Performance Analysis for Unsolicted Grant Service in 802.16 Networks Using a Discrete-Time Model

Chia-Chuan Chuang 1, Shang-Juh Kao 2

Department of Computer Science and Engineering, National Chung Hsing University
250, Kuo Kuang Rd., Taichuang 402, Taiwan
1ccchuang@cs.nchu.edu.tw
2sjkao@cs.nchu.edu.tw

Abstract—The standards of IEEE 802.16 define the unsolicted grant service (UGS) for the stringent delay requirement of real-time applications. In this paper, using a discrete-time GI-D-c model to describe the behaviors of UGS in 802.16 networks is presented. The performance metrics of mean number of PDUs, PDU queueing delay, and bandwidth utilization subject to the change in traffic load, allocated bandwidth are investigated. The results verify that if the required bandwidth for the real time CBR traffic is larger than the granted capacity, the mean number of PDUs will grow unexpectedly as the traffic becomes saturated. When the granted bandwidth is larger, the queue length keeps empty but suffers from unexpected growth when the ON state probability is closed to 1. Similar performance effects of PDU queueing delay are resulted from various bandwidth allocation. The experiments also show that the bandwidth utilization is linearly increased along with the traffic load.

I. INTRODUCTION

Since IEEE 802.16 [1] specifies the communication standard for wireless access, wireless metropolitan area network (WMAN) intends to provide the last mile wireless broadband access. An IEEE 802.16 network is usually composed of a base station (BS) and multiple subscriber stations (SSs). The unsolicted grant service (UGS) algorithm defined in IEEE 802.16 is designed to support real-time applications that generate fixed-size data packets periodically. Since the promise of transmission from BS is guaranteed, the access delay and bandwidth request overhead can be avoided. The delay performance of the IEEE 802.16 random access protocol has been carried out by using continuous queueing model, such as reported in [2], [3], [4], [5], [6]. Since the IEEE 802.16 is a frame-based protocol and the system state only changes at frame boundary, the discrete-time queueing model could be more suitable for characterizing the 802.16 network, especially for the study of UGS service.

The differences between discrete and continuous queueing model are in two aspects, the arrival process and the departure behavior, and are shown in Fig. 1. The figure depicts two single-servers of FIFO queue. The top diagram denotes a discrete D-D-1 model and the bottom one represents a continuous D/D/1 model. We assume that when the customer n arrives, there are no other customers waiting and the server is available. The service time of both models are deterministic of one frame length, and the arrival processes are also deterministic. However, the definitions of the arrival processes of these two model are totally different. In the continuous model, the arrival process denotes the inter-arrival time between two consecutive arrivals. Nevertheless, the arrival process of discrete model describes the probability that how many arrivals occur in a single frame. In the discrete D-D-1 model, when the customer n arrives in current frame, it stays in the buffer and gets service at the beginning of next frame. The customer n finishes service and departures at the end of next frame. Although customer n+1 arrives in the same frame as customer n, customer n+1 can receive service immediately after customer n leaves from server. As the discrete model shown in Fig. 1, the deterministic arrival process represents that two arrivals come during every frame, and the deterministic service time denotes that any arrival will take one frame for service. With the continuous model, the server can serve the customer n immediately when he arrives. The customer n+1 will get service as soon as the customer n departures. Similarly, the continuous model in Fig. 1 has the deterministic arrival process, that is, all inter-arrivals times between any two consecutive arrivals have fixed length of two frame length. Furthermore, the server in continuous model consumes service time of one frame length for all arrivals. In Fig. 1, the difference between departure times of arrival i in continuous model and those in discrete model are represented by δi, where i = n, n+1, n+2, n+3. As time goes infinity, the differences of arrival process definitions and departure behaviors make δi of these two models not analalogical. Thus, the continuous model is less accurate than discrete model and the discrete model is adopted in this paper to capture behaviors
of the frame-based IEEE 802.16 network.

The remainder of the paper is organized as follows. Section II describes UGS behavior and traffic model. Section III introduces the discrete-time $GI-D-m$ model and performance metrics such as queueing length and delay are derived. Section IV illustrates the simulation results which are compared with the model outputs. Finally, Section V concludes the study.

II. SYSTEM MODEL

A. Unsolicited Grant Service (UGS)

With UGS service, the BS grants a fixed size bandwidth to an SS periodically, which should be sufficient for the maximum number of voice packets. Since the period and the bandwidth are negotiated during the connection initialization phase, SSs do not have to make any explicit bandwidth request during the transmission. Therefore, UGS service can avoid the latency causing by bandwidth request and is suitable for real-time constant bit rate traffic. However, UGS may suffer from the low bandwidth utilization [7]. When PDUs arrive in the current frame, they are buffered in the SS. The SS is allowed to use the fixed allocated bandwidth in the next frame. If the bandwidth is larger than the arrived PDUs in a frame, all PDUs are transmitted immediately. Otherwise, SS transmits the PDUs until the allocated bandwidth is consumed.

B. Discrete-Time $GI-D-m$ Model

In the frame-based MAC protocol, the time axis is divided into frames of equal length. As packets arrive within a frame, they are stored in a buffer. After scheduling, the packets will be transmitted in the following. Packets which arrive in a frame are eligible for transmission at beginning of next frame. As elaborated in [8] and [9], discrete-time queueing model is suitable to characterize the time-slotted communication system. Accordingly, a discrete-time $GI-D-m$ model is adopted to investigate the performance of the UGS service. The $GI-D-m$ model has the following characteristics.

- The buffer has unlimited storage capacity.
- The number of servers in the queueing model via which the PDUs are removed is equal to $m > 0$.
- Time is divided into fixed-length intervals, as referred by frames. PDUs which arrive in a frame can not leave the buffer during the same frame. They would get transmitted at the next frame if any server is available. The transmission starts at the beginning of the frame and ends at the end of the same frame.
- The numbers of arrivals during the consecutive frames are modeled as i.i.d. random variables with a general and independent probability distribution. The distribution can be characterized by a probability generating function (p.g.f.), $A(z)$, and the mean number of arrivals can be derived by $A'(1)$.

With the above assumptions, it is clear that a steady state exists only if $A'(1) < m$.

Similar to the Kendall’s notation [10] in the continuous-time queueing model, the notation of $A-B-m$ is used for the discrete-time model, as defined in [8]. $A$ and $B$ denote the distribution of the number of arrivals in a frame and service time respectively and $m$ specifies the number of servers. Accordingly, the $GI-D-m$ model represents a discrete-time queueing system with $m$ identical servers where each server has the same constant service time, and the number of PDUs arrived in a frame follows a general and independent probability distribution.

C. Traffic Model

As described in the previous section, the notation $A$ of $A-B-m$ denotes the number of arrivals distribution of the source traffic. As shown in Fig. 2, the traffic pattern of real time applications, such as VoIP, are characterized by bursty ON-Off traffic, which presented by [11]. During each ON period, the fixed size packets arrive in a constant rate in the ON state. During each OFF period, no packet is generated and the SS is idle. Since these applications also have latency constraints on packet delivery, IEEE 802.16 defines UGS to support these CBR real time applications.

Let $T_{on}$ and $T_{off}$ represent the average length of the ON and OFF intervals, respectively. The probability that a SS is being ON state and generates data during a given frame, denoted by $P_{on}$, is expressed as

$$P_{on} = \frac{T_{on}}{T_{on} + T_{off}}$$

We assume that CBR applications generate fixed $N$ PDUs during each ON interval. Since fixed $N$ PDUs are generated in the ON period, the number of PDU generated from a given SS during each frame forms a random variable of a binomial distribution with parameter $N$ and $P_{on}$, and the mean value of $NP_{on}$.

Based on the assumptions of $GI-D-m$ model and the binomial distributed arrival process, the SS can be modeled as a specified Binom-D-m queueing system. Furthermore, if the number of PDUs, $N$, is large enough, the binomial distribution can be approximated by Poisson distribution and the queueing model becomes Pois-D-m. The corresponding p.g.f. of Poisson distribution is given by

$$A(z) = e^{-\alpha(1-z)}, \alpha = NP_{on},$$

and the mean number of arrived PDU is $NP_{on}$.

III. ANALYTICAL RESULTS

A. Probability Generation Function for Random Variable of Queue Length

We follow the similar derivation technique in [8] and [12] to obtain the probability generating function of the random
variable \( Q \), the queue length. \( Q \) denotes the number of PDUs in the SS buffer just after the transmission following the probability generating function (p.g.f.), \( Q(z) \). Two p.g.f’s are defined for random variable \( A \) and \( S \). The random variable \( A \) defines the number of arrived PDUs during any frame and the random variable \( S \) represents the number of PDUs in the system at the beginning of any frame, which including the PDUs in the buffer and in transmission. The probability that there are exactly \( i \) PDUs arrived during a frame is denoted by \( a_i \). The \( s_i \) is the probability that there are \( i \) PDUs in a given frame. Accordingly, the p.g.f.s of \( A \) and \( S \) are given as \( A(z) = \sum_{i=0}^{\infty} a_i z^i \) and \( S(z) = \sum_{i=0}^{\infty} s_i z^i \). Let \( S_m \) denote the number of PDUs in the system at the beginning of the \( m \)-th frame and \( A_m \) be the number of arrived PDUs during the \( m \)-th frame. We also define that \( Q_m \) as the number of PDUs in the buffer just after the transmission during the \( m \)-th frame. Since at most \( c \), the allocated bandwidth, PDUs can be transmitted during a given frame, the number of PDUs in the \( m+1 \)-th frame can be obtained from

\[
S_{m+1} = S_m - c + A_m
\]

If \( A'(1) < c \), the steady state (i.e. \( m \to \infty \)) of the number of PDUs in the system exists, and, we have

\[
S = S - c + A = Q + A. \tag{5}
\]

The underlying stochastic process \( \{ S_m | m = 0, 1, 2, \ldots \} \) forms a discrete time Markov chain, and the state transition diagram is shown in Fig. 3. The state transition probability \( P_{ij} = P(S_{m+1} = j | S_m = i) \) of the Markov chain is given by

\[
P_{ij} = \begin{cases} 
  a_{j-i+c} & \text{if } i > c, j \geq i - c, \\
  0 & \text{if } i \geq c, j < i - c, \\
  a_j & \text{if } 0 \leq i \leq c, j \geq 0.
\end{cases}
\]

Let \( S = (s_0, s_1, s_2, \ldots) \) denote the steady-state probability vector, which can be derived from \( S = SP \). According the balance requirement, \( s_j \) can be written as

\[
s_j = a_j \sum_{i=0}^{c-1} s_i + \sum_{i=c}^{c+j} a_{j-i+c} s_i, \quad j = 0, 1, 2, \ldots
\]

such that

\[
\sum_{j=0}^{\infty} s_j = 1.
\]

To obtain the p.g.f of random variable \( S \), \( S(z) \), the equations (6) are multiplied by \( z^j, j = 0, 1, 2, \ldots \). By adding these equations term by term for all \( j \), we obtain

\[
S(z) = z^{-c} A(z) \left[ \sum_{i=0}^{c-1} s_i z^i + S(z) - \sum_{i=0}^{c-1} s_i z^i \right]. \tag{7}
\]

This yields the solution of \( S(z) \) as

\[
S(z) = A(z) \frac{\sum_{i=0}^{c-1} (z^c - z^i) s_i}{z^c - A(z)}. \tag{8}
\]

According to \( S = Q + A \) in equation (5) and the convolution theorem [13], we have \( S(z) = A(z)Q(z) \) and \( Q(z) \) can be expressed as

\[
Q(z) = \frac{\sum_{i=0}^{c-1} (z^c - z^i) s_i}{z^c - A(z)}. \tag{9}
\]

By means of Rouchê’s theorem [14], it can be shown that the denominator of the equation (9) has exactly \( c - 1 \) roots, say \( z_j, j = 1, 2, \ldots, c - 1 \), inside the unit disk, and the numerator of the equation (9) must also be zero at these points. Hence, we have

\[
\sum_{i=0}^{c-1} (z^c - z^i) s_i = 0, \quad j = 1, 2, \ldots, c - 1. \tag{10}
\]

Furthermore, applying the L’Hôpital’s rule on \( Q(z) |_{z=1} \) and using \( Q(1) = 1 \), we obtain that

\[
\sum_{i=0}^{c-1} (c - i) s_i = c - A'(1). \tag{11}
\]

By solving the equations (10) and (11), we can determine the \( s_i, i = 0, 1, \ldots, c - 1 \), and finally obtain the p.g.f. of random variable \( Q \).

\[
Q(z) = (c - A'(1)) \frac{z - 1}{z^c - A(z)} \prod_{j=1}^{c-1} \frac{z - z_j}{1 - z_j}, \tag{12}
\]

where \( z_j, j = 1, 2, \ldots, c - 1 \), are the \( c - 1 \) complex roots of the \( z^c - A(z) \) inside the unit disk.

### B. Mean System Length and Mean Queue Length

The average queue length \( E[Q] \) for the UGS flows can be obtained by computing \( Q'(z) |_{z=1} \) and apply the L’Hôpital’s rule twice. It yields

\[
E[Q] = \sum_{j=1}^{c-1} \frac{1}{(1 - z_j)} - \frac{c(c - 1) - A''(1)}{2[c(c - 1)]}. \tag{13}
\]

Since \( S = Q + A \), the mean system length \( E[S] \) can be obtained by summing mean queue length and mean number of arrivals. Hence,

\[
E[S] = E[Q] + A'(1). \tag{14}
\]
C. PDU Queueing Delay

The delay of a PDU is defined as the number of frames between the end of the frame of arrival of the PDU, and the end of the frame when this PDU leaves the buffer. To find the mean queuing delay, we employ Little’s formula to calculate $E[D]$. The arrival rate of the queuing system is the mean arrived PDUs in each frame. Thus, we can express $E[D]$ as

$$E[D] = \frac{E[S]}{A'(1)}$$

$$= \frac{1}{A'(1)} \left[ A'(1) + \sum_{j=1}^{c-1} \frac{1}{1 - z_j} - \frac{c(c - 1) - A''(1)}{2(c - A'(1))} \right]$$

D. Bandwidth Utilization

Owing to the silence periods of the voice connection, the voice sources do not always generate the voice packets. In the UGS flows, the BS always assigns fixed allocated bandwidth that are sufficient to send voice packets to the voice user. The utilization of the allocated bandwidth denoted by $U$ is adopted to study the percentage of the bandwidth used for the voice connection. Since the system length $S$ is the queue length plus the PDUs in the channels, the equation (14) implies that the number of the busy channels is $A'(1)$. Because $c$ is the total fixed allocated bandwidth, the utilization ($U$) can be expressed as:

$$U = \frac{A'(1)}{c}.$$  

IV. Simulation Results and Comparisons

To validate the accuracy of the analytical model, an event-driven simulator is developed to evaluate the UGS performance. Since $P_{on}$ denotes the proportion that each source resides in the ON states, the increase on $P_{on}$ indicates that the traffic load becomes heavier. With UGS service, the BS reserves a fixed bandwidth for SSs at each frame, it can be expected that the buffer is nearly empty when the mean number of arrived PDUs in a frame is less than the allocated bandwidth. As $P_{on}$ increases, the mean number of arrived PDUs also increases. If the number of arrived PDUs in a single frame is larger than the allocated bandwidth, they can not be transmitted in a frame time. The excess PDUs are left in the buffer until the next frame time. The left PDUs cause the growth of the queue length. In contrast, the lower $P_{on}$ causes the fewer PDUs arriving in a single frame, and the arrived PDUs can be transmitted immediately if the number of arrived PDUs is less than the allocated bandwidth. With the given connections or allocated bandwidth, the buffer will be filled drastically depending upon the value of $P_{on}$. When the PDUs from connections are more than the allocated bandwidth, the buffer will grow soon as the traffic load is saturated. Otherwise, with more allocated bandwidth than connections, the buffer is almost empty until $P_{on}$ is closed to 1. The phenomena are illustrated in Fig. 4.

In Fig. 4, we study the mean number of PDUs in the queue under different parameter configurations which $c = 100$ and $N$ is set as 90, 100, and 160. Since we assume that fixed $N$ PDUs are generated during the ON period, the mean number of PDUs in each frame is $NP_{on}$. When the mean number of arrivals, $NP_{on}$, exceeds the allocated bandwidth at certain $P_{on}$, the untransmitted PDUs result in the larger buffer length. If the mean number of arrival PDUs keep less than the allocated bandwidth, the PDUs can be transmitted immediately and make the buffer empty. Fig. 4 demonstrates the As $N = 160, c = 100$ in Fig. 4, the buffer becomes suddenly increased at $P_{on} = 0.6$, when the mean number of PDUs is more than the allocated bandwidth, Otherwise, in the cases with more allocated bandwidth than the PDUs from associated connections, as $c = 100$, $N = 90$ in Fig. 4, the buffer is empty all the time but suffers from unexpected growth when $P_{on}$ is closed to 1.

The Fig. 5 depicts the delay variations as opposed to different $P_{on}$ in both case: $N = 128, c = 128$ and $N = 128, c = 120$. Once an UGS connection is admitted, BS would promise maximum available bandwidth and no further bandwidth request is needed. Thus PDUs can be transmitted with the minimum delay. If the newly PDUs arrive in the current frame, these PDUs are placed in the SS buffer. The SS is allowed to use the promised maximum bandwidth in the next frame to transmit these PDUs. This results that the delay is at least single frame. In Fig. 5, with $N = 128$ and $c = 120$, the mean number of PDUs is larger than the allocated bandwidth at $P_{on} = 0.94$, and the queueing delay also starts raising at $P_{on} = 0.94$. The figure also shows that when $N$ and $c$ are identical, the queueing delay of this case increases when $P_{on}$ is higher than 0.93. In this case, the excess PDUs will be generated at $P_{on} = 0.98$, which is higher than 0.93, and the excess PDUs are delayed until the next frame transmission.

The bandwidth utilization shown in Fig. 6 is the usage of allocated bandwidth for UGS service. The lower utilization means that more allocated bandwidth is not used. More allocated bandwidth are used to transmit the PDUs with the
larger $P_{on}$. In all the combinations of $c$ and $N$, bandwidth utilization in linearly increased along with the value of $P_{on}$. The connections would soon run out of the allocated bandwidth when $N$ is larger than $c$. This is because that the number of incoming PDUs will soon exceed the allocated bandwidth even at lower $P_{on}$. Similarly, if $N$ is smaller than $c$, the allocated bandwidth is still wasted even at the heavy traffic.

Fig. 5. Mean delay with different combinations of $N$ and $c$

Fig. 6. Bandwidth utilization with different combinations of $N$ and $c$

becomes larger, traffic becomes heavier. Bandwidth utilization is linearly increased along with the growth of $P_{on}$. The queueing delay and queue length grow rapidly if the traffic becomes saturated. However, the performance is not sensitive to the $P_{on}$ if the allocated bandwidth is larger than the number of connections. Although the UGS service is very steady at light traffic, the drawback of UGS service is that lower bandwidth utilization would result in the waste of allocated bandwidth. Simulation experiments have been conducted to verify the discrete-time queueing formula and consistent results were concluded.

V. CONCLUSION

In this paper, a discrete-time queueing system is proposed to investigate the performance of UGS service. Under the discrete model, traffic load, allocated bandwidth, and the number of SSs are varied for performance evaluation. Both analytical and simulation results show that UGS connections are stable in most circumstances. When the allocated bandwidth is less than the maximum arrived PDUs, the $P_{on}$ is crucial. When $P_{on}$