

Bandwidth Allocation for Video Streaming in WiMax Networks

Alessandra Scicchitano
DEIS, Università della Calabria

Andrea Bianco, Carla-Fabiana Chiasserini, Emilio Leonardi
Dipartimento di Elettronica, Politecnico di Torino

Abstract—We describe an analytical model, based on a Markov chain, suitable to study different bandwidth allocation policies for video streams over a WiMax access link. The Markov chain models an MPEG source, wireless channel conditions derived from a model compliant with WiMax specifications, and different bandwidth allocation policies. Validation with simulation results shows the correctness of the analytical model. The model permits to discuss the properties of various bandwidth allocation policies in terms of wasted slots, amount of lost data and access delays.

I. INTRODUCTION

WiMax provides a wireless alternative to cable, such as ISDN or DSL. High data rates and low cost make this system the answer to the ever increasing demand for conventional types of multimedia services, such as digital video.

The air interface and medium access control (MAC) protocol of the WiMax technology have been specified by the IEEE 802.16 standard, specifically 802.16a and 802.16d. In this work, we focus on 802.16d wireless metropolitan area networks (WMANs) using the OFDM modulation. We consider the traffic transfer on the uplink direction in the Point-to-Multipoint (PMP) configuration, and study bandwidth allocation schemes to ensure QoS in WiMax networks. The 802.16 standard defines a fixed bandwidth allocation only for unsolicited grant service (UGS) flow, the remaining radio resources are devoted to real-time and non-real-time polling service (rtPS and nrtPS) and best effort (BE) flows. It does not specify, however, any policy for traffic scheduling at the base station (BS). Designing a scheduling scheme for uplink traffic that is able to reduce bandwidth waste while fulfilling the desired QoS, is thus a fundamental issue, especially for real-time traffic such as MPEG video.

Several proposals appeared in the literature [1], [2], [3], [4]. In particular, in [1] a scheme for the polling service is evaluated through a queueing analytical framework based on a discrete-time Markov chain. In [2], an efficient algorithm for scheduling uplink grants to subscriber stations (SS) with VoIP traffic is presented. Several research works, e.g., [3] or [4], propose schedulers, or even a hierarchy of schedulers, so that the BS can efficiently allocate radio resources for uplink.

In our paper, we analyze and compare various allocation schemes for MPEG video traffic, ranging from a fixed bandwidth allocation based only on average source rate and worst channel conditions, to fully dynamic schemes taking into account the current channel conditions and buffer occupation at the SSs. To analyze the different strategies, we develop an

analytical framework based on a discrete-time Markov chain, which models the wireless channel behavior and integrates the non-controlled MPEG video source proposed in [5]. We validate our analytical model through simulation.

Our work differs from the previous ones since it accounts for many aspects of the wireless system, such as channel conditions and transmission rate adaptation, besides including a detailed model of an MPEG source. Also, it presents an accurate and flexible analytical model, which can be used to assess the performance of various scheduling techniques.

II. SYSTEM DESCRIPTION

We consider one SS transmitting MPEG video traffic to a BS, using the TDMA/TDD. According to the 802.16 standard, time is divided into frames of duration T_f , each of which includes an uplink and downlink sub-frame.

In this work, we focus on traffic transmission on the uplink direction only. The 802.16 specifications define, as basic time unit, the physical slot (PS), which depends on the physical layer characteristics; n PSs form a mini-slot, which is used as the uplink bandwidth allocation unit. Bandwidth allocation on the uplink takes place as follows. Consider time frame m ; at the beginning of the uplink sub-frame, each SS wishing to transmit its traffic sends a bandwidth request (BW-Request) to the BS. An SS can transmit its BW-request through either contention mode or contention-free (polling) mode. In the contention mode, an SS sends a BW-request during a predefined contention period, and a back off mechanism is used to resolve contention among BW-requests from multiple SSs. On the contrary, in the contention-free mode, the BS polls each SS; upon receiving the polling message, the SSs respond by sending their BW-request. To notify the bandwidth assignments to the SSs, at the beginning of the $m + 1$ -th frame the BS broadcasts a UL-MAP message containing the so-called Information Elements (IE). The IEs indicate which mini-slots are reserved for the SS's transmission during the $m + 1$ -th uplink sub-frame.

We now describe the traffic source model, the wireless channel model and the bandwidth allocation strategies.

A. Traffic source

We consider a non-controlled MPEG video source, modeled through a discrete-time Markov chain as in [5].

The output of an MPEG encoder is a packet sequence determined by the group of picture (GoP) structure, which, in

the case of real-time transmission of MPEG video, is typically composed of six frames (*IBBPBB*). Each frame is quantized according to the desired quality of the reconstructed video and the bit rate of the traffic source. The step size employed by the quantizer is determined by the quantize scale parameter (qsp) – an integer value ranging from 1 to 31.

Following [5], we set the video frame duration to 40 ms and the size of the packets generated by the source to 576 bytes. Furthermore, we model the packet emission process at the encoder output as a Switched Batch Bernoulli Process (SBBP), whose parameters are determined by the following three elements: type of video frame, qsp value and activity level of the video sequence. Four video sequence activity levels are considered in [5]: very low, low, high, very high. We take the qsp as the input parameter and we obtain, as output of the SBBP, the probability density function (pdf) of the number v of packets generated by the source in a time interval of 40 ms.

B. Wireless channel

To model the channel conditions, we first simulated the SUI-3 channel model, which is compliant with the WiMax specifications. We set the delay spread equal to $0.5 \mu\text{s}$ and the Doppler maximal frequency parameter equal to 0.3 Hz. A Reed Solomon code with the following parameter setting is considered: $n=255$, $k=239$, $t=8$. An inner convolutional code is also implemented, with constraint length equal to 7 and coding rate that varies as a function of the employed modulation. We identify the channel state with the data rate, i.e., the combination of modulation and coding rate that is used by an SS during one time frame. Under the previously described scenario, we consider only five channel states out of the possible seven states, since the states corresponding to the two highest data rates never occur.

C. Bandwidth allocation strategies

We analyze various schemes for uplink bandwidth allocation in 802.16 networks. As already mentioned, bandwidth allocation is done on a time frame basis, and is expressed as the number of mini-slots assigned to an SS by the BS.

Let A denote the maximum value of bandwidth that can be granted to the SS in one uplink sub-frame. Recall that the bandwidth allocation computed at frame n is used for the SS's transmission at frame $n + 1$.

We start by considering two simple *fixed bandwidth allocations* (*FA*) policies, FA_1 and FA_2 , which are based on the average packet generation rate, \bar{v} , of the MPEG source. The former allocates as many mini-slots as needed to transmit \bar{v} packets per frame; the latter overestimates such an allocation by an over-allocation parameter, set to 20% in the final comparison among allocation policies.

We then study two *adaptive bandwidth allocations* (*AA*) strategies, AA_1 and AA_2 , that allocate as much bandwidth as needed to transmit all data units currently in the SS queue. To compute the required number of slots, the former assumes that the lowest data rate can be used (i.e., worst channel conditions), while the latter considers the transmission to take

place at the highest data rate (i.e., best channel conditions). As an example, note that the number of mini-slots needed to transmit a packet generated by the MPEG source is equal to 20 when the BPSK and coding rate 1/2 are used, while it is equal to 5 in the case of 16-QAM and coding rate 3/4.

Finally, two *fully adaptive bandwidth allocations* (*FAA*) are studied, FAA_1 and FAA_2 , which consider both buffer occupation and actual channel conditions. The former allocates exactly what is needed to transmit all buffer content at the data rate allowed by the current channel conditions, while the latter overestimates such allocation by an over-allocation parameter set to 10% in the final comparison among allocation policies.

III. SYSTEM MODEL AND ASSUMPTIONS

We describe the system dynamic using a discrete-time Markov chain (DTMC), in which the time is discretized into *time steps*. Each time step lasts four time frames, i.e., it corresponds to the MPEG video inter-frame generation period. The DTMC represents the following fundamental aspects of the system: (i) the MPEG traffic source, (ii) the bandwidth allocation strategy, (iii) the transmission buffer, and (iv) the channel state.

To develop our model, we make the following assumptions:

- the basic time unit is named *slot*, and it is composed by 5 mini-slots;
- the data unit D is set to 75 bytes; this corresponds to the greatest common divisor of the amounts of data bits that can be transmitted in one slot using any of the combinations of modulation and code rate;
- the SS stores data into a buffer of capacity equal to B data units; note that each packet generated by the MPEG source corresponds to the arrival of 7 data units at the SS's buffer;
- the channel may change status on a time step basis; this is justified by the relatively slow time evolution of the channel conditions, as reported by the channel traces we obtained;
- error-free packet transmissions.

The above assumptions will be validated by comparing analytical and simulation results.

A. The DTMC Model

The DTMC state space is defined by the following vector:

$$\bar{s} = (g, a, b, c) \quad (1)$$

where:

- g denotes the state of the MPEG source [5], which is a function of the specific frame (I, B or P) and of the specific activity level. g also specifies the pdf, $P_g(v)$, of the number v of packets generated in one time step. Note that $0 \leq v \leq V$, where V is the maximum number of packets generated by the source in one time step;
- a is the bandwidth allocated for the next SS's transmission, in number of slots; the allocation determined at time t is used for the transmission at time $t + 1$;

- b is the transmission buffer occupancy, in number of data units; $0 \leq b \leq B$;
- c denotes the channel state, i.e., the data rate in bits/s allowed by the channel conditions.

Let $P(\bar{s}_o, \bar{s}_d)$ be the probability that the chain moves in one time step from the origin state $\bar{s}_o = (g_o, a_o, b_o, c_o)$ to the destination state $\bar{s}_d = (g_d, a_d, b_d, c_d)$. To evaluate, the probabilities $P(\bar{s}_o, \bar{s}_d)$'s, we observe that the behavior of the traffic source and of the channel are independent of the rest of the system, while the buffer occupancy and the allocated bandwidth can be deterministically derived once the allocation scheme and the source and channel behavior are known. Hence, we write $P(\bar{s}_o, \bar{s}_d)$ as:

$$P(\bar{s}_o, \bar{s}_d) = P\{g: g_o \rightarrow g_d\} \cdot P\{c: c_o \rightarrow c_d\} \cdot P_{g_o}(v) \cdot \delta(\bar{s}_o, \bar{s}_d) \quad (2)$$

where

- $P\{g: g_o \rightarrow g_d\}$ is the probability that the traffic source changes from g_o to g_d ;
- $P\{c: c_o \rightarrow c_d\}$ is the probability that the channel state changes from c_o to c_d ;
- $P_{g_o}(v)$ is the probability that v packets are generated by the MPEG source in one time step, given that the source state is g_o ;
- $\delta(\bar{s}_o, \bar{s}_d)$ is an indicator function, which is equal to 1 if state \bar{s}_d is a possible successor of state \bar{s}_o , and equal to 0 otherwise.

Let us now describe how the values of b and a depend on the channel and source behavior, as well as on the chain current state. The number of data units in the SS queue in the destination state is calculated as:

$$b_d = \max\left(0, \min(b_o + 7 \cdot v, B) - \frac{c_d}{D}(a_o \cdot \text{slot_time})\right) \quad (3)$$

where b_o and a_o are the buffer occupation and bandwidth allocation in the origin state, respectively, c_d is the data rate in the destination state, slot_time is the slot duration and D is the data unit size. The adopted convention is the following: at the beginning of a time step, the generation of a new packet of 7 data units occurs first, followed by the removal of the data units from the buffer. This implies, for example, that the transition from state (g_o, a_o, B, c_o) to state $(g_d, a_d, B-1, c_d)$ corresponds to the successful transmission of one data unit and to the loss of v packets with probability $P_{g_o}(v)$ or it corresponds to the successful transmission of one packet and no data loss with probability $P_{g_o}(0)$.

As for the bandwidth allocation, this is computed according to the strategies described before. For instance, in the case of the AA₁ scheme, we have:

$$a_d = \min\left(\left\lceil \frac{Db_d}{c_1 \text{slot_time}} \right\rceil, A\right)$$

where b_d is the buffer occupation in the destination state and c_1 is the worst channel state. Recall that the slots allocated at time t will be used for transmission at time step $t+1$. Thus, even when fully adaptive strategies are considered, the

channel conditions at the time of the transmission (i.e., c_d) may be different from the ones considered when computing the bandwidth allocation. As an example, when channel conditions worsen, the slots allocated for an SS might be not enough to transmit all data units in the buffer.

The total number of states of the DTMC is $N = 24 \cdot A \cdot B \cdot 5$, where 24 is the number of states of the MPEG video source and 5 is the number of channel states. The steady-state probabilities $\Pi = \{\pi(\bar{s})\}$, with \bar{s} belonging to the state space, are computed by employing standard techniques. From the steady-state probabilities many interesting performance metrics can be derived, as shown in the following section.

B. Performance metrics

The performance of the schemes for bandwidth allocation can be expressed in terms of average buffer occupation, packet loss probability, average allocated bandwidth, and average delay. We compute these metrics as follows.

- Average transmission buffer occupancy: $E[b] = \sum_{\bar{s} \in \mathcal{S}} b\pi(\bar{s})$ where \mathcal{S} is the state space and b is the buffer occupancy in state \bar{s} ;
- Average data unit loss rate:

$$E[\lambda_{loss}] = \sum_{\bar{s} \in \mathcal{S}} \sum_{v=1}^V [\max(0, 7v - (B - b))] P_g(v)\pi(\bar{s})$$

where $7v - (B - b)$ is the number of data units that are lost due to buffer overflow, and the subscript g refers to the source being in state g ;

- Average bandwidth waste, i.e., number of allocated slots that are not used either because of improved channel conditions or for excessive bandwidth allocation:

$$E[a_{waste}] = \sum_{\mathcal{T}} \sum_{v=1}^V \left[a_o - \left\lceil \frac{D \min(b + 7v, B)}{c_d \cdot \text{slot_time}} \right\rceil \right] P_{g_o}(v) P(\bar{s}_o, \bar{s}_d) \pi(\bar{s}_o) \quad (4)$$

where \mathcal{T} is the set of transitions from \bar{s}_o to \bar{s}_d .

- Average data unit delay, i.e., the time interval from the time instant when a data unit arrives at the buffer till it is transmitted: $E[T] = E[t] + \bar{\tau}$. $E[t]$ represents the time interval from when a data unit is generated till the time step at which its transmission is scheduled; it is obtained by using Little law as follows:

$$E[t] = \frac{E[b]}{\sum_{\bar{s} \in \mathcal{S}} \sum_{v=1}^V 7v P_g(v)\pi(\bar{s}) - E[\lambda_{loss}]}$$

where the denominator represents the average rate at which data units enter the buffer. The term $\bar{\tau}$, instead, accounts for the delay due to the fact that not all of the transmitted data units are sent at the beginning of the time step, rather their transmission may take place in any of the four frames corresponding to the time step at which their transmission is scheduled. By assuming that at each frame of the time step one fourth of the scheduled data

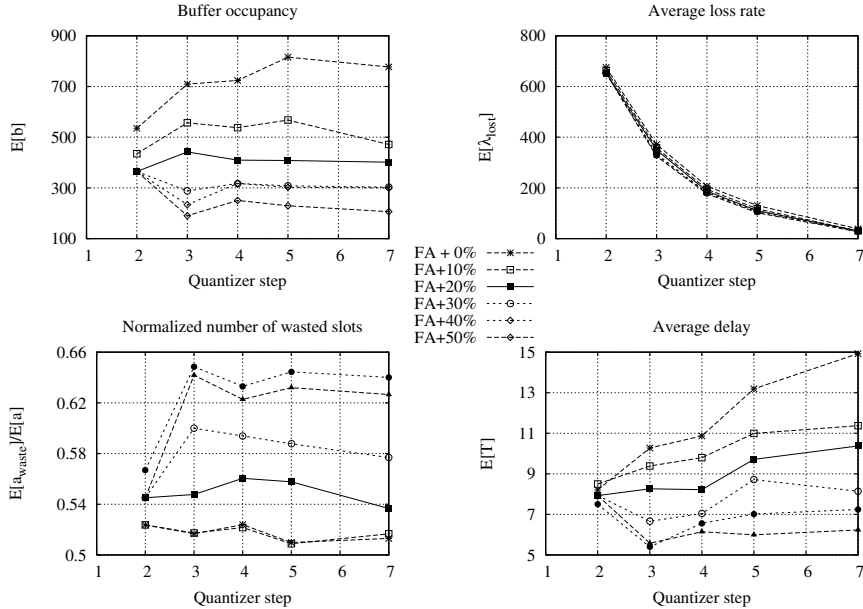


Fig. 1. Sensitivity of the fixed bandwidth allocation policy to the over-allocation parameter

units are transmitted and that such transmission occurs at the beginning of the frame, we can write:

$$\bar{\tau} = \sum_{\mathcal{T}} \sum_{j=0}^3 \delta(\beta_j) \cdot jT_f \cdot P(\bar{s}_o, \bar{s}_d) \pi(\bar{s}_o) \quad (5)$$

where T_f is the radio frame duration and β_j is the number of data units in the buffer at the beginning of the j -th frame of the time step; we have: $\delta(\beta_j) = 0$ if $\beta_j = 0$, i.e., if the buffer is empty, and 1 otherwise. We write β_j as the buffer occupancy at the beginning of the $j - 1$ th frame minus the number of data units already sent in frame $j - 1$, i.e.,

$$\beta_j = \begin{cases} \min(B, b_o + 7v) & j = 0 \\ \beta_{j-1} - \left(\frac{a_o}{4} \cdot \text{slot_time} \cdot \frac{ca}{D}\right) & \text{else} \end{cases}$$

IV. PERFORMANCE RESULTS

Here we compare the various allocation policies and validate the analytical model by simulation when one MPEG source is considered. We plot performance results as a function of the quantizer step of the single MPEG source. Smaller quantizer steps imply a higher source rate. The performance metrics we adopt are the average number of slots allocated in a time frame, the percentage of wasted slots, defined as the ratio between the number of wasted slots and allocated slots, the average data delay $E[T]$ and the average number of bytes lost.

We first analyze the sensitivity of the fixed and adaptive allocation policies to the over-allocation parameter, respectively in Fig. 1 and Fig. 2. In the case of fixed allocation, an over-allocation of 20% seems to provide the best compromise between delay and wasted slots. Indeed, the wasted slots do not increase for an over-allocation of 10%, whereas the delay is significantly reduced, specially for large quantizer

steps. Clearly increasing the over-allocation parameter further improves the delay, but not so significantly if considering the increase in the amount of wasted slots. Similar consideration hold for the adaptive policy, where an over-allocation of 10% provides the best compromise between reduced delay and increase in wasted slots.

In Fig. 3 we compare the various allocation policy and we validate the analytical model against simulation. Simulation runs exploit a proprietary simulation environment developed in C language. Statistical significance of the results is assessed by running experiments with an accuracy of 3% under a confidence interval of 95%. Simulation results are shown with dots, whereas lines refer to performance figures obtained by the analytical model. A very good agreement between simulation and analytical model is evident for any allocation policy. The most critical parameter is the amount of lost bytes, which is slightly overestimated by the analytical model.

Fixed rate policies are not able to deal with source burstiness. As such, the percentage of wasted slot with respect to the allocated one is likely unacceptable; obviously, losses and delay are very low. Wasted slots decrease for all adaptive policies, which show similar performance. The AA_2 policy, which allocates the slots according to the best channel conditions, reduces both the allocated and the wasted slots; however, delay performance are really unsatisfactory. The AA_1 policy, which allocates the slots according to the worst channel conditions, provides minor delay advantages at the cost of a non marginal increase in the percentage of wasted slots. The best policies are the two FAA, which exploit the knowledge of current channel status to request the allocated slots. No benefits are evident when applying a 10% over-allocation in this case.

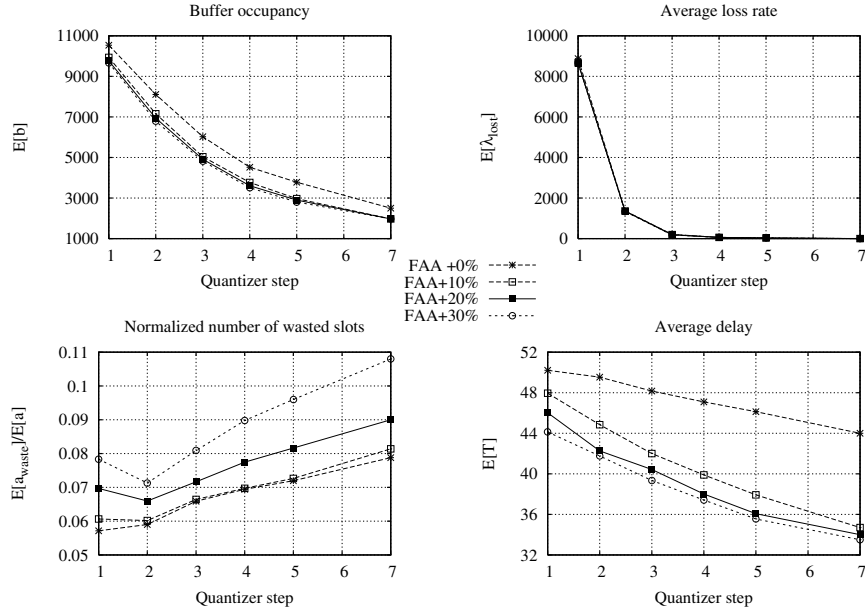


Fig. 2. Sensitivity of the adaptive bandwidth allocation policy to the over-allocation parameter

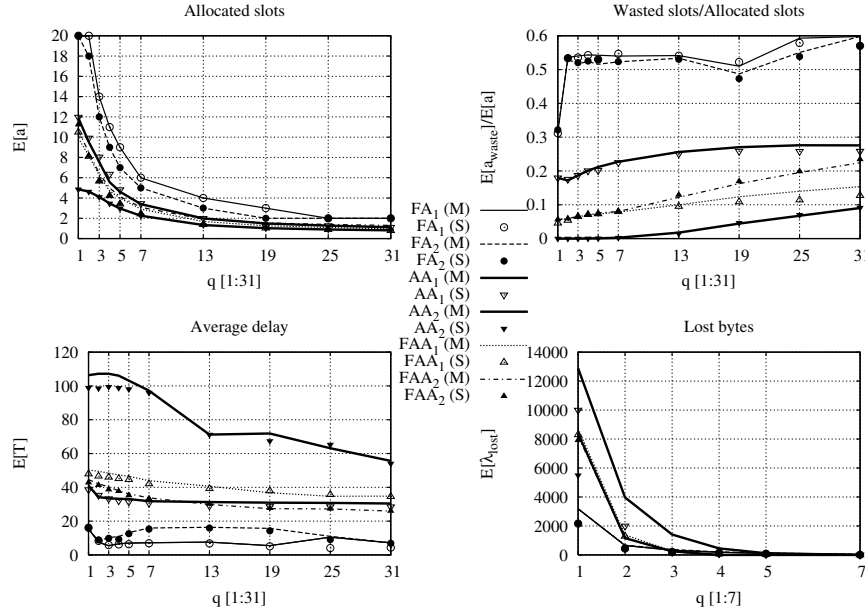


Fig. 3. Comparison among the various allocation policies

V. CONCLUSIONS

We describe an analytical model suitable to study the performance of bandwidth allocation algorithms for video streaming in WiMax networks. The model is validated against simulation. Among the proposed bandwidth allocation policies, the most promising seems to be the fully adaptive policy FAA_1 with no over-allocation, which provides the best compromise between wasted slots and access delay. We plan to extend the model to support more sources (currently we are able to solve the model with two MPEG sources) and to study performance results in a heterogeneous environment, i.e., with sources generating traffic at different rates and using

different allocation policies.

REFERENCES

- [1] D. Niyato, and E. Hossain, "Queue-Aware Uplink Bandwidth Allocation and Rate Control for Polling Service in IEEE 802.16 Broadband Wireless Networks", IEEE Trans. on Mobile Computing, 2006.
- [2] H. Lee, T. Kwon, and D.-H. Cho, "An Efficient Uplink Scheduling Algorithm for VoIP Services in IEEE 802.16 BWA Systems", IEEE VTC-Fall, 2004.
- [3] J. Chen, et al., "A service flow management strategy for IEEE 802.16 broadband wireless access systems in TDD mode", IEEE ICC 2005.
- [4] H. Zhang, "Service disciplines for guaranteed performance service in packet-switching networks", Proc. of IEE, 2005.
- [5] A. Lombardo, and G. Schembra, "Performance Evaluation of an Adaptive-Rate MPEG Encoder Matching IntServ Traffic Constraints", IEEE Trans. on Networking, 2003.