) when the CSMA/CA protocol is used. We assume that all involved wireless links compete for the channel at the beginning of each time frame, and the one winning the competition will transmit in the remaining part of the time frame. Consider a particular link l of the considered connection. For notational convenience, let $\varphi^{(l)}(k)$ be the probability that the channel is in state k at link l. Let p_s be the probability that link l wins the channel access in any time frame. Then, we have

$$p_i^{(l)}=\begin{cases}p_s\times\varphi^{(l)}(k), & i=c_k,\\ (1-p_s)\times\varphi^{(l)}(k), & i=0,\\ 0, & \text{otherwise},\end{cases} \tag{25}$$

where, as before, $p_i^{(l)}$ is the probability that i packets are transmitted on link l in any time frame. Using $p_i^{(l)}$ derived in (25), the presented queuing models can be used to calculate all performance measures of interest.

Now, we show how we can calculate p_s for the CSMA/CA protocol. Let W denote the contention window of the CSMA/CA protocol. If no exponential backoff is employed, the probability that link l transmits in any time frame can be calculated as [38]

$$\tau=\frac{2}{W+1}. \tag{26}$$

When exponential backoff is used, a more complicated procedure can be used to calculate τ. We refer the interested readers to [38] for more details. Now, if link l competes with other (N_a-1) links in its neighborhood, the probability that it wins the access can be calculated as

$$p_s=\tau(1-\tau)^{N_s-1}. \tag{27}$$

Hence, all performance measures of interest can be calculated for wireless systems using the CSMA/CA protocol.

10 Conclusions

We have proposed both exact and approximated decomposition approaches to solve a general tandem queue system. The proposed tandem queuing models capture realistic physical and link layer designs, where the multirate transmission feature due to adaptive modulation and coding in the physical layer and the ARQ-based error recovery in the link layer have been incorporated. The proposed decomposition approach achieves very accurate queuing performance measures with much lower computational complexity, as compared to the exact approach. Using the decomposition queuing approach, we have developed a unified queuing and QoS routing framework, which is able to satisfy the QoS requirements in multihop wireless networks. The numerical results have shown that the framework works efficiently for finding feasible routing paths in the network if they exist. The extension of the framework to wireless networks with class-based queuing implementation has also been presented.

Appendix A

Derivations of Transition Probabilities for MC $X(t)$

We derive the transition probabilities for MC $X(t)$ in Section 4. Before deriving $\Pr\{(x_1, y_1) \to (x_2, y_2)\}$, let us define $\gamma^{(l)}(j, i)$ as the probability that i packets are correctly received, given that j packets were transmitted over link l. Assuming that packet errors are independent, we can calculate $\gamma^{(l)}(j, i)$ as follows:

$$\gamma^{(l)}(j,i) = \binom{j}{i} \left(\beta^{(l)}\right)^{j-i} \left(1-\beta^{(l)}\right)^{i}, \tag{28}$$

where $\beta^{(l)}$ is the probability of transmission error on link l, as defined in Section 3.2.

Let s be the number of packets arriving at queue 1 in a particular time frame, and the transmission capability on link 1 is k packets. We need to find the conditions under which a general transition $(x_1, y_1) \to (x_2, y_2)$ occurs. The number of packets in queue 1 after accepting newly arriving packets is $\min(x_1 + s, Q_1)$, and the number of packets transmitted on link 1 is $\min(x_1, k)$. Assuming that, among these transmitted packets, i packets are correctly received at the receiving end (i.e., these i successfully transmitted packets will enter queue 2), we have $x_2 = \min(x_1 + s, Q_1) - i$. Note that, due to the employment of an infinite-persistent ARQ protocol in the link layer, all the erroneous packets will stay in the buffer for retransmission until they are successfully transmitted. Similarly, assuming that the transmission capability of the second link is l packets, and j packets among $\min(y_1, l)$ transmitted packets are correctly received at the receiving end of link 2, we have $y_2 = \min(y_1 + i, Q_2) - j$. Hence, we can calculate $\Pr\{(x_1, y_1) \to (x_2, y_2)\}$ as

$$\begin{aligned}\Pr\{(x_1,y_1)\to(x_2,y_2)\} = &\sum_s \sum_k \sum_l \sum_i \sum_j a_s p_k^{(1)} p_l^{(2)} \\ &\times \gamma^{(1)}(\min\{x_1,k\}, i) \times \gamma^{(2)}(\min\{y_1,l\}, j),\end{aligned} \tag{29}$$

where all possible cases such that $x_2 = \min(x_1 + s, Q_1) - i$ and $y_2 = \min(y_1 + i, Q_2) - j$ are included in the sum.

Appendix B

Derivations of Transition Probabilities for MC $X_k(t)$

We derive the transition probabilities for MC $X_k(t)$ for a particular queue k of the tandem system defined in Section 5. Let us consider a general transition probability $\Pr\{x_1 \to x_2\}$. Let s be the number of packets arriving at queue k, and the transmission capacity of the wireless link is l packets during the considered time frame. Let us also assume that i packets among $\min\{x_1, l\}$ number of transmitted packets are correctly received. Then, we have $x_2 = \min\{x_1 + s, Q_k\} - i$. Thus, the transition probability $\Pr\{x_1 \to x_2\}$ can be written as

$$\Pr\{x_1 \to x_2\} \approx \sum_s \sum_l \sum_i a_s^{(k)} p_l^{(k)} \times \gamma^{(k)}(\min\{x_1, l\}, i), \tag{30}$$

where all combinations of s, l, and i, for which $x_2 = \min\{x_1 + s, Q_k\} - i$, are included in the sum. This transition probability is exact only for queue 1, whereas it is approximated for queue $k \geq 2$ since the arrival probabilities are approximated.

References

[1] I.F. Akyildiz and X. Wang, "A Survey on Wireless Mesh Networks," *IEEE Comm. Magazine*, vol. 43, no. 9, pp. 23-30, Sept. 2005.

[2] L.B. Le and E. Hossain, "Multihop Cellular Networks: Potential Gains, Research Challenges, and a Resource Allocation Framework," *IEEE Comm. Magazine*, vol. 45, no. 9, pp. 66-73, Sept. 2007.

[3] L.B. Le and E. Hossain, "Cross-Layer Optimization Frameworks for Multihop Wireless Networks Using Cooperative Diversity," *IEEE Trans. Wireless Comm.*, to appear.

[4] L. Le and E. Hossain, "An Analytical Model for ARQ Cooperative Diversity in Multihop Wireless Networks," *IEEE Trans. Wireless Comm.*, vol. 7, no. 5, pp. 1786-1791, May 2008.

[5] L.B. Le, E. Hossain, and M. Zorzi, "Queuing Analysis for GBN and SR ARQ Protocols Under Dynamic Radio Link Adaptation with Non-Zero Feedback Delay," *IEEE Trans. Wireless Comm.*, vol. 6, no. 9, pp. 3418-3428, Sept. 2007.

[6] L.B. Le, E. Hossain, and A.S. Alfa, "Service Differentiation in Multi-Rate Wireless Networks with Weighted Round-Robin Scheduling and ARQ-Based Error Control," *IEEE Trans. Comm.*, vol. 54, no. 2, pp. 208-215, Feb. 2006.

[7] L.B. Le, E. Hossain, and A.S. Alfa, "Delay Statistics and Throughput Performance for Multirate Wireless Networks under Multiuser Diversity," *IEEE Trans. Wireless Comm.*, vol. 5, no. 11, pp. 3234-3243, Nov. 2006.

[8] B. Zhang and H.T. Mouftah, "QoS Routing for Wireless Ad Hoc Networks: Problems, Algorithms, and Protocols," *IEEE Comm. Magazine*, vol. 43, no. 10, pp. 110-117, Oct. 2005.

[9] E. Royer and C.-K. Toh, "A Review of Current Routing Protocols for Ad Hoc Mobile Wireless Networks," *IEEE Personal Comm.*, pp. 46-55, Apr. 1999.

[10] C.R. Lin and J.-S. Liu, "QoS Routing in Ad Hoc Wireless Networks," *IEEE J. Selected Areas in Comm.*, vol. 17, no. 8, pp. 1426-1438, Aug. 1999.

[11] C. Zhu and M. Scott, "QoS Routing for Mobile Ad Hoc Networks," *Proc. IEEE INFOCOM '01*, 2001.

[12] S. Chen and K. Nahrstedt, "Distributed Quality-of-Service Routing in Ad-Hoc Networks," *IEEE J. Selected Areas in Comm.*, vol. 17, no. 8, pp. 1488-1505, Aug. 1999.

[13] M.S. Alouini and A.J. Goldsmith, "Adaptive Modulation over Nakagami Fading Channels," *Kluwer J. Wireless Comm.*, vol. 13, nos. 1-2, pp. 119-143, May 2000.

[14] Q. Liu, S. Zhou, and G.B. Giannakis, "Cross-Layer Combining of Adaptive Modulation and Coding with Truncated ARQ over Wireless Links," *IEEE Trans. Wireless Comm.*, vol. 3, no. 5, pp. 1746-1755, Sept. 2004.

[15] G. Kulkarni, S. Adlakha, and M. Srivastava, "Subcarrier Allocation and Bit-Loading Algorithms for OFDMA-Based Wireless Networks," *IEEE Trans. Mobile Computing,* vol. 4, no. 6, pp. 652-662, Nov./Dec. 2005.
[16] J.-H. Song, V.W.S. Wong, and V.C.M. Leung, "Efficient On-Demand Routing for Mobile Ad Hoc Wireless Access Networks," *IEEE J. Selected Areas in Comm.,* vol. 22, no. 7, pp. 1374-1383, Sept. 2004.
[17] A. Iwata et al., "Scalable Routing Strategies for Ad Hoc Wireless Networks," *IEEE J. Selected Areas in Comm.,* vol. 17, no. 8, pp. 1369-1379, Aug. 1999.
[18] D.S.J.D. Couto, D. Aguayo, J. Bicket, and R. Morris, "A High-Throughput Path Metric for Multihop Wireless Routing," *Proc. ACM MobiCom 2003,* 2003.
[19] D. Kim, C.-H. Min, and S. Kim, "On-Demand SIR and Bandwidth-Guaranteed Routing with Transmit Power Assignment in Ad Hoc Mobile Networks," *IEEE Trans. Vehicular Technology,* vol. 22, no. 7, pp. 1301-1321, Sept. 2004.
[20] C.E. Perkins and P. Bhagwat, "Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for Mobile Computers," *Proc. ACM SIGCOMM '94,* pp. 234-244, 1994.
[21] D. Johnson and D. Maltz, "Dynamic Source Routing in Ad Hoc Wireless Networks," *Mobile Computing,* E. Imielinski and H. Korth, eds. Kluwer Academic Publishers, 1996.
[22] V.D. Park and M.S. Corson, *Temporally Ordered Routing Algorithms (TORA) Version 1 Functional Specification,* IETF Internet draft, work in progress, July 2001.
[23] C. Perkins, E. Belding-Royer, and S. Das, *Ad Hoc On-Demand Distance Vector (AODV) Routing,* IETF RFC 3561, July 2003.
[24] P. Pham and S. Perreau, "Performance Analysis of Reactive Shortest Path and Multipath Routing Mechanism with Load Balance," *Proc. IEEE INFOCOM '03,* 2003.
[25] A. Tsirigos and Z.J. Haas, "Analysis of Multipath Routing—Part I: The Effect on the Packet Delivery Ratio," *IEEE Trans. Wireless Comm.,* vol. 3, no. 1, pp. 138-146, Jan. 2004.
[26] X. Lin and N.B. Shroff, "An Optimization-Based Approach for QoS Routing in High-Bandwidth Networks," *Proc. IEEE INFOCOM '04,* Mar. 2004.
[27] C.-K. Toh, M. Delwar, and D. Allen, "Evaluating the Communication Performance of an Ad Hoc Wireless Network," *IEEE Trans. Wireless Comm.,* vol. 1, no. 3, pp. 402-414, July 2002.
[28] J.A. Morrison, "Two Discrete-Time Queues in Tandem," *IEEE Trans. Comm.,* vol. 27, no. 3, pp. 563-573, Mar. 1979.
[29] M. Xie and M. Haenggi, "Delay Performance of Different MAC Schemes for Multihop Wireless Networks," *Proc. IEEE Global Telecomm. Conf. (GLOBECOM '05),* Dec. 2005.
[30] F.P. Kelly, "The Throughput of a Series of Buffers," *Advances in Applied Probability,* vol. 14, pp. 633-653, 1982.
[31] V. Anantharam and P. Tsoucas, "Stochastic Concavity of Throughput in Series of Queues with Finite Buffers," *Advances in Applied Probability,* vol. 22, pp. 761-763, 1990.
[32] A. Brandwajn and Y.L.L. Jow, "An Approximation Method for Tandem Queues with Blocking," *J. Operational Research Soc.,* vol. 36, no. 1, pp. 73-83, Jan./Feb. 1988.
[33] A. Burchard, J. Liebeherr, and S.D. Patek, "A Min-Plus Calculus for End-to-End Statistical Service Guarantees," *IEEE Trans. Information Theory,* vol. 52, no. 9, pp. 4105-4114, Sept. 2006.
[34] A.S. Alfa, "Algorithmic Analysis of the BMAP/D/k System in Discrete Time," *Advances in Applied Probability,* vol. 35, pp. 1131-1152, 2003.
[35] L. Hu, "Distributed Code Assignments for CDMA Packet Radio Networks," *IEEE/ACM Trans. Networking,* vol. 1, pp. 668-677, Dec. 1993.
[36] P. Kyasanur and N.H. Vaidya, "Routing and Interface Assignment in Multi-Channel Multi-Interface Wireless Networks," *Proc. IEEE Wireless Comm. and Networking Conf. (WCNC '05),* Mar. 2005.
[37] H.E. Inamura, R. Ludwig, A. Gurtov, and F. Khafizov, *TCP over Second (2.5G) and Third (3G) Generation Wireless Networks,* IETF RFC 3481, Feb. 2003.
[38] G. Bianchi, "Performance Analysis of the IEEE 802.11 Distributed Coordinated Function," *IEEE J. Selected Areas in Comm.,* vol. 18, no. 3, pp. 535-547, Mar. 2000.

Long Le received the BEng degree (with highest distinction) from Ho Chi Minh City University of Technology in 1999, the MEng degree from the Asian Institute of Technology (AIT) in 2002, and the PhD degree from the University of Manitoba in 2007. He is currently a postdoctoral fellow in the Department of Electrical and Computer Engineering, University of Waterloo. His research interests include cognitive radio, network coding, link and transport layer protocol issues, cooperative diversity and relay networks, stochastic control, and cross-layer design for communication networks. He is a member of the IEEE. He received a University Gold Medal from Ho Chi Minh City University of Technology, the Keikyu Scholarship, a University of Manitoba Graduate Fellowship, the Edward R. Toporeck Graduate Fellowship in Engineering, the University of Manitoba Students' Union Scholarship, and the IEEE Student Travel Awards for the 2003 IEEE Wireless Communications and Networking Conference (WCNC) and the 2005 IEEE International Conference on Communications (ICC).

Ekram Hossain received the PhD degree in electrical engineering from the University of Victoria, Canada, in 2000. He is currently an associate professor in the Department of Electrical and Computer Engineering at the University of Manitoba, Winnipeg, Canada. Dr. Hossain's current research interests include the design, analysis, and optimization of wireless communication networks and cognitive radio systems. He was a coeditor of the books *Cognitive Wireless Communication Networks* (Springer, 2007, ISBN: 978-0-387-68830-5), *Wireless Mesh Networks: Architectures and Protocols* (Springer, 2007, ISBN: 978-0-387-68839-8), and *Heterogeneous Wireless Access Networks* (Springer, 2008, ISBN: 978-0-387-09776-3), and a coauthor of the book *Introduction to Network Simulator NS2* (Springer, 2008, ISBN: 978-0-387-71759-3). Dr. Hossain serves as an editor for the *IEEE Transactions on Mobile Computing*, the *IEEE Transactions on Wireless Communications*, the *IEEE Transactions on Vehicular Technology*, *IEEE Wireless Communications*, and several other international journals. He served as a guest editor for special issues of *IEEE Communications Magazine* (cross-layer protocol engineering for wireless mobile networks) and *IEEE Wireless Communications* (radio resource management and protocol engineering for IEEE 802.16). He served as a technical program cochair for IEEE Globecom '07 and IEEE WCNC '08. Dr. Hossain served as the technical program chair for workshops on "Cognitive Wireless Networks" (CWNets '07) and "Wireless Networking for Intelligent Transportation Systems" (WiN-ITS '07) held in conjunction with QShine '07: the International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness. He served as the technical program cochair for the Symposium on "Next Generation Mobile Networks" (NGMN '06), NGMN '07, and NGMN '08 held in conjunction with the ACM International Wireless Communications and Mobile Computing Conference (IWCMC '06), IWCMC '07, and IWCMC '08, and the First IEEE International Workshop on Cognitive Radio and Networks (CRNETS '08) in conjunction with the IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '08). He is a senior member of IEEE. Dr. Hossain is a registered Professional Engineer in the province of Manitoba, Canada.