Performance of the IEEE 802.16e Sleep Mode Mechanism in the Presence of Bidirectional Traffic

Koen De Turck, Sergey Andreev, Stijn De Vuyst, Dieter Fiems, Sabine Wittevrongel and Herwig Bruneel
Ghent University, Department of Telecommunications and Information Processing

Abstract—We refine existing performance studies of the WiMAX sleep mode operation to take into account uplink as well as downlink traffic. This as opposed to previous studies which neglected the influence of uplink traffic. We obtain numerically efficient procedures to compute both delay and energy efficiency characteristics. A test scenario with an Individual Subscriber Internet traffic model in both directions shows that even a small amount of uplink traffic has a profound effect on the system performance.

I. INTRODUCTION

The IEEE 802.16e standard (WiMAX) [1] is an emerging standard that has the potential to become the major standard for wireless communication in the near future. It regulates the communication between mobile stations (MS) and base stations (BS) in a metropolitan area wireless network. Energy-saving mechanisms in wireless communications are currently a hot topic. Short battery life is one of the main impediments to a more widespread use of wireless devices. Hence, understandably, a lot of research is directed at solving or at least lessening this problem. In the first place, this can be done by making more efficient batteries, but lately there is also a lot of interest in including energy-saving measures in the communication protocols themselves. On that account, it is no wonder that the WiMAX committee has opted to incorporate various energy-saving elements which are commonly referred to as ‘sleep mode’ and ‘idle mode’. Power saving is generally achieved by turning off parts of the MS in a controlled manner when there is neither traffic from the MS (uplink traffic) nor to the MS (downlink traffic). Whereas a MS in sleep mode is still registered to a BS and still performs hand-off procedures, idle mode operation (which is optional in current WiMAX standards) goes further and allows the MS to be completely switched off and unregistered with any BS, while still receiving broadcast traffic.

In this paper, we consider mainly the sleep mode mechanism, in which the MS turns itself off for predetermined periods of time which are negotiated with the BS. These periods are often referred to as ‘sleep windows’. Three sleep mode classes are defined by the WiMAX standard. When in Sleep Mode Class I, the sleep window is progressively doubled in size from a prenegotiated minimum to a prenegotiated maximum value. Having reached the maximum sleep window length, this sleep window is repeated incessantly, until traffic arrives. This class is considered to be suitable for best-effort and non-real-time traffic. Class II features fixed-length sleep windows, that is the same sleep window size is repeated continuously until there is uplink or downlink traffic to be transmitted. This class is typically employed for UGS (unsolicited grant services) traffic. Finally, class III negotiates only a one-time sleep period. This is typically used for management traffic, when the MS knows when the next traffic is to be expected.

Although class I is the most interesting from a modeling point of view (and will as such get the most attention), the analysis in this paper is general enough to encompass all three classes. That is, in our model we assume that a general sequence of sleep window sizes $t_i, i \geq 1$, has been negotiated between MS and BS. We furthermore assume that this sequence is kept fixed during the entire operation of the system, instead of being negotiated.

Sleep mode operation has received quite a lot of interest lately from the performance modeling community. In [2], the average energy consumption of the MS is obtained in case of downlink traffic only, as well as an approximate expression for the mean packet delay. The energy consumption of the MS in case of both downlink and uplink traffic is considered in [3]. Both [2] and [3] model the incoming (and outgoing) traffic as a Poisson process. An accurate assessment of the delay experienced at the BS buffer however, requires a queueing model. For IEEE 802.16e, in [4] the BS buffer is modeled as a continuous-time finite-capacity queue with a Poisson arrival process and deterministic service times. A semi-Markov chain analysis leads to expressions for the mean packet delay and the mean energy consumption by the MS. The analysis in [5] is based on an $M/G/1/K$ queueing model with multiple vacations and exhaustive service, where the vacations represent the sleep periods. Similar work can also be found in [6], where the length of a vacation is assumed to depend on the previous vacation length. In [7], the sleep mode operation in Cellular Digital Packet Data (CDPD) services is evaluated. The difference with IEEE 802.16e is that the subsequent sleep periods do not increase in length. The system can thus be modeled as a queueing system with multiple vacations and exceptional first vacation. The loss probability in both [5] and [7] is obtained as well. A simulation study of CDPD

1Sergey Andreev was on a research visit at Ghent University during the realization of this work. He is affiliated with the Saint-Petersburg State University of Aerospace Instrumentation.
sleep mode performance is found in [8]. An alternative to the exponential increase of the sleep period lengths is evaluated by simulation in [9]. Finally, in our previous work on the sleep mode mechanism [11], we considered a general D-BMAP arrival process, and we found that traffic correlation, which was hitherto neglected in almost every study, has an important influence on the sleep mode performance.

In the present paper, we investigate how the presence of uplink traffic affects the performance of the sleep mode operation, which is another aspect that has received little attention in existing studies. In fact, only in [3], uplink traffic has thus far been given attention, although in that paper it was restricted to Poissonian traffic only.

The influence of uplink traffic is considerable, as the MS immediately ceases the sleep mode operation whenever there is uplink traffic to be transmitted. Thus, during uplink transmission, no extra sleep mode delay is experienced by the antenna (as the antenna is already switched on). We model the uplink activity by a finite but otherwise general Markovian background process.

The structure of this paper is as follows. In Section II, we expound on the details of the model. The analysis is spelled out in Section III. We show how to employ our analysis for a realistic WiMAX scenario in Section IV together with some numerical results. Finally, we draw some conclusions in Section V.

II. MODEL

We model the buffer in the base station (from now on denoted as simply ‘the buffer’) as a discrete-time queue with infinite capacity and a first-come-first-served (FCFS) server discipline. This buffer models the traffic in the downlink direction (i.e. from the base station to the mobile station). A sample evolution of the system is shown in Figure 1.

![Figure 1](image-url)

During each slot, packets arrive in the buffer in batches according to an arrival process that is detailed later on. The service times of the packets (i.e. the times needed to transmit a packet to the MS) are considered to be independent and have probability generating function (pgf) \( S(z) = \sum_i \Pr[s = i]z^i \).

During the course of our analysis we refer to slots during which there is uplink traffic as ‘active’; the rest of the slots are called ‘passive’. Both the uplink traffic and the downlink traffic are modulated by the same (background) Markov chain with \( M \) phases, characterized by transition matrix \( P \), with entries as follows:

\[
[P]_{ij} = \Pr[\varphi_k = j | \varphi_{k-1} = i],
\]

where \( \varphi_k \) is the random variable that denotes the phase (or background state) at time instant \( k \). The probability that a slot is active is phase dependent; during phase \( i \), this probability is \( q_i \). The distribution of the number of arrivals is also modulated by the background Markov chain. The pgf of the batch size during phase \( i \) is given by \( A_i(z) \). For convenience, we define three diagonal matrices with entries as follows:

\[
[Q]_{ij} = \delta_{ij} q_i, \quad [\overline{Q}]_{ij} = \delta_{ij} (1 - q_i), \quad [A(z)]_{ij} = \delta_{ij} A_i(z),
\]

where the notation \( \delta_{ij} \) denotes Kronecker’s delta. Note that this scheme is very versatile and permits the uplink traffic to be correlated to the downlink traffic. The load \( \rho \) is defined as the average number \( E[a] \) of packet arrivals per slot times the average service duration \( E[s] \). The system is stable if the condition \( \rho < 1 \) is fulfilled.

When the queue is empty and the system is active, the operation is not different from usual queueing models: the server resumes work in the slot immediately after the first packet arrival. However, when the buffer becomes empty and the system is passive, it enters sleep mode: an internal timer is then started and the server awakes to check the queue content after a sleep period of \( t_1 \) time slots. When upon awakening the server finds that there are still no packets, it goes to sleep again, this time for a sleep period of \( t_2 \) slots. The next sleep period would be of length \( t_3 \), and so on. In general, \( t_n, n \geq 1 \), denotes the length (expressed in slots) of the \( n \)th sleep period after the queue has become empty. If after a sleep period, the server finds a non-empty queue, it serves all packets present at that point and also all new packets that arrive while the server is working, until the queue becomes empty again and the whole procedure is repeated.

When the system becomes active during a sleep period, sleep mode is immediately abandoned, and the system is ready to serve incoming packets immediately. Note that, when the system becomes passive again while the buffer is still empty, sleep mode is resumed, but starting again with a sleep period of length \( t_1 \), and so on.

Let us agree upon some terminology: the system alternates busy periods (during which the downlink traffic is being processed) with idle periods. Idle periods consist of sleep periods and ‘idle but active’ periods. Note that some of our terminology is standard in queueing theory, but may have another meaning in the IEEE community. For example, the term
‘idle period’ is fairly standard queueing theory terminology but should not be confounded with the idle-mode mechanism of IEEE WiMAX. Secondly, we have opted for the neutral term ‘slots’ but will explain in Section IV that the most natural translation into a WiMAX context is a WiMAX frame. Lastly, the sleep periods we consider include the listening intervals at the end. Hence, sleep periods in this paper, are WiMAX sleep intervals plus the corresponding listening interval.

III. ANALYSIS

We aim to determine the vector generating function \( U(z) \) of the buffer content during random slots. We do so by introducing ‘event tags’ at special time instants, and subsequently finding relationships of the stationary distribution of the buffer content at such time instants.

For our first set of special time instants we look inside idle periods. We mark the beginning of an idle period with \( I \), the beginning of the \( n \)th sleep period by \( S_n, \ n > 0 \), and active but idle time instants by event tag \( A \). Note that, necessarily, the buffer at such time instants is empty, and hence, the equilibrium state during such time instants consists only of the phase variable. The probability vectors \( u_I, u_A \) and \( u_s_n \) have entries defined as follows:

\[
[u_I]_j = \Pr[\varphi = j, I].
\]

The other vectors are defined analogously. We state three relationships that lead to formulas involving these probability vectors.

A. A time instant is marked by tag \( S_1 \) (i.e. first slot of the first sleep period) if one of two (mutually exclusive) conditions are fulfilled: (1) the time instant is marked with tag \( I \) and it is a passive time instant as well; (2) the previous time instant is active but idle, no arrivals occur in the slot, and the current time instant is passive. This leads to the following formula:

\[
u_{S_1} = u_I Q + u_A PQA(0).
\]

B. A time instant is marked by tag \( S_{n+1} \) if the time instant \( t_n \) slots earlier is marked by tag \( S_n \), and the system has remained passive and without arrivals during those \( t_n \) slots,

\[
u_{S_{n+1}} = u_s_n (PQA(0))^n.
\]

C. The third formula clarifies the occurrence of active but idle time instants. They occur if one of three mutually exclusive conditions is fulfilled: (1) the time instant is marked with tag \( I \) and it is an active time instant; (2) the previous time instant is marked by \( A \), no arrivals occur during the previous slot and the current slot is active; (3) the system becomes active during a sleep period, that is there exists a time instant marked with \( S_1 \), no arrivals have occurred in the interval up until the current time instant, and all but the current time instant in the enclosed interval are passive. This gives

\[
u_A = u_I Q + u_A PQA(0)
\]

\[+ u_{S_1} (I - PQA(0))^{-1}PQA(0).
\]

Now we focus our attention to sets of special time instants inside busy periods. We discern three types: the first slot of a busy period (marked by tag \( B \)); the time instants at which a new service is started (marked by tag \( b \)); and the time instants at which a service is completed (marked by tag \( C \)). They are characterized by vector generating functions of the buffer content \( U_B(z) \), \( U_b(z) \) and \( U_C(z) \) respectively, with entries defined as follows:

\[
[U_B(z)]_{j} = \sum_{k=0}^{\infty} \Pr[u = k, \varphi = j, B] z^k.
\]

Their behavior is captured in the following three relations:

D. The system has three distinct possibilities of entering a busy period at a time instant: (1) the previous instant is active but idle, and there was at least one arrival during the previous slot; (2) the system has become active during a sleep period in which at least one arrival occurred; (3) at the end of a sleep period, in which the system has remained passive but during which arrivals have occurred. Each of the possibilities corresponds to a term in the following equation:

\[
U_B(z) = u_A P(A(z) - A(0))
\]

\[+ \sum_{n=1}^{\infty} u_s_n \sum_{i=0}^{t_n-1} (PQ)^i PQ(A(z)^{i+1} - A(0)^{i+1})
\]

\[+ \sum_{n=1}^{\infty} u_s_n (PQ)^n (A(z)^{n} - A(0)^{n}).
\]

E. The evolution of the buffer content from the beginning to the completion of a service is as follows. In a single server FCFS system, from the start to the completion of a service, only one packet leaves the buffer. The number of arriving packets during a service time depends on the length of a service. When the service length is \( k \), the pgf of the number of arrivals is \( A(z)^k \), and the transition of the background phase is given by transition matrix \( P^k \).

Therefore, we have

\[
U_C(z) = U_b(z) \left( \frac{1}{z} \sum_{k=1}^{\infty} \Pr[S = k](PA(z))^k \right) = U_b(z) \frac{S(PA(z))}{z}.
\]

F. The set of time instants that is marked as \( C \) but not as \( I \), is equal to the set of time instants that is marked as \( b \) but not as \( B \). Hence, the distribution of the buffer content at such a subset of time instants must be equal as well:

\[
U_C(z) - u_I = U_b(z) - U_B(z).
\]

There are two final relations that give the distributions during a random slot and a random busy slot.

G. We can prove that

\[
U(z) = U(z) PA(z) + (1 - z) U_C(z).
\]
H. The vector generating function of the buffer content during a random busy slot is given by

\[ U_β(z)(I - PA(z)) = U_δ(z)(I - S(PA(z))). \]  \hspace{1cm} (12)

We can solve each of the occurring functions and variables in terms of the constant vector \( u_β \), which can subsequently be found by standard techniques to compute the boundary vector of an \( M/G/1 \)-type queueing model [12]. We can then get such performance measures as the moments of the buffer content at various time instants, as well as probabilities that slots of a certain type occur. As we will see, this is crucial for the computation of power consumption measures. By Little’s law, we can get from the mean buffer content also the mean packet delay.

IV. APPLICATION TO A WiMAX SCENARIO

In order to provide a numerical example to the analytical approach of this paper we conduct a simplified but representative power consumption investigation of an IEEE 802.16e mobile station. We restrict further derivations to the Time Division Duplex (TDD) scheme together with the Orthogonal Frequency Division Multiplexing (OFDM) physical layer.

Following many works on the performance analysis of IEEE 802.16 networks (see, for example [13] and [14]) we set the necessary operation parameters as follows:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T_f )</td>
<td>5.000 ( \mu s )</td>
<td>Frame duration</td>
</tr>
<tr>
<td>( T_h )</td>
<td>400 ( \mu s )</td>
<td>Header duration</td>
</tr>
<tr>
<td>DL/UL</td>
<td>50:50</td>
<td>Downlink-uplink ratio</td>
</tr>
<tr>
<td>( T_{RTG} )</td>
<td>22.9 ( \mu s )</td>
<td>Receive transition gap</td>
</tr>
<tr>
<td>MCS</td>
<td>16-QAM ( \frac{1}{2} )</td>
<td>Modulation and coding scheme</td>
</tr>
</tbody>
</table>

We denote \( T_d = \frac{T_f}{2} \) and \( T_u = \frac{T_f}{2} \) to be the durations of the DownLink (DL) and UpLink (UL) sub-frames, respectively. As we focus on the Best Effort (BE) and Non Real-Time Polling Service (nrtPS) traffic classes only, we select packet arrival model for the downlink traffic according to IEEE recommendations [15]. In the “Individual Subscriber Internet model” from [15] the packet arrivals are assumed to conform to the Interrupted Poisson Process (IPP). This is a two-state Markov chain with a Poisson arrival process in the ON state, and no arrivals in the OFF state. The aforementioned source recommends an arrival rate of \( \lambda_p = 22.79 \) packets per time unit in the ON state. The data packet length is \( L = 192 \) bytes. The resulting model represents a single user interacting with the Internet and gives the resulting data flow with the average arrival rate of 15 Kbps. Since the model approximates the aggregation of HTTP/TCP and FTP traffic, it is subject to the Class I power saving mechanism.

We follow the approach of [16] to obtain the data packet duration for the 16-QAM \( \frac{1}{2} \) modulation and coding scheme. The raw data rate in this case is approximately equal to \( R = 11.29 \) Mbps. Therefore, packet duration \( (T_p) \) is readily obtained as:

\[ T_p = \frac{8 \cdot L}{R}. \]  \hspace{1cm} (13)

The above expression allows us to calculate the number of packets per DL sub-frame \( (N_d) \) as:

\[ N_d = \frac{\left( T_d - T_h - T_{RTG} \right)}{T_p}. \]  \hspace{1cm} (14)

As the derived analytical model is discrete with a slot duration being equal to the frame duration, it is convenient to rescale the packet arrival rate from packets per time unit to blocks per time unit. Here each block is exactly \( N_d \) packets necessary to fill the DL sub-frame. The rescaled arrival rate is, therefore, \( \lambda_b = \frac{N_d}{T_p} \) blocks per time unit in the ON state. Service times of such blocks equal exactly one frame, and hence, \( S(z) = z \). The uplink traffic model is, again following the recommendations of [15] a rescaled version of the downlink traffic model. That is, it carries 10 times less traffic (i.e. 1.5 Kbps) and transition rates are 10 times smaller. As the uplink traffic rate is very small compared to the maximal capacity, we neglect the uplink queuing effect. Combining the two traffic models, we get a background process with four states. In this example, the two traffic streams are assumed to be independent, although our model can incorporate such correlation as well.

To the best of our knowledge, no definitive real-world power consumption values have been reported for IEEE 802.16e MS. Therefore, we borrow the respective values from the corresponding paper on IEEE 802.11 analysis [17] and summarize them in the following table:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P_s )</td>
<td>0.045 W</td>
<td>Sleep mode power</td>
</tr>
<tr>
<td>( P_l )</td>
<td>1.15 W</td>
<td>Listening mode power</td>
</tr>
<tr>
<td>( P_r )</td>
<td>1.40 W</td>
<td>Receive mode power</td>
</tr>
<tr>
<td>( P_t )</td>
<td>1.65 W</td>
<td>Transmit mode power</td>
</tr>
</tbody>
</table>

Notice that there are five types of frames (slots) in the considered system:

1) Sleep frame: wireless interface is inactive and consumes \( P_s \).
2) Listening frame: wireless interface is active and consumes \( P_l \), no packet receptions/ transmissions.
3) Transmit frame: wireless interface is active \( (P_t) \) and there is a packet transmission in the UL sub-frame \( (P_r) \).
4) Receive frame: wireless interface is active \( (P_t) \) and there is a reception of \( N_d \) packets in the DL sub-frame \( (P_r) \).
5) Receive-transmit frame: wireless interface is active \( (P_t) \), there is a reception of \( N_d \) packets in the DL sub-frame \( (P_r) \) and there is a packet transmission in the UL sub-frame \( (P_r) \).

The above frame types lead to the following simple expressions for the energy costs per slot, which are denoted by \( E_s, E_l, E_r, E_r \) and \( E_{rt} \) for the sleep, listening, transmit, receive and receive-transmit frame type, respectively:
We set the listening interval at the end of each sleep interval to be equal to $t^{(f)} = 2$ frames and we keep the final sleep interval fixed at a value of $t_{\text{max}} = 256$ frames.

Using the results from our analysis we can derive the following expression for the mean energy consumption per frame:

$$E = \sum_{n=1}^{\infty} u_{\delta_n} 1((t_n - t^{(f)})E_s + t^{(f)}E_d) + E_{rt}pU_{\beta}(1)Q1 + E_{s}pU_{\beta}(1)Q1 + E_{f}u_{\delta}1,$$

where $1$ is a column vector of length $M$, with all entries equal to 1. Finally, we obtain the power consumption of the radio interface by dividing $E$ by the frame length $T_f$.

In Figure 2, we compare a bidirectional scenario with a scenario in which the uplink traffic is neglected. The left subplot shows the mean delay of the uplink traffic versus the length of the initial sleep interval. The delay is considerably lower for the bidirectional scenario. In the right subplot, we see that a downlink traffic-only model underestimates the power consumption considerably.

V. CONCLUSION

We analyzed the sleep mode mechanism of IEEE 802.16e WiMAX protocol, by means of a queuing model that takes into account both downlink and uplink traffic. We are thus able to compute efficiently such crucial performance measures as the mean packet delay and the expected power consumption. We applied our theoretical model to a realistic scenario and found that even a modest amount of uplink traffic has a tremendous influence on the system performance. The versatility of the chosen downlink traffic model allows us to model a multitude of traffic situations, and to assess the influence on the power delay and the power consumption.

ACKNOWLEDGEMENTS

The fourth author is a Postdoctoral Fellow with the Research Foundation–Flanders (FWO–Vlaanderen), Belgium.

REFERENCES


