This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE GLOBECOM 2005 proceedings.

Delay Analysis of 802.16 based Last Mile Wireless Networks

Rajagopal Iyengar
RPI,Troy, NY
Email: iyengr@rpi.edu

Prakash Iyer
Intel, Hillsboro, OR
Email:prakash.iyer@intel.com

Biplab Sikdar
RPI,Troy, NY
Email: sikdab@rpi.edu

***Abstract*— In this paper, we present a delay analysis of IEEE 802.16 based broadband wireless access networks, a promising technology capable of supporting both fixed and fully mobile operations while offering integrated voice, video and data services. Our work develops analytic models to evaluate the performance of IEEE 802.16 based networks in terms of their latencies as a function of various scheduling policies. The model also allows us to explore the impact of various system parameters like the subframe lengths on the performance and thereby aid in system design. The results of our model can also be used for providing probabilistic quality of service guarantees and determining the number of nodes that can be accommodated while satisfying a given delay constraint. The analytic models are verified using extensive simulations.**

I. Introduction

The IEEE 802.16 is an emerging suite of standards for point to multipoint (PMP) broadband wireless access (BWA). The 802.16e amendment to the 802.16-2004 specification enables support for combined fixed and mobile operation for licensed and license-exempt frequencies below 11 GHz. IEEE 802.16 is likely to emerge as a preeminent technology for cost-competitive ubiquitous broadband wireless access supporting fixed, nomadic, portable and fully mobile operations offering integrated voice, video and data services. The technology is likely to be considered in a variety of deployment scenarios, such as standalone IP core based networks (including NGN) and as a data overlay for PS services over existing broadband and cellular networks. Initial deployments are likely to be based on fixed/nomadic operation with fully mobile usage to follow. Of the three different PHYs specified in the standard, OFDM multi-access (OFDMA) is likely to emerge as the most preferred PHY supporting all usage models. The 802.16 standard supports four different scheduling service classes for QoS and the MAC supports an Request-Grant mechanism for uplink transmissions from a Subscriber station (SS) to its Service Base station (BS). In order to send data in the next frame, the SS reserves bandwidth for it in the current frame. Though a number of techniques for Bandwidth Allocation and QoS are provided, detailed scheduling methods, and reservation techniques are not standardised, and provide a method for vendors to differentiate their equipment. Interactions between the MAC and the PHY enable modulation and coding schemes in a burst profile adaptively per Subsciber Station (SS) - consequently depending on the condition of the link, different burst profiles may be used. A key contributor to the performance and capacity utilization of the air link is the delay due to contention and/or polling schemes employed for uplink channel allocation. Minimizing this delay across a set of fixed or mobile SS is critical to meeting QoS goals for delay-sensitive applications. In this paper, we analyze three different polling schemes and present comparative results based on both our analysis as well as simulations. The analytical models derive closed form expressions for the queueing delays at each SS as a function of systems parameters. The developed models can be used for purposes like determining optimal frame lengths , admission control, and determining the number of connections that can be supported for a given delay constraint. To the best of our knowledge, there is no existing literature addressing the delay related performance of 802.16 networks. The rest of the paper is organized as follows. In Section II we present a brief overview of the IEEE 802.16 MAC protocol. In Section III we present the analytic model and simulation results for the case when the nodes are polled at the end of the uplink subframe while IV addresses the case of polling at the beginning of the uplink subframe. We address the case of piggybacked polls in Section V while Section VI presents the concluding remarks.

II. Overview of the 802.16 MAC

The frame structure in the 802.16 standard is divided into two main regions, for Downlink and Uplink respectively. The overall frame structure of the 802.16 MAC is shown in Figure 1 and shows the operation of 802.16 MAC in the Time Division Duplex (TDD) mode. The Downlink Map (DL-Map) and the Uplink Map (UL-Map) contain information pertinent to the channel allocation for various nodes (i.e. SSs) in the coming Uplink and Downlink subframes respectively and are broadcast to all nodes. Each SS receives and decodes the DL-Map and looks for information indicating data directed to that SS in the DL subframe. In Figure 1, the first portion of the Uplink Subframe is reserved for the BS to poll the SSs. The BS defines the duration when a given SS can request bandwidth in the UL-Map. In it's allocated duration, an SS transmits a *Bandwidth Request (BR)* MAC header to the BS. Based on the scheduling algorithm used, the BS allocates chunks of the uplink subframe to different SSs. This information becomes available when the SSs decode the UL-Map. In this paper we assume that stations are polled sequentially in some portion of the Uplink Subframe. Note that the BS can grant the bandwidth request made by a SS in the current frame only in the next frame. Thus in it's assigned slot, a packet from a SS is transmitted only if the queue is not empty and bandwidth has been reserved for a packet from this

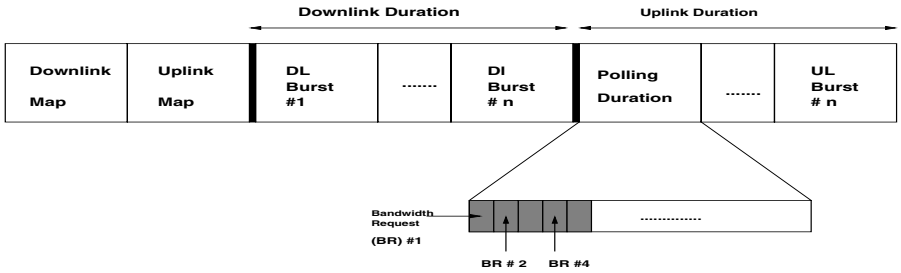

Fig. 1. Frame Structure of 802.16 MAC

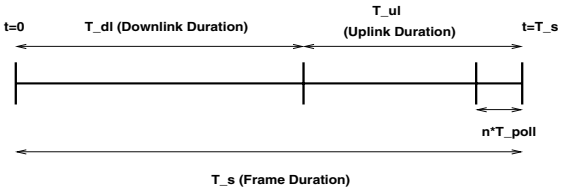

Fig. 2. Polling at end of Uplink: Frame Schematic

SS in the previous frame. This paper focuses on finding average delay for a packet in an 802.16 system under these conditions. For simplicity it is assumed that all packets are of the same length.

III. Delay Analysis: Polling at End of Uplink Subframe

In this section we develop an analytic model for evaluating the average delay for a packet when the following polling scheme is used: nodes are polled sequentially at the end of every Uplink subframe. If a node has a non-empty queue then it is allocated time sufficient to transmit one packet in the next Uplink subframe. The frame structure when nodes are polled at the end of an Uplink subframe is shown in Figure 2. We assume that an arbitrary number of SS nodes n exist in the system. The packet interarrival times at the queue of any node are assumed to exponentially distributed with rate λ. We denote the duration of the Uplink subframe by T_{ul} and Downlink subframe by T_{dl}. The time it takes for a station to send a Bandwidth Request packet on the Uplink is T_{poll}. and time taken to transmit a packet from the head of any queue when the service begins by L. We now calculate the delay experienced by a packet arriving at the ith node, $1 \leq i \leq n$. In any given uplink subframe, each of the n queues can exist in one of three states:

- S0: The queue is empty. Let the *utilisation* of a given node be ρ. The probability that an arbitrary arrival sees an empty queue is $(1-\rho)$. Hence the probability of an arbitrary arrival seeing a full queue is ρ.
- S1: The queue has packets in it, but bandwidth has not been reserved for the Head of Queue (HOQ). The polling interval in this frame isused to reserve bandwidth for the HOQ in the *next* frame.
- S2: The queue has packets in it, and bandwidth has been reserved in the previous subframe for the current HOQ. The queue transmits a packet in the current uplink subframe and uses the polling interval to reserve bandwidth for the next packet.

Given that the queue is non-empty, let the probability of it being in state S1 be p. Then, the conditional probability of the queue being in state S2 is $(1-p)$. Also, since each SS gets a chance to transmit a packet in each frame, it is easy to see that $\rho = \lambda(T_{ul} + T_{dl})$. All the analysis that follows is with respect to a "tagged arrival" at the ith station, $1 \leq i \leq n$. In any given Uplink subframe, the delay experienced by a packet at the head of the ith queue depends on the number of stations amongst those polled before it that have packets to send and are in state S2. If there are j such stations, then j is Binomially distributed with parameters $i-1$ and $\rho(1-p)$, where $\rho(1-p)$ is the probability of a queue existing in state S2.

A. Arrival at an Empty Queue

Here we consider two subcases for the tagged arrival: when it arrives before and after the ith SS has been polled. The case C1 where an arrival occurs before the poll is illustrated in Figure 3. C2 refers to the case where arrival occurs after the poll. For case C1 (Figure 3) we can write the expected delay as:

$$E[D_{S0}|C1] = \frac{T_s + (n-i+1)T_{poll}}{2} + T_{dl} + (i-1)\rho(1-p)L \quad (1)$$

The details of how the expression is derived can be understood from Figure 3. The first term of the delay expression accounts for the average delay incurred by the arriving packet till the frame in which it arrived finishes. Given that the packet arrives before the node is polled, this arrival can occur uniformly in the interval $[0, T_s-(n-i+1)T_{poll}]$. This can be seen as follows: It is well known that with exponential arrivals in a slotted departure system (for example a classical M/D/1 queue), an arrival is equally likely to occur anywhere in a slot or frame [4], [5]. In our case, *given* that an arrival occurs in a given frame, the arrival is thus uniformly distributed over $[0, T_s]$, relative to the start of the frame. If t denotes the arrival time (measured from start of frame) and we are given that $t < T < T_s$, for some T, we then need the distribution of $T-t$. Note that $P[t \leq T] = \frac{T}{T_s}$. The probability distribution function of t given that it occurs before T is given by:

$$P[t \leq \tau | t \leq T] = \frac{P[t \leq \tau, t \leq T]}{P[t \leq T]} = \frac{P[t \leq \tau]}{P[t \leq T]} = \frac{\tau}{T}$$

which is a uniform distribution in the range $[0, T]$. Using $T = T_s - (n-i+1)T_{poll}$, we get the desired result. Hence

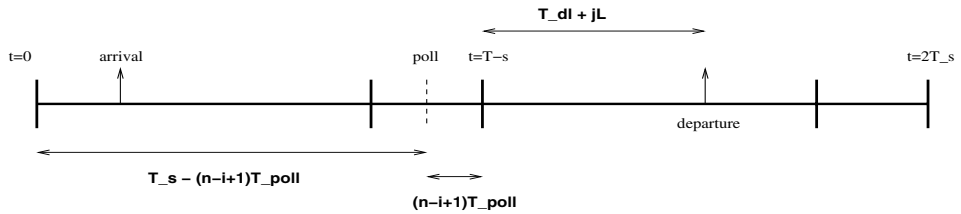

Fig. 3. Packet arrival at Empty queue before Poll

on an average, the arrival must wait $\frac{T_s-(n-i+1)T_{poll}}{2}$ before the node is polled. In addition it must wait for $(n-i+1)T_{poll}$ time after the node is polled, till the end of the current frame. In the next frame, this packet is served. Here, it must wait for the duration of the Downlink subframe T_{dl} and the delay due to the packets from other nodes which transmit before the node with the tagged arrival. Assume there are j such packets served as shown in figure 3. Since j is binomially distributed, as discussed earlier, we can write the average delay due to this effect as $(i-1)\rho(1-p)$. The total average delay is sum of these constant and computed average delays. For case C2 we can write the expected delay in the same way as discussed for case C1, except that an extra frame time T_s must be accounted for, since the packet arrives after the node has been polled in the current frame and must wait till the next polling instant. Given that the packet arrived after the node has been polled (and before the end of the frame), it can arrive uniformly in a duration of $(n-i+1)T_{poll}$, using arguments similar to the one use for case C1, causing an average delay of $\frac{(n-i+1)T_{poll}}{2}$. Thus the total delay in this case is given by:

$$E[D_{S0}|C2] = \frac{(n-i+1)T_{poll}}{2} + T_s + T_{dl} + (i-1)\rho(1-p)L \tag{2}$$

The probability of the two cases C1 and C2 is given by:

$$\begin{aligned} P_{C1} &= 1 - \frac{(n-i+1)T_{poll}}{T_s} \\ P_{C2} &= 1 - P_{C1} \end{aligned} \tag{3}$$

These probabilities can be easily derived by noting the respective time durations in the frame where these cases may occur. Case C2 occurs when the packet arrives during or after the node has been polled, and before the end of the current frame. This corresponds to a duration of $(n-i+1)T_{poll}$ seconds from a total frame length T_s. An arrival in the remainder of the frame falls under case C1. Hence we can write the overall expected delay in the case of an arrival at an empty queue as:

$$E[D_{S0}] = E[D_{S0}|C1]P_{C1} + E[D_{S0}|C2]P_{C2} \tag{4}$$

B. Arrival at a Non-Empty Queue

In case of an arrival at a non empty queue, we identify 2 cases: (1) the queue is in state S1 and (2) the queue is in state S2. We now address these two cases.

1) Queue is in state S1: Here, again we identify two subcases, if the tagged arrival occurs before or after the ith node is polled, at the end of a given frame. Again we refer to these cases as C1 and C2. Let N_q denote the number of packets in the queue seen by the tagged arrival. The tagged arrival must wait till these N_q packets are served. We make the simplifying assumption that the tagged arrival sees the long term average of the number of packets in the queue when it arrives and this is denoted by N_q. The important difference in the case when the arrival at the queue sees it to be occupied is that it must wait till the packets ahead of it in the queue are served. For case C1, we can write expected delay as:

$$\begin{aligned} E[D_{S1}|C1] = \frac{T_s + (n-i+1)T_{poll}}{2} + T_{dl} \\ +(i-1)\rho(1-p)L + N_qT_s \end{aligned} \tag{5}$$

For the case C2, we can write:

$$\begin{aligned} E[D_{S1}|C2] = \frac{(n-i+1)T_{poll}}{2} + T_s + T_{dl} \\ +(i-1)\rho(1-p)L + (N_q-1)T_s \end{aligned} \tag{6}$$

Note that the derivations of the expressions above follow arguments similar to those Section III-A. The expressions for the probabilities of the two cases C1 and C2 remain the same as earlier (Equation (3)). We can then write the expected delay for state S1 as:

$$\begin{aligned} E[D_{S1}] &= E[D_{S1}|C1]P_{C1} + E[D_{S1}|C2]P_{C2} \\ &= E[D^*_{S1}] + N_qT_s \end{aligned} \tag{7}$$

where $E[D^*_{S1}]$ is used to represent all terms without N_qT_s in the expression for $E[D_{S1}]$ once the values for $E[D_{S1}|C1]$, $E[D_{S1}|C2]$, P_{C1} and P_{C2} have been substituted.

2) Queue is in state S2: As in states S0 and S1, we consider the same two cases here, C1 and C2. In state S2, the HOQ has already been polled but not yet transmitted when the tagged arrival occurs. We can write delay expressions for each case as:

$$\begin{aligned} E[D_{S2}|C1] = \frac{(n-i+1)T_{poll}}{2} + T_s + T_{dl} \\ +(i-1)\rho(1-p)L + N_qT_s \end{aligned} \tag{8}$$

$$\begin{aligned} E[D_{S2}|C2] = T_{dl} + (i-1)\rho(1-p)L - \frac{T_s}{2} \\ + \frac{(n-i+1)T_{poll}}{2} + N_qT_s \end{aligned} \tag{9}$$

As in the case of state S1, we can write:

$$E[D_{S2}] = E[D^*_{S2}] + N_qT_s \tag{10}$$

C. Overall Delay

We denote the overall delay by D. We can calculate the expected overall delay $E[D]$, by unconditioning on the probabilities of the queue being in state S0, S1 or S2 respectively. Also,

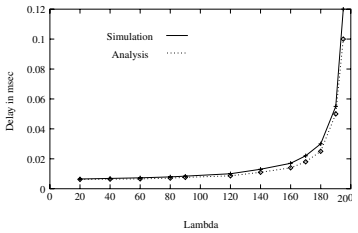

Fig. 4. Average delay, $n = 10$, $i = 5$

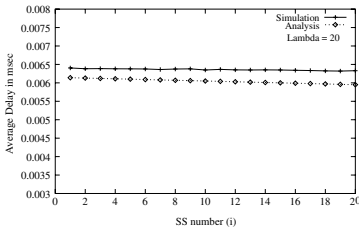

Fig. 5. Average delay across all nodes

from Little's theorem, $N_q = \lambda E[D]$. Now we can write the overall delay as:

$$E[D] = \frac{E[D_{S0}](1-\rho) + E[D^*_{S1}]p\rho + E[D^*_{S2}](1-p)\rho}{1 - \lambda T_s \rho} \tag{11}$$

D. Evaluating p

The quantity p is the probability that the queue is in state S1 given that the queue is not empty. We present an approximate analysis for p and the approximation is accurate for low loads but loses accuracy for higher loads. The approximations are necessary since an exact closed form for evaluating p requires knowledge of the queue state distributions, which is hard to find. We assume that the queue changes state to S1 when the tagged arrival occurs at a queue with one other packet ahead of it. Since we have exponential arrivals in a slotted system, the instance of the tagged arrival relative to the start of the frame is uniformly distributed in the duration of a frame. We further assume that the packet at the head of the queue arrives just after the start of the frame, and that polling intervals are very small compared to the frame time T_s so that $nT_{poll} << T_s$. Since we assume no other arrivals between the packet at the head of the queue and the tagged arrival, the tagged arrival must occur within a frame duration to ensure that no bandwidth was reserved for the packet at the head of the queue. Thus we have $p = \int_0^{T_s} e^{-\lambda t} \frac{1}{T_s} dt$, where t is the time when the tagged arrival occurs at the queue, measured from the start of the frame. Hence, $p = \frac{1-e^{-\lambda T_s}}{\lambda T_s}$. This expression is also used in the case where polling occurs at the start of the frame, since the duration in which the tagged arrival can occur and see the HoQ present but not polled remains T_s.

E. Simulation Results: Polling at End of Uplink Subframe

In order to simulate the 802.16 operation, we implemented a MAC module in NS-2 with 802.16 functionality. The frame duration is 5msec, with a Downlink to Uplink duration ratio of 2:1. The data rate is 50Mbps, and the polling duration is 0.01msec. In the simulation topology used in all simulations, each wireless node (SS) is associated with a single 802.16 base station (BS). We carried out a number of simulations with varying number of wireless nodes. In Figure 4, we compare the results from simulation and analysis for a network of 10 nodes. The average delay seen at node $i = 5$ is plotted here. In Figures 4 the simulation and analytical results match well. In Figure 5, we compare the simulation and analysis for the average delay for each wireless SS associated with the Base Station, for $n = 20$ nodes. We observe that the average delay across all nodes is nearly the same for our simulation settings, and moreover, the analytical results are are well corroborated by the simulation.

IV. Delay Analysis: Polling at the Start of the Uplink Subframe

In this section we analyze the case where stations are polled at the start of the Uplink period, instead of at the end, as we have considered so far. The technique used for analysis remains the same in principle as before (Section III) and the details of the derivations are omitted because of space limitations. The same queue states are considered. For case C1 in state S0 the expected delay is given by:

$$E[D_{S0}|C1] = \frac{3T_{dl}}{2} + \frac{(i-1)T_{poll}}{2} + (n-i+1)T_{poll} + T_{ul} + (i-1)\rho(1-p)L \tag{12}$$

For case C2 in state S0 we can write the expected delay is given by:

$$E[D_{S0}|C2] = \frac{T_{ul} - (i-1)T_{poll}}{2} + T_s + T_{dl} + nT_{poll} + (i-1)\rho(1-p)L \tag{13}$$

The probability of the two cases C1 and C2 is given by:

$$P_{C1} = \frac{T_{dl} + (i-1)T_{poll}}{T_s} \tag{14}$$

$$P_{C2} = 1 - P_{C1} \tag{15}$$

The overall delay for this queue state is given by Equation (4). With the queue in state S1, for case C1, the expected delay is given by:

$$E[D_{S1}|C1] = \frac{T_{dl}}{2} + T_{ul} - \frac{(i-1)T_{poll}}{2} + N_q T_s + T_{dl} + nT_{poll} + (i-1)\rho(1-p)L \tag{16}$$

and for case C2, it is given by:

$$E[D_{S1}|C2] = \frac{T_{ul} - (i-1)T_{poll}}{2} + (N_q+1)T_s + T_{dl} + nT_{poll} + (i-1)\rho(1-p)L \tag{17}$$

The expressions for the probabilities of the two cases C1 and C2 remain the same as earlier. We can write the expected delay for state S1 as:

$$E[D_{S1}] = E[D_{S1}|C1]P_{C1} + E[D_{S1}|C2]P_{C2} \tag{18}$$

$$= E[D^*_{S1}] + N_q T_s \tag{19}$$

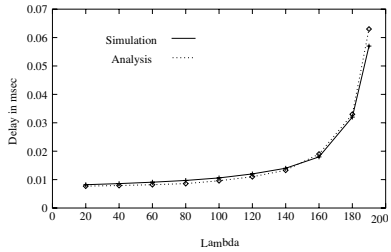

Fig. 6. Average Delay, $n = 20$, $i = 10$

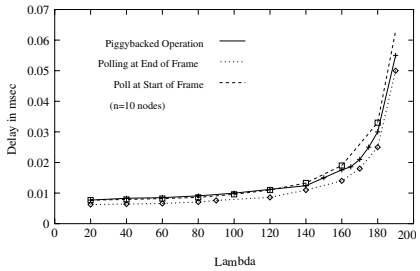

Fig. 7. Average Delay, $n = 20$, $i = 10$, Comparison with Piggybacked Operation

The overall delay for state S1 is given by Equation (7). When the tagged arrival occurs when the queue is in state S2, the delay expressions are as follows:

$$E[D_{S2}|C1] = \frac{T_{dl} + (i-1)T_{poll}}{2} + T_{ul} - (i-1)T_{poll} + (N_q - 1)T_s + T_{dl} + nT_{poll} + (i-1)\rho(1-p)L \quad (20)$$

$$E[D_{S2}|C2] = \frac{T_{ul} - (i-1)T_{poll}}{2} + N_q T_s + T_{dl} + nT_{poll} + (i-1)\rho(1-p)L \quad (21)$$

As in the case of state S1, we can write:

$$E[D_{S2}] = E[D^*_{S2}] + N_q T_s \quad (22)$$

The overall delay is given by

$$E[D] = \frac{E[D_{S0}](1-\rho) + E[D^*_{S1}]p\rho + E[D^*_{S2}](1-p)\rho}{1 - \lambda T_s \rho}. \quad (23)$$

A. Simulation Results: Polling at Start of Uplink Subframe

Simulation and analysis for this case are compared in Figure 6. The simulation scenario is the same as that described in Section III. The simulation results closely match the analytically calculated values.

V. Piggybacked Polling Operation

Here we consider the case where polling operation is more like that in the IEEE 802.11 PCF mode where the node is polled immediately after the packet from that node is transmitted. When a node has a packet to transmit in the current uplink subframe, polling takes place by piggybacking information on the packet sent out during the slot allocated to that node. Intuitively, of the three polling scenarios considered, we expect the case where nodes are polled at the start of the frame to show the largest delay. If we assume that the Uplink and Downlink subframes are of roughly equal size, and polling time is small compared to frame time, then an arriving packet misses the poll at the node roughly half the time. This means that if it must wait till the next frame to reserve bandwidth and only then obtain service. Using the same reasoning, polling at the end of the frame will see most arrivals in time for poll at their nodes. Hence service is possible in the next frame. It is expected that the piggybacked polling case shows delay behaviour in between these two earlier cases. Thus our analytical results for these two cases act as the upper and lower bounds for the delays in the piggybacked polling case. The simulation results in Figure 7 where we compare the delays of the three polling strategies for a 20 node case verifies that the delay in the piggybacked case is indeed bounded by the delay in the other two cases considered in the paper. An exact analysis of the piggybacked polling case is not attempted here. At low loads, the performance is almost identical to when stations are polled at the start of the uplink frame, since not many stations on the average have a packet to transmit when the load is low.

VI. Conclusions

In this paper we presented analytic models to evaluate the performance of IEEE 802.11 based broadband wireless access networks in terms of their packet delays. With IEEE 802.16 based networks being expected to support both fixed and fully mobile scenarios with integrated voice, video and data services, the evaluation of their delay characteristics and understanding the impact of various system settings on the delays is of critical importance and our work addresses these issues. In a system where the BS polls every node in each frame to determine their bandwidth requirements, we considered three different scheduling strategies. We developed closed form analytical expressions for the delays in the cases when the nodes are polled either at the beginning or end of the Uplink subframe and showed that these delays form upper and lower bounds for the case of piggybacked polling. Our analytical results were verified using simulations carried out using the NS-2 simulator.

References

[1] The IEEE 802.16 Working Group on Broadband Wireless Access Standards, *http://grouper.ieee.org/groups/802/16/*
[2] Draft IEEE standard for Local and metropolitan area networks, Part 16: Air Interface for Fixed Broadband Wireless Access Systems
[3] Carl Eklund, Roger Marks, Kenneth Stanwood, Stanley Wang, "IEEE Standard 802.16: A Technical Overview of the WirelessMAN Air Interface for Broadband Wireless Access", IEEE Communications Magazine, June 2002.
[4] L. Kleinrock, *Queueing Systems, Vol. I,* John Wiley and Sons, 1975.
[5] D. Bertsekas and R. Gallaher, *Data Networks,* Prentice-Hall, Englewood Cliffs, NJ, 1987.