# Call-level Performance Modeling of Voice over IEEE 802.16 Systems

B. P. Tsankov[1], P. H. Koleva[2], K. M. Kassev[3], V. K. Poulkov[4]

*Abstract* — **Call and burst-level performance modeling of IEEE 802.16 based networks for voice communications is considered. An analytical method for carrier grade voice traffic over IEEE 802.16 system evaluation for UL transmission is proposed. The results demonstrate an application of the method used in the study for IEEE 802.16 network design, such as CAC deployment and envisage some system characteristics.**

*Keywords* — **Backhaul network, burst-level traffic, call-level traffic, IEEE 802.16, traffic performance, VoIP.**

## I. Introduction

It is expected the packetized voice to be the most popular application for IEEE 802.16 systems (known as WMAN or WiMAX) and major revenue earner for network service providers. Apart from the possibility of providing wireless broadband connections to home and small-business users, replacing DSL and cable modems, the IEEE 802.16 systems can also be used in backhaul networks for cellular base stations, bypassing the public switched telephone network as well as for backhaul connections to the Internet for WiFi hotspots. In this case, VoIP is a carrier grade service with stringent QoS requirements. In order to provide the required quality of service for any particular traffic type (in our case a voice transmission), a suitable scheduling algorithm for a real-time application should be applied. Hence, there are intensive investigations [1–3] of voice traffic performance under different standardized and proposed scheduling algorithms. A special attention to the problems, concerning a voice transmission is paid in the current version of the IEEE 802.16e standard, in which the quality of service (QoS) is supported by allocating each connection between the SS and the BS (called a service flow in the 802.16 terminology) to a specific QoS class – unsolicited grant service (UGS), real-time polling service (rtPS), enhanced real-time polling service (ertPS), non-real-time polling service (nrtPS), and best-effort service (BE).

For a carrier grade voice service the network operator has to offer QoS similar to that of circuit switched networks, such as fixed (PSTN) and mobile (GSM) networks.

The assessment of voice degradation due to packet delay and packet losses is a subject of separate investigations, as it has already been presented in [4] and [5], facing the problem of WiMAX environment and concluding that the "VoIP call quality is more sensitive to packet losses rather than packet delay".

For the purpose of analysis, we accept the maximum delay of 60 ms and packet losses of 0.5%, introduced by an IEEE 802.16 sysem [4].

Most of the published IEEE 80.16 voice traffic performance investigations consider overloaded or nearly overloaded conditions. This is specifically true if simulation is used [1], [6]. It should be noted the QoS norms for a carrier grade voice traffic service restrict the system load far before an overloaded condition occurs. Thus, the network service providers are interested in the system performance evaluation under normal load conditions where QoS measures, such as blocking, packet losses, etc. are rare events, which are often difficult to be estimated by the means of simulation.

In this paper we propose an analytical method for carrier grade voice traffic evaluation over IEEE 802.16 system uplink transmission, considering the application of particular scheduling services.

## II. Carrier Grade Performance Analysis

The most important features of queuing systems [7] and particularly those serving a superposition of independent sources [8], are the waiting time $t_q$ and the packet losses $P_{PL}$. In our case, the packet losses (the probability the buffer overflows a finite length) are closely approximated by the probability the infinite buffer contains more packets than given finite buffer length. This is particularly true when the system is not overloaded, as it is with the carrier grade VoIP systems.

The total delay time $t_d$ for IEEE 802.16 systems is defined with two components: $t_d = t_{MAC} + t_q$, where $t_{MAC}$ is a time delay inherent to the MAC protocol, taking into account that $t_{MAC}$ is independent of the traffic carried and depends upon the scheduling algorithm used. The case in which the numbers of connections, which are simultaneously in an active state (number of simultaneous talk spurts or bursts, when voice activity detection is supported) are more than the IEEE 802.16 system can support, corresponds to the buferless burst-scale packet losses - $P_{ER}$, and therefore some packets have to be stored in the buffer.

[1]B. P. Tsankov is with the Faculty of Telecommunications, Technical University of Sofia, 8 Kliment Ohridski Blvd., 1000 Sofia, Bulgaria, E-mail: bpt@tu-sofia.bg

[2]P. H. Koleva is with the Faculty of Telecommunications, Technical University of Sofia, 8 Kliment Ohridski Blvd., 1000 Sofia, Bulgaria, E-mail: p_koleva@tu-sofia.bg

[3]K. M. Kassev is with the Faculty of Telecommunications, Technical University of Sofia, 8 Kliment Ohridski Blvd., 1000 Sofia, Bulgaria, E-mail: kmk@tu-sofia.bg

[4]V. K. Poulkov is with the Faculty of Telecommunications, Technical University of Sofia, 8 Kliment Ohridski Blvd., 1000 Sofia, Bulgaria, E-mail: vkp@tu-sofia.bg

In order to meet the QoS requirements, the buffer size is restricted by the maximum allowable time delay $t_{d,\max}$, so that the corresponding queuing packet delay is:

$$t_{q,\max} = t_{d,\max} - t_{MAC} \quad (1)$$

Our aim is to dimension the IEEE 802.16 system for VoIP traffic, meeting the QoS requirements. We determine $t_{MAC}$ in dependence of the scheduling algorithm and knowing maximum permitted delay $t_{d,\max}$ we calculate $t_{q,\max}$ from Eq. (1) and the corresponding buffer size - $k$. In the next section, we present an analytical tool forming the relation connecting the buffer size $k$, the probability $P_{PL}$ of packet losses, the transmission capacity $C$ and the traffic load.

## III. Fluid-Flow Approach Application

We assume the voice packets tend to be of fixed size and generated by homogenous and independent traffic sources. Taking into account the fact the most of voice codecs support a voice activity detection (VAD) algorithm as a means of reducing average bit rate and enhancing overall coding quality of speech, thus the VoIP packet sources are ON-OFF sources with exponentially distributed state period durations - $T_{on}$ and $T_{off}$. The activity factor is $\alpha = T_{on}/(T_{on}+T_{off})$, which together with the number of connections (calls) $N$ and packet rate $c$ during source ON period forms the traffic load.

We analyze the burst-scale queue state probability and corresponding packet delays and losses. We apply the "fluid-flow" approach in a way similar to that used in [9], [10]. In order to simplify the traffic model, the aggregation of $N$ traffic flows is substituted by a single equivalent source having two states, as well (Fig. 1).

The aggregate process is in the ON state when the number of active voice sources is more than $n = C/c$ with a mean output packet rate $C_{ON}^{Eq}$. The equivalent source is in OFF state with a mean output rate $C_{OFF}^{Eq}$ in case less than $n$ voice source are active. The distributions of both the ON and OFF periods $T_{ON}^{Eq}$ and $T_{OFF}^{Eq}$ of the equivalent source are modeled exponentially, as well.

The buffer overflow probability $Q(k)$ for traffic process aggregated by multiplexing $N$ independent sources and feeding buffer of length $k$ is presented as [8]:

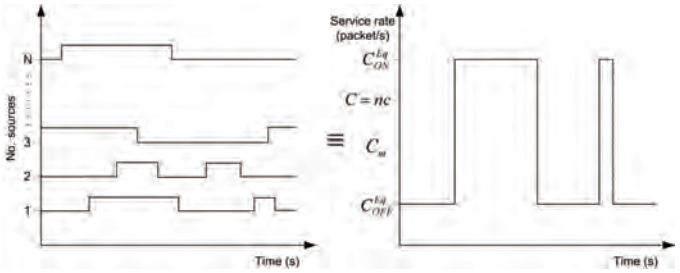

Fig. 1. State-space reduction for aggregate traffic

$$Q(k) = P_{ER}.d^{k+1} \quad (2)$$

From the burst-scale point of view, if more packet flows are active, the queue increases in size, because of the excess rate. The probability a packet to be an excess rate arrival is denoted by $P_{ER}$. The value $d$ of the *decay rate* is obtained by [9]:

$$d = \{1-1/T_{ON}^{Eq}.(C_{ON}^{Eq}-C)\}/\{1-1/T_{OFF}^{Eq}.(C-C_{OFF}^{Eq})\} \quad (3)$$

We need now expressions for $C_{ON}^{Eq}$, $C_{OFF}^{Eq}$, $T_{ON}^{Eq}$ and $T_{OFF}^{Eq}$.

The overall rate of bursts from the sources (subscriber stations –SSs) to BS is $\lambda_B = N/(T_{on}+T_{off})$, and the overall offered burst traffic in terms of Erlangs is $A = \lambda_B.T_{on}$. Assuming a memoryless process for the arrival of packet flows, this situation is equivalent to a system modeled by Erlang's waiting-call analysis. According to Erlang-C formula the probability a call (in our case a burst) to wait is $P(>0) = E_C(n,A)$. It could be also expressed by Erlang-B formula $E_B(n,A)$, which is easy calculated ([11], p. 257). According to Erlang's model of waiting systems the mean queue length ("waiting" bursts) given the queue is then greater than zero (excess rate condition) – $L_q = A/(n-A)$, and the excess rate itself is:

$$C_{ON}^{Eq} = C + L_q.c = C + A.c/(n-A) \quad (4)$$

Unconditioned mean queue length is $L = P(>0).L_q$ and applying Little's theorem we have $L = \lambda_B.W$, where $W$ is the mean delay for all offered calls (bursts). The mean delay for delayed (excess rate) burst only is the average duration of the state of excess rate, and is expressed as:

$$T_{ON}^{Eq} = W/P(>0) = T_{on}/(n-A) \quad (5)$$

The obvious relation $T_{ON}^{Eq}/(T_{ON}^{Eq}+T_{OFF}^{Eq}) = P(>0)$ gives the following expression:

$$T_{OFF}^{Eq} = T_{ON}^{Eq}.(1-P(>0))/P(>0) \quad (6)$$

The mean packet rate upward to BS is obtained by:

$$C_m = c.A = c.N.T_{on}/(T_{on}+T_{off})$$

It also holds:

$$C_m = P(>0).C_{ON}^{Eq} + (1-P(>0)).C_{OFF}^{Eq}$$

Therefore:

$$C_{OFF}^{Eq} = (C_m - P(>0).C_{ON}^{Eq})/(1-P(>0)) \quad (7)$$

The probability $P_{ER}$ of a packet to be an excess rate packet is:

$$P_{ER} = (C_{ON}^{Eq} - C)T_{ON}^{Eq} / C_m(T_{ON}^{Eq} + T_{OFF}^{Eq}) \quad (8)$$

After substituting of Eqs. (4) to (7) in Eqs. (3) and (8), we obtain $d$ and $P_{ER}$, and thus, we can determine the overflow probability *Q(x)*.

$$P_{ER} = cP(>0)/(C - C_m) \quad (9)$$

## IV. Numerical Results

We consider an IEEE 802.16 based backhaul network, in which time division duplexing mode (TDD) for data transmission is applied, and the physical layer is OFDM. Without loosing generality, we accept that all system bandwidth is allocated to voice services. The number of SSs connected to the BS is denoted as *S*, and for the purpose of the numerical experiment, an equal number of voice connections are associated with each of the SSs. We have also accepted the maximum packet delay and packet losses introduced by the IEEE 802.16 system to be 60 ms and 0.5 %, respectively [4].

In order to obtain reasonable quantitative results, we apply a PHY and MAC framework exactly as it is in [6]. The coding and modulation scheme accepted in [6] is BPSK modulation and channel coding rate of ½ at the PHY layer is the most reliable, but with fewer throughputs.

A bit sequence with a rate $R_{cod}$ from an active source is packed every $T_{cod}$ second into a voice packet, and thus, the time to transmit a voice packet is given by:

$$T_p = (R_{cod}T_{cod} + H_{head})/R_{BS} + T_{pre}$$

where $R_{BS}$ is the PHY transmission rate in bit/s, $H_{head}$ is the total packet header size at PHY and all upper layers and $T_{pre}$ is uplink burst preamble.

A fixed amount of time $T_{cont}$ in the UL subframe is allocated for contention-based transmission, initial ranging of SS connection and other functions. The time left is used for voice traffic (packet transmission and bandwidth request if applicable). For a single SS the time allowed to use per frame, depending on the scheduling scheme used, is given by:

$$T^{rtPS} = (T_{MAC} - T_{cont})/S - T_{BWrq}$$

$$T^{ertPS} = T^{UGS} = (T_{MAC} - T_{cont})/S$$

where $T_{BWrq}$ is the time for each request message.

The transmission capacity $C$ per SS, in dependence of the scheduling scheme used, is given by:

$$C^{rtPS} = \frac{T^{rtPS}}{T_p.T_{MAC}} \quad \text{and} \quad C^{ertPS} = C^{UGS} = \frac{T^{ertPS}}{T_p.T_{MAC}}$$

The packet rate $c$ per active voice source is $c = 1/T_{cod}$ for all scheduling services and thus the maximum number $n$ of simultaneous active connections is the following:

$$n^{UGS} = \lfloor C^{UGS}.T_{cod} \rfloor$$

$$n^{rtPS} = C^{rtPS}.T_{cod} \quad \text{and} \quad n^{ertPS} = C^{ertPS}.T_{cod}$$

It should be noted that $n^{UGS}$, $n^{rtPS}$ and $n^{ertPS}$ may not be integer values.

In polling services the packets are first stored in the SS buffer before the SS requests bandwidth. The resource request and grant process takes maximum 1 frame time and on average – half a frame. If the polling interval is 1 MAC frame for the rtPS, it holds $t_{MAC}^{rtPS} = 1½\,T_{MAC}$ for a substitution in Eq. (1). For the ertPS it holds $t_{MAC}^{ertPS} = ½\,T_{MAC}$. The buffer size, for a particular scheduling scheme used, is consequently given by the following expressions:

$$k^{rtPS} = t_{q,\max}^{rtPS}.C^{rtPS} \quad \text{and} \quad k^{ertPS} = t_{q,\max}^{ertPS}.C^{ertPS}$$

As we mentioned above, the coding and modulation scheme (CMS) applied is BPSK ½ and the main parameters for the analysis are: $R_{BS}$ = 6.91 Mbit/s; $T_{cont}$ = 312 μs; $T_{BWrq}$ = 27.78 μs; $T_{pre}$ = 11.11 μs; $H_{head}$ = 48 B; $T_{on}$ = 240 ms; $T_{off}$ . = 400 ms; $R_{cod}$ = 64 kb/s; $T_{cod}$ = 20 ms.

As a result of analysis, Fig. 2 depicts the packet loss probability $P_{PL}$ as a function of the number of voice connections *N* per SS, considering different polling services and coding rates. Results obtained from analytical research show the advantage of using ertPS polling service (which combines the simplicity of UGS and flexibility of the rtPS for supporting voice services with voice activity detection scheme), especially in case of low traffic load.

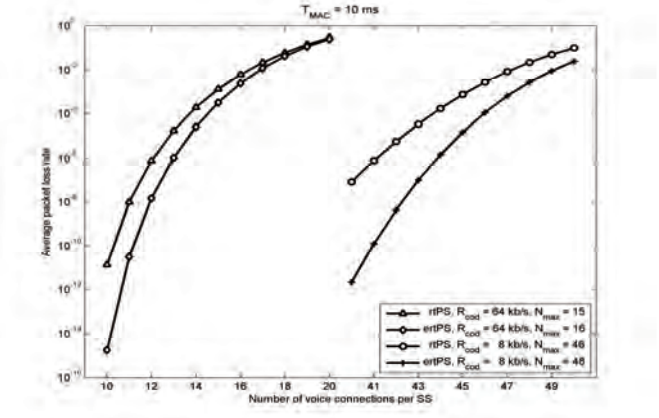


Fig. 2. Packet loss probability $P_{PL}$ as a function of the number of voice connections per SS, different polling services, and coding rates

The same figure also depicts the advantages of using the more effective voice coding algorithm, based on recommendation G.729. It is also of particular interest for us to investigate the influence of the MAC frame duration on the overall packet loss probability $P_{PL}$. The significance of the results of this research under normal traffic load conditions can be seen on Fig. 3.

Because signal strength of the radio spectrum allocated for data transmission in IEEE 802.16 networks falls off sharply with distance from the base station (BS), the signal-to-noise ratio drops with distance, as well. For this reason, IEEE 802.16 standard employs different coding and modulation schemes (CMS), depending on the distance between the SS and the BS. Based on this realization, the results of the evaluation of the number of admitted calls *N* for more realistic CMS than BPSK modulation, using more effective voice coding algorithm (G.729), are depicted on Fig. 4.

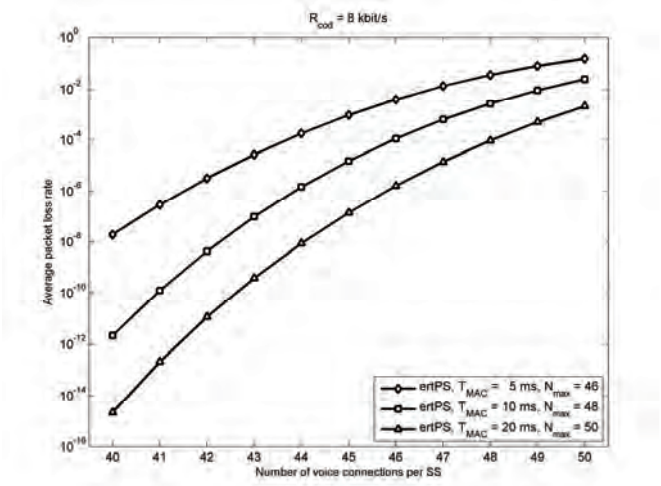


Fig. 3. Packet loss probability $P_{PL}$ as a function of the number of voice connections per SS and different MAC frames

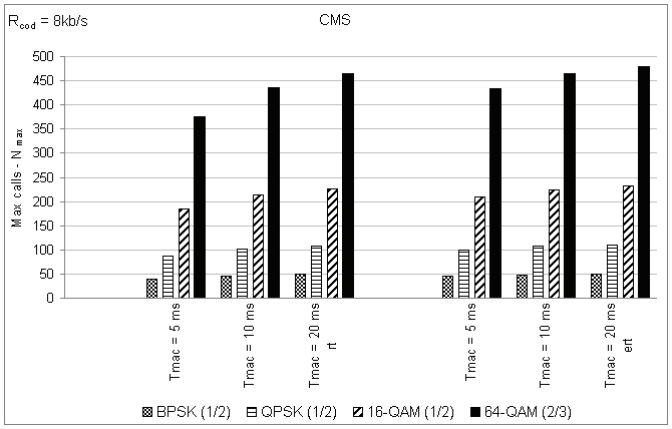


Fig. 4. Maximum admitted call numbers as a function of CMS

## V. Conclusion

The proposed analytical method is applicable for quick determination of the number of voice connections per SS, as the maximum admitted calls in a call admission control (CAC) procedure.

The authors intend to extend the model, taking into account some additional details, such as packet generation during the silence periods due to availability of so called comfort noise.

## References


[1] C. Cicconetti, A. Erta, L. Lenzini and E. Mingozzi, "Performance evaluation of the IEEE 802.16 MAC for QoS support," *IEEE Transactions on mobile computing*, vol.. 6, pp. 26-38, Jan. 2007.

[2] C Cicconetti, C. Eklund, L. Lenzini and E. Mingozzi, "Quality of Service Support in IEEE 802.16 networks," *IEEE Network, vol. 20, no. 2*, March/April 2006, pp.50-55.

[3] S.-E. Hong and O.-H. Kwon, "Considerations for VoIP services in IEEE 802.16 broadband access systems," *IEEE 63rd Vehicular Technology Conference, VTC 2006 Spring*, vol. 3, pp. 1296-1230.

[4] A. P. Makropolou, F. A. Tobagi and M. J. Karam, "Assessing the quality of voice communications over Internet backbone," *IEEE/ACM Transactions on Networking*, vol. 11, no. 5, Oct. 2003, pp. 747-760.

[5] S. Sengupta, M. Chatterjee, S. Ganguly and R. Izmailov, "Improving R-score of VoIP streams over WiMax", *IEEE International Conference on Communications (ICC),"* June 2006, pp. 1-6.

[6] D. Zhao and X. Shen, "Performance of packet voice transmission using IEEE 802.16 protocol," *IEEE Wireless Communications*, vol. 13, pp. 44-51, Feb. 2007.

[7] J. N. Daigle, *Queueing Theory with Applications to Packet Telecommunication*, Springer, 2005.

[8] J. Roberts, U. Mocci, and J.Virtano, *COST 242 Final Report of Action*, Springer, 1996

[9] J. A. Schormans, and J. M. Pitts, "Decay rate (ER) modeling of G/D/1 queue, and results for ATM telecommunications," *Electronics Letters*, vol. 34, no. 10, 14 May 1998, pp. 943-945.

[10] J. A. Schormans, J. M. Pitts, E. M. Scharf, A. J. Pearmain and C. I. Philips, "Buffer overflow probability for multiplexed on-off VoIP sources," *Electronics Letters*, vol. 36, no. 6, 16 March 2000 pp. 523-524.

[11] G. Fiche and G. Hebuterne, *Communicating Systems & Networks: Traffic & Performance*. Kogan Page Ltd, UK, 2004