JOURNAL OF INDUSTRIAL AND
MANAGEMENT OPTIMIZATION
Volume **7**, Number **3**, August **2011**

doi:10.3934/jimo.2011.7.XX

pp. **X–XX**

ANALYSIS OF GLOBALLY GATED MARKOVIAN LIMITED CYCLIC POLLING MODEL AND ITS APPLICATION TO UPLINK TRAFFIC IN THE IEEE 802.16 NETWORK

Zsolt Saffer and Miklós Telek

Department of Telecommunications
Budapest University of Technology and Economics, Budapest, Hungary

Abstract. In this paper we introduce the *globally gated Markovian limited* service discipline in the cyclic polling model. Under this policy at most K customers are served during the server visit to a station among the customers that are present at the start of the actual polling cycle. Here the random limit K is the actual value of a finite state Markov chain assigned to the actual station. The model enables asymmetric Poisson arrival flows and each station has an individual Markov chain. This model is analyzed and the numerical solution for the mean of the stationary waiting time is provided.

This model is motivated by the problem of dynamic capacity allocation in Media Access Control of wireless communication networks with Time-Division Multiple Access mechanism. The "globally gated" character of the model is the consequence of the applied reservation mechanisms. In a fixed length frame after allocating the required capacity for the delay sensitive real-time traffic the random remaining capacity is shared among the subscriber stations for the non real-time traffic. The Markovian character of the random limits enables to model the inter frame dependencies of the required real-time capacity at each station individually.

In the second part of the paper the application of this model to the uplink traffic in the IEEE 802.16 network is discussed.

1. Introduction.

Polling models have been applied in the performance modeling of telecommunication systems from the beginning of 1980s. In the classical cyclic polling model the single server attends the stations in cyclic manner and the customer arrival process is Poisson at each station. Polling models are differentiated according to the service discipline, which determines the duration of the service at a station. The most common disciplines are the exhaustive, the gated and the G-limited disciplines. For the analysis of cyclic polling models we refer to the excellent book of Takagi [11].

In this paper we introduce a new service discipline for better modeling of a dynamic capacity allocation mechanism in Media Access Control (MAC) of wireless communication networks with Time-Division Multiple Access (TDMA) mechanism.

2000 *Mathematics Subject Classification.* Primary: 60K25, 68M20; Secondary: 90B22.

Key words and phrases. Queueing theory, polling model, waiting time, capacity allocation, IEEE 802.16.

This paper was presented at the QTNA2010 conference which was held in Beijing, China, during July 24-26, 2010. An earlier and brief version of this paper was published in the Conference Proceedings. The reviewing process of the paper was handled by Wuyi Yue and Yutaka Takahashi as Guest Editors.

Under the *globally gated Markovian limited* service policy at most K customers are served during the server visit to a station among the customers that are present at the start of the actual polling cycle. Here each station has an individual Markov chain and the random limit K of the considered station is determined by the help of the actual values of these finite state Markov chains. The "globally gated" character of the model is the consequence of the applied contention-free reservation mechanisms, since for every stations the capacity allocation is ensured only once in a cycle. After allocating the required capacity for the delay sensitive real-time traffic in a fixed length frame the random remaining capacity is shared among the stations for the non real-time traffic. This is realized by the dependency of the random limit K of a station on the actual values of the finite state Markov chains, which represent the capacity demands for the real-time traffic at the stations.

The principal goal of this paper is to introduce and to analyze the cyclic polling model with globally gated Markovian limited service policy and to show its application to IEEE 802.16 network [10].

Related works on delay analysis of IEEE 802.16 network are [5], [3] and [12]. In [1] an analytical model is established for the exact overall delay of the non real-time service flow with unicast polling in the IEEE 802.16 system. In contrast to these references the polling model presented in this paper enables to incorporate the effect of the real-time traffic capacity on the delay of the non real-time traffic.

The queueing theoretic contribution of this paper is the analysis and the results for the polling model with the newly introduced globally gated Markovian limited service discipline. For the analysis we use service discipline independent results from [6] and the numerical solution takes several elements from the computational procedure described in [8]. The model counts for the capacity allocation of both the real-time and non real-time traffic. The capacity allocation for the non real-time traffic of a station is dynamic in the dependency of the capacity needs of the real-time traffic at every stations. The model enables also priorities among the stations for their non real-time traffic flows. Furthermore the Markovian character of the random limits enables to model the inter frame dependencies of the required real-time capacity at each station individually.

We demonstrate the application of this polling model to the uplink non real-time traffic in the IEEE 802.16 network. It enables to study the effect of the mean and the maximum of the real-time capacity and the correlation of its consecutive values on the delay of the non real-time traffic. We also describe how to take into account an upper bound on mean delay in setting the mean or the maximum of the reserved capacity for the real-time traffic flows. Furthermore we introduce a cost model, which takes into account the Quality of Service (QoS) on delay constraint and the real-time capacity parameters. These tunings have potential applications in network control, since they facilitate the setting of the service flow parameters to the requirements of the actual application scenario.

The rest of this paper is organized as follows. In section 2 we introduce the model and the notations. The joint probabilities at different epochs are derived in section 3. The probability-generating function (PGF) of the stationary number of customers is given in section 4. The Laplace-Stieljes transform (LST) of the stationary waiting time and its mean are provided in section 5. In section 6 the details of the numerical solution are described. Section 7 closes the paper with the discussion of the application to IEEE 802.16 network.

2. Model and notation.

2.1. **The basic cyclic polling model.** We consider a continuous-time asymmetric polling model with N stations [11]. A single server attends the stations in cyclic manner and serves their infinite buffer queues during their visits. If no customer is present at a station at server arrival, the server leaves the station and attends the next station. At station i customers arrive according to Poisson arrival process with arrival rate λ_i for $i = 1, \ldots, N$. The customer who arrives to station i is called i-customer. The customer service time at station i is constant and it is the same for every $i = 1, \ldots, N$ and thus it is denoted by b. Random switchover time is enabled at switching from station i to the next one. The switchover times are integer multiple of the constant customer service time. Let R_i, for $i = 1, \ldots, N$, stand for the length of the switchover time after the service of station i in number of constant customer service times, i.e. the switchover time equals $R_i b$. The server utilization at station i and the overall utilization are $\rho_i = \lambda_i b$ and $\rho = \sum_{i=1}^{N} \rho_i$, respectively.

The *cycle time* of the system is defined as the time elapsed between the starts of two consecutive visits to station 1. The cycle time is also called as polling cycle. Let the length of the cycle time be fixed, which is denoted by c in number of constant customer service times, i.e. the cycle time is cb. The arrival of the server to a station and the departure of the server from a station are called *polling epoch* and *departure epoch*, respectively. We call the polling epoch of station i as i-polling epoch. Similarly the departure epoch of station i is an i-departure epoch. The *station time* of a given station is defined as the time elapsed from the arrival of the server to station i until its next departure. The station time of station i is called i-station time.

2.2. **Globally gated Markovian limited service discipline.** We introduce the *globally gated Markovian limited* service discipline, in which the service is both globally gated and limited as well as the random limit is determined from cycle to cycle on Markovian manner.

In the globally gated service (introduced by Boxma, Levy and Yechiali in [2]) only those i-customers can be served during a visit to station i that are present at the start of the cycle. Thus the starts of the cycles represent a global gate. Every i-customer arriving to the system after this epoch must wait until the start of the next cycle to get a service opportunity. Hence the start of the polling cycle we also call as *global gate epoch*. We refer to the start of the m-th cycle as m-th global gate epoch, for $m \geq 1$.

According to the limited service the number of i-customers that can be served during a server visit to station i is limited by a limit $K_i > 0$.

The random limit K_i is governed by background discrete-time Markov chains (DTMCs) for each $i = 1, \ldots, N$. Let $t_0^f(\ m)$ be the global gate epoch at the start of the m-th polling cycle, for $m \geq 1$. For each $i = 1, \ldots, N$ let $\{Y_i(t_0^f(\ m));\ m \in \{1, \ldots\}\}$ homogenous DTMC on the state space $\Omega = \{\omega_1, \ldots, \omega_L\}$, where $\omega_1, \ldots, \omega_L$ are positive integers. We call $\{Y_i(t_0^f(\ m));\ m \in \{1, \ldots\}\}$ the i-th background Markov chain. Let $u \leq c$ stand for a length of a fixed portion of the cycle time in number of constant customer service times. The values $Y_i(t_0^f(\ m))$, for $i = 1, \ldots, N$, are disjunct parts of u and $\sum_{i=1}^{N} Y_i(t_0^f(\ m)) < u$, for $m \geq 1$. Let $K_i(\ m)$ be the random limit in the m-th cycle. It is determined as a function of the values of the background Markov chains as follows

$$K_i(\ m) = \left\lfloor \zeta_i(u - \sum_{j=1}^{N} Y_j(t_0^f(\ m))) \right\rfloor, \quad m \in \{1, \ldots\},$$

$$\zeta_i \geq 0, \quad i = 1, \ldots, N, \quad \text{and} \quad \sum_{i=1}^{N} \zeta_i = 1, \tag{1}$$

where $\lfloor d \rfloor$ stands for the integer part of d (flooring operation). In (1) the function $\lfloor\rfloor$ is necessary because the random limit $K_i(\ m)$ must be an integer as it represents number of i-customers. The random limits $K_i(\ m)$, for $i = 1, \ldots, N$, are disjunct parts of $u - \sum_{j=1}^{N} Y_j(t_0^f(\ m))$, for $\ m \geq 1$.

In the stationary analysis we use the limiting version of (1), which is given by

$$K_i = \left\lfloor \zeta_i(u - \sum_{j=1}^{N} Y_j) \right\rfloor, \ \zeta_i \geq 0, \quad i = 1, \ldots, N \ \text{ and } \sum_{i=1}^{N} \zeta_i = 1, \tag{2}$$

where $K_i = \lim^d{}_{m\to\infty} K_i(\ m)$ and $Y_j = \lim^d{}_{m\to\infty} Y_j(t_0^f(\ m))$ and $\lim^d$ stands for the convergence in distribution. It follows from (2) that the sum of all K_i-s is upper bounded, i.e. $\sum_{i=1}^{N} K_i \leq u - \sum_{i=1}^{N} Y_i$.

Y_i can represent a reserved capacity at station i from the total capacity u, where the capacity is in the number of constant customer service times. Thus $u - \sum_{j=1}^{N} Y_j$ is the total remaining capacity in the system, from which station i gets K_i according to its priority weight ζ_i. The non-covered capacity $c - u = T + D$ is reserved for model specific purpose, where T is typically used for protocol overhead purposes and D is reserved for the traffic in the reverse direction. Note that the capacity values c, u, T and D are not necessarily integers.

Note that in the *globally gated Markovian limited* service discipline the random K_i represents both capacity and limit on the number of i-customers that can be served. It follows that the constant customer service time assumption is crucial in this model, since otherwise the capacity K_i would represent a limit on the service time, which would lead to another service discipline.

2.3. **Globally gated Markovian limited cyclic polling model.** The globally gated Markovian limited cyclic polling model is a cyclic polling model in which the service discipline at each of the N stations is the globally gated Markovian limited one. Additionally in this model a *cycle setup time* is inserted between the global gate epoch and the start of the server visit to station 1. The length of the cycle setup time, in the number of constant customer service times, is denoted by R_0 and it is defined as

$$R_0 = T + \sum_{j=1}^{N} Y_j.$$

Thus the cycle setup time at the begin of the cycle represents T plus the total reserved capacity. The capacity remaining from the cycle time is divided among the random limits K_i, for $i = 1, \ldots, N$, the difference $u - \sum_{i=1}^{N}(Y_i + K_i)$ due to the flooring operation in (2) and the reserved capacity part D. The random limit K_i is a part of this remaining capacity allocated to station i according to its priority weight. For $m \geq 1$ let $G_i(m)$ stand for the number of served i-customers during the m-th cycle. Then the stationary number of served i-customers during a cycle is given as $G_i = \lim^d{}_{m\to\infty} G_i(m)$. In fact G_i is also the stationary length of the i-station time in the number of constant customer service times, since it equals the stationary number of served i-customers during a cycle. The switchover time R_i is defined as

$$\begin{aligned} &R_i = K_i - G_i, \quad i = 1, \dots, N-1, \\ &R_N \geq K_N - G_N + D. \end{aligned} \tag{3}$$

Thus for every $i = 1, \dots, N$ the switchover time R_i is an unused part of the random limit K_i. Note that besides of the unused capacity at station N the last switchover part (R_N) incorporates also the difference $u - \sum_{i=1}^{N}(Y_i + K_i)$ due to the flooring operation in (2) and the reserved capacity part D. The globally gated Markovian limited cyclic polling model is illustrated for the case of $R_N = K_N - G_N + D$ in Fig. 1.

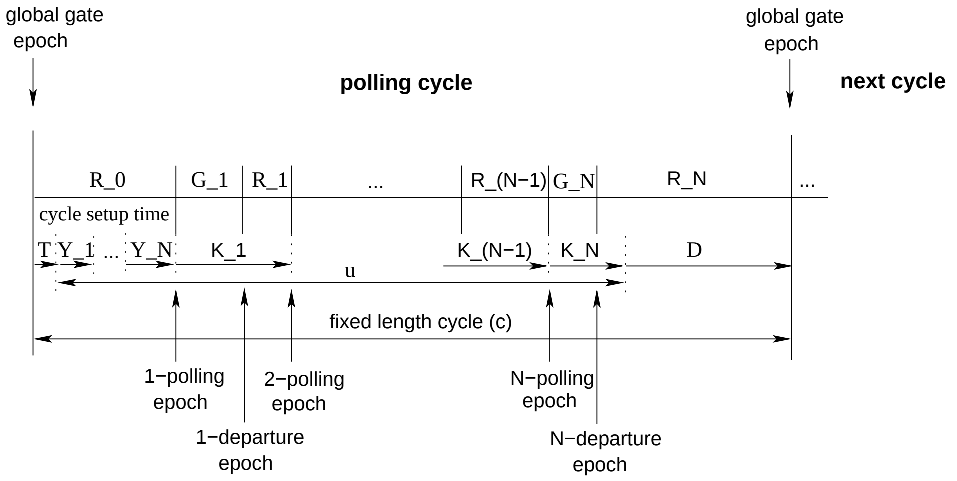

FIGURE 1. Globally gated Markovian limited cyclic model

Note that it follows from the expression (3) and from the model definition that the switchover time R_i is independent of the arrival processes during it for $i = 1, \dots, N-1$. Additionally on the globally gated Markovian limited cyclic polling model we impose the following assumptions:

A.1 For every $i = 1, \dots, N$ the i-th background Markov chain is irreducible. This ensures the existence of the limiting distributions of these Markov chains.

A.2 At each station the arrival rate and the customer service time is positive and finite, $0 < \lambda_i < \infty$, $0 < b < \infty$.

A.3 The arrival processes, and the service times and are mutually independent. Moreover the switchover time R_i is independent of the arrival processes during it for $i = 1, \dots, N$.

A.4 The following relation holds for each $i = 1, \dots, N$, for which $\zeta_i > 0$

$$\lfloor \zeta_i (u - N\omega_{max}) \rfloor \geq 1,$$

where ω_{max} is the maximal among the values $\omega_1, \ldots, \omega_L$.

This ensures that the average remaining capacity is at least one for each station, for which $\zeta_i > 0$, which implies that the traffic at these stations can not be blocked by the total reserved capacity.

A.5 The model is stable.

A.6 The queueing discipline is the First-In-First-Out (FIFO) order at each station.

2.4. **Stability of the model.** Due to the memoryless property of the Poisson arrival process and the fixed length cycle the number of arriving i-customers during a cycle is not dependent on the cycle index m. Thus let A_i stand for the number of arriving i-customers during a cycle. In addition we define $a_i = E[A_i]$, which can be interpreted as the mean (stationary) number of i-customers arriving during a cycle. Let $N_i(t)$ be the number of i-customers in the system at time t for $t \geq 0$ and $i = 1, \ldots, N$. We define the state vector $\mathbf{Z}(t_0^f(\ m))$ in the m-th global gate epoch as

$$\begin{aligned}\mathbf{Z}(t_0^f(\ m)) = \big(\ & N_1(t_0^f(\ m)), \ldots, N_N(t_0^f(\ m)), \\ & Y_1(t_0^f(\ m)), \ldots, Y_N(t_0^f(\ m))\big).\end{aligned} \tag{4}$$

$\mathbf{Z}(t_0^f(\ m))$ describes the state of the system at global gate epochs. It follows from the model definition that $\{\mathbf{Z}_i(t_0^f(\ m));\ m \in \{1, \ldots\}\}$ is a homogenous embedded Markov chain. Each component of $\mathbf{Z}(t_0^f(\ m))$ has either finite valued or countable infinite state space. This ensures that the stability analysis and results in [7] can be extended for this model. According to it the sufficient and necessary condition of the whole stability of the model is that, for each $i = 1, \ldots, N$, the mean stationary number of i-customers arrivals (a_i) must be less than the maximum of the mean number of i-customers, which can be served during an i-station time. This leads to

$$a_i < E[K_i] \quad \text{for every} \quad i = 1, \ldots, N. \tag{5}$$

Applying (2) in (5) leads to

$$a_i < E\left[\left\lfloor \zeta_i(u - \sum_{j=1}^{N} Y_j)\right\rfloor\right] \quad \text{for every} \quad i = 1, \ldots, N.$$

Using $a_i = \lambda_i bc = \rho_i c$ and rearranging results in the condition of the whole stability as

$$\rho_i < \frac{E\left[\left\lfloor \zeta_i(u - \sum_{j=1}^{N} Y_j)\right\rfloor\right]}{c} \quad \text{for every} \quad i = 1, \ldots, N. \tag{6}$$

3. **The stationary joint probabilities.** From now on $[\mathbf{Y}]_{j,l}$ stands for the j, l-th element of matrix $\mathbf{Y}$. Similarly $[\mathbf{y}]_j$ denotes the j-th element of vector $\mathbf{y}$. We define the transition probability matrix of the i-th background Markov chain, $\mathbf{\Pi}_i$, by its (j, y)-th element as

$$[\mathbf{\Pi}_i]_{j,y} = Pr\{Y_i(t_0^f(\ m+1)) = y \mid Y_i(t_0^f(\ m)) = j\},$$
$$j, y \in \Omega \ \ m \geq 1, \ \ i = 1, \ldots, N.$$

3.1. **The stationary joint probabilities at global gate epoch.** It follows from the model description that the number of i-customers together with the values of every background Markov chains at a global gate epoch determine the number of i-customers and the values of every background Markov chains at the next global gate epoch in stochastic sense. It means that for every $i = 1, \ldots, N$ the $\{(N_i(t_0^f(\ m)), Y_1(t_0^f(\ m)), \ldots, Y_N(t_0^f(\ m))); \ m \in \{1, \ldots\}\}$ is also a homogenous embedded Markov chain. Thus it is enough to establish relations among the joint probabilities of the components of $(N_i(t_0^f(\ m)), Y_1(t_0^f(\ m)), \ldots, Y_N(t_0^f(\ m)))$ instead of relating joint probabilities of every components of the state vector $\mathbf{Z}(t_0^f(\ m))$.

We define the joint probabilities of the stationary number of i-customers and the values of every background Markov chains at the global gate epoch as

$$\begin{aligned} &p_{0,i}^f(n_i, y_1, \ldots, y_N) = \lim_{m \to \infty} Pr\big\{N_i(t_0^f(\ m)) = n_i, \\ &Y_1(t_0^f(\ m)) = y_1, \ldots, Y_N(t_0^f(\ m)) = y_N\big\}, \\ &n_i \in \{0, 1, \ldots\}, \ \ y_1, \ldots, y_N \in \Omega, \ \ i = 1, \ldots, N. \end{aligned} \tag{7}$$

Theorem 3.1. *The relations among the joint probabilities of the stationary number of i-customers and the values of every background Markov chains at the global gate epoch are given as*

$$\begin{aligned} &p_{0,i}^f(n_i, y_1, \ldots, y_N) = \sum_{j_1 \in \Omega} \cdots \sum_{j_N \in \Omega} [\mathbf{\Pi}_1]_{j_1, y_1} \ldots [\mathbf{\Pi}_N]_{j_N, y_N} \\ &\sum_{k_i=0}^{n_i+K_i} p_{0,i}^f(k_i, j_1, \ldots, j_N) \frac{(\lambda_i bc)^{n_i - k_i + s_i}}{(n_i - k_i + s_i)!} e^{-\lambda_i bc}, \\ &s_i = min(k_i, K_i) \quad and \quad K_i = \left\lfloor \zeta_i (u - \sum_{\ell=1}^{N} j_\ell) \right\rfloor, \\ &n_i \in \{0, 1, \ldots\}, \ \ y_1, \ldots, y_N \in \Omega, \ \ i = 1, \ldots, N, \end{aligned} \tag{8}$$

where $min(k_i, K_i)$ stands for the smallest value of a set (k_i, K_i).

Proof. Assuming that k_i i-customers are present at the actual global gate epoch, the number of remaining i-customers after the next service is $k_i - s_i$, where $s_i = min(k_i, K_i)$ is the actual value of the number of i-customers served during a cycle. The number of i-customers at the next global gate epoch is n_i, therefore the number of i-customers arriving during a cycle is $n_i - k_i + s_i \geq 0$. Due to the fix cycle length c, this has the probability $\frac{(\lambda_i bc)^{n_i - k_i + s_i}}{(n_i - k_i + s_i)!} e^{-\lambda_i bc}$. Additionally $n_i - k_i + s_i \geq 0$ implies that $k_i \leq n_i + s_i \leq n_i + K_i$. Thus k_i must be summed up to $n_i + K_i$. Putting all these together leads to

$$\sum_{k_i=0}^{n_i+K_i} p_{0,i}^f(k_i, j_1, \ldots, j_N) \frac{(\lambda_i bc)^{n_i - k_i + s_i}}{(n_i - k_i + s_i)!} e^{-\lambda_i bc}.$$

Using it the theorem comes by taking also into account the probabilities of every possible transitions of every background Markov chains to states y_ℓ, for every $\ell = 1, \dots, N$. □

Relations (8) defines a system of linear equations for computing the joint probabilities $p^f_{0,i}(n_i, y_1, \dots, y_N)$ for $n_i \in \{0, 1, \dots\}$, $y_1, \dots, y_N \in \Omega$, $i = 1, \dots, N$.

3.2. **The stationary joint probabilities at polling and departure epochs.** For $m \geq 1$ let $t^f_i(m)$ and $t^m_i(m)$ be the i-polling and the i-departure epoch in the m-th polling cycle, respectively. We define the joint probabilities of the stationary number of i-customers and the values of every background Markov chains at the i-polling epoch as

$$\begin{aligned} &p^f_i(n_i, y_1, \dots, y_N) = \lim_{m \to \infty} Pr\big\{N_i(t^f_i(m)) = n_i,\\ &Y_1(t^f_i(m)) = y_1, \dots, Y_N(t^f_i(m)) = y_N\big\},\\ &n_i \in \{0, 1, \dots\}, \quad y_1, \dots, y_N \in \Omega, \quad i = 1, \dots, N. \end{aligned} \tag{9}$$

Similarly we define the joint probabilities of the stationary number of i-customers and the values of every background Markov chains at the i-departure epoch as

$$\begin{aligned} &p^m_i(n_i, y_1, \dots, y_N) = \lim_{m \to \infty} Pr\big\{N_i(t^m_i(m)) = n_i,\\ &Y_1(t^m_i(m)) = y_1, \dots, Y_N(t^m_i(m)) = y_N\big\},\\ &n_i \in \{0, 1, \dots\}, \quad y_1, \dots, y_N \in \Omega, \quad i = 1, \dots, N. \end{aligned} \tag{10}$$

Theorem 3.2. *The joint probabilities at the i-polling and i-departure epochs defined in (9) and (10) can be expressed by the joint probabilities at global gate epoch defined in (7) as*

$$\begin{aligned} &p^f_i(n_i, y_1, \dots, y_N) = \sum_{k_i=0}^{n_i} p^f_{0,i}(k_i, y_1, \dots, y_N)\\ &\frac{(\lambda_i b(R_0 + \sum_{\ell=1}^{i-1} K_\ell))^{n_i - k_i}}{(n_i - k_i)!} e^{-\lambda_i b(R_0 + \sum_{\ell=1}^{i-1} K_\ell)},\\ &p^m_i(n_i, y_1, \dots, y_N) = \sum_{k_i=0}^{n_i + K_i} p^f_{0,i}(k_i, y_1, \dots, y_N)\\ &\frac{(\lambda_i b(R_0 + \sum_{\ell=1}^{i-1} K_\ell + s_i))^{n_i - k_i + s_i}}{(n_i - k_i + s_i)!} e^{-\lambda_i b(R_0 + \sum_{\ell=1}^{i-1} K_\ell + s_i)},\\ &R_0 = T + \sum_{\ell=1}^{N} y_\ell, \quad K_i = \left\lfloor \zeta_i(u - \sum_{\ell=1}^{N} y_\ell) \right\rfloor, \quad s_i = min(k_i, K_i)\\ &n_i \in \{0, 1, \dots\}, \quad y_1, \dots, y_N \in \Omega, \quad i = 1, \dots, N. \end{aligned} \tag{11}$$

Proof. Assuming that k_i i-customers are present at the global gate epoch and the number of i-customers at the next i-polling epoch is n_i, it follows that the number of i-customers arriving in between is $n_i - k_i \geq 0$. The length of this interval is $R_0 + \sum_{\ell=1}^{i-1} K_\ell$. As this length does not depend on the Poisson arrivals during it, the probability that $n_i - k_i$ i-customers arrive during this interval is $\frac{(\lambda_i b(R_0+\sum_{\ell=1}^{i-1} K_\ell))^{n_i-k_i}}{(n_i-k_i)!} e^{-\lambda_i b(R_0+\sum_{\ell=1}^{i-1} K_\ell)}$. Additionally $n_i - k_i \geq 0$ implies that $k_i \leq n_i$, thus k_i must be summed up to n_i. Putting all these together gives the first relation of (11).

Similarly assuming that k_i i-customers are present at the global gate epoch, $k_i - s_i$ i-customers remains among them at the next i-departure epoch. The number of i-customers at this i-departure epoch is n_i, therefore the number of i-customers arriving in between is $n_i - k_i + s_i \geq 0$. As the length of the interval from the global gate epoch to the next i-departure epoch does not depend on the Poisson arrivals during it, the probability that $n_i - k_i + s_i$ i-customers arrive during this interval is $\frac{(\lambda_i b(R_0+\sum_{\ell=1}^{i-1} K_\ell+s_i))^{n_i-k_i+s_i}}{(n_i-k_i+s_i)!} e^{-\lambda_i b(R_0+\sum_{\ell=1}^{i-1} K_\ell+s_i)}$. Additionally $n_i - k_i + s_i \geq 0$ implies that $k_i \leq n_i + s_i \leq n_i + K_i$. Thus k_i must be summed up to $n_i + K_i$. Putting all these together results in the second relation of (11). □

4. **The stationary number of customers.** Let $p_i^f(n_i)$ and $p_i^m(n_i)$ denote the probabilities of the stationary number of i-customers at i-polling and i-departure epochs, respectively. They can be calculated from the joint probabilities $p_i^f(n_i, y_1, \ldots, y_N)$ and $p_i^m(n_i, y_1, \ldots, y_N)$ as

$$\begin{aligned} p_i^f(n_i) &= \sum_{y_1 \in \Omega} \ldots \sum_{y_N \in \Omega} p_i^f(n_i, y_1, \ldots, y_N), \\ p_i^m(n_i) &= \sum_{y_1 \in \Omega} \ldots \sum_{y_N \in \Omega} p_i^m(n_i, y_1, \ldots, y_N), \\ & \qquad n_i \in \{0, 1, \ldots\}, \quad i = 1, \ldots, N. \end{aligned} \tag{12}$$

Based on these quantities we define the PGFs of the stationary number of customers at i-polling and i-departure epochs as

$$\begin{aligned} \widehat{F}_i(z) &= \sum_{n=0}^{\infty} p_i^f(n) z^n, \\ \widehat{M}_i(z) &= \sum_{n=0}^{\infty} p_i^m(n) z^n, \quad |z| \leq 1, \quad i = 1, \ldots, N. \end{aligned}$$

Furthermore we define the PGF of the stationary number of customers at an arbitrary instant as

$$\widehat{Q}_i(z) = \lim_{t \to \infty} \sum_{n=0}^{\infty} Pr\{N_i(t) = n\} z^n, \quad |z| \leq 1, \quad i = 1, \ldots, N.$$

Let f_i and m_i stand for the means of the stationary number of i-customers at i-polling epoch and at i-departure epoch, respectively.

Theorem 4.1. *The PGF of the stationary number of i-customers at a random instant is given by*

$$\begin{aligned} \widehat{Q}_i(z) \quad = \quad & \frac{(1-\rho_i)\,(1-z)\,e^{-(\lambda_i - \lambda_i z)b}}{e^{-(\lambda_i - \lambda_i z)b} - z} \\ & \cdot \frac{\widehat{M}_i(z) - \widehat{F}_i(z)}{(f_i - m_i)\,(1-z)}. \end{aligned} \tag{13}$$

Proof. Since the state space of the multidimensional embedded Markov chain describing the state of the system at the global gate epochs $\{\mathbf{Z}_i(t_0^f(\ m));\ m \in \{1, \ldots\}\}$ is countable, the Markov regenerative process (MRP) framework in [6] can be extended to the globally gated Markovian limited cyclic polling model. The statement (13) is proven in [6] for the classical cyclic polling model for the case of more general customer service times. As the assumptions used for the proof hold also for the globally gated Markovian limited cyclic polling model, the statement holds also for this model. □

5. The stationary waiting time.

The waiting time of an i-customer is defined as the time elapsed from the arrival of the i-customer to the start of its service. Let $W_{i,\ell}$ denote the waiting time of the i-customer that arrives as the ℓ-th into the system, $\ell \geq 1$. We define the cumulated distribution function of the stationary waiting time of an i-customer, $W_i(t)$, as

$$W_i(t) = \lim_{\ell \to \infty} Pr\{W_{i,\ell} \leq t\}, \quad t \geq 0, \quad i = 1, \ldots, N.$$

The LST of the stationary waiting time of an i-customer is defined as

$$\widetilde{W}_i(s) = \int_{t=0}^{\infty} e^{-st} dW_i(t), \quad Re(s) \geq 0, \quad i = 1, \ldots, N.$$

Theorem 5.1. *The LST of the stationary waiting time of an i-customer is given by*

$$\begin{aligned} \widetilde{W}_i(s) \quad &= \quad \frac{s\,(1-\rho_i)}{s - \lambda_i + \lambda_i e^{-sb}} \\ &\quad \cdot \frac{\widehat{M}_i\left(1 - \frac{s}{\lambda_i}\right) - \widehat{F}_i\left(1 - \frac{s}{\lambda_i}\right)}{\frac{s}{\lambda_i}\,(f_i - m_i)}. \end{aligned} \tag{14}$$

Proof. Due to the FIFO queueing discipline the argument can be used that the number of i-customers left in the system at service completion of a tagged i-customers equals with the number of i-customers arrived during the sojourn time of that i-customer in the system. Due to the model assumptions a new arriving i-customers do not affect the time in the system of the previously arrived i-customers, i.e. their waiting and service time. Additionally the waiting time of an i-customer and its service time are independent. Using the above argument it is shown in [6] that under these conditions (14) can be derived from (13). It follows that (14) holds also in the globally gated Markovian limited cyclic polling model. □

Let $f_i^{(2)}$ and $m_i^{(2)}$ stand for the second factorial moments of the stationary number of i-customers at i-polling epoch and at i-departure epoch, respectively.

Corollary 1. *The mean stationary waiting time of an i-customer is given by*

$$E[W_i] = \frac{\lambda_i b^2}{2(1-\rho_i)} + \frac{f_i^{(2)} - m_i^{(2)}}{2\lambda_i (f_i - m_i)} \tag{15}$$

Proof. (15) can be derived from (14). □

6. The numerical solution.

6.1. Computation of the joint probabilities.

To keep the computation of the joint probabilities in relations (8) tractable, we apply an upper limit $n_i \leq X_i$ on the number of i-customers, which results in finite number of unknowns and equations in the system of linear equations. This technique is similar to the one used in [8]. An appropriate value of X_i depends on the required precision level. In an iterative realization X_i is increased until the difference of consecutive values of $p^f_{0,i}(n_i, y_1, \dots, y_N)$ becomes less than the allowed error. In the computational steps the probabilities $p^f_{0,i}(n_i, y_1, \dots, y_N)$ for $n_i > X_i$ can be neglected, therefore these probabilities are set 0. This leads to

$$\begin{aligned}
&p^f_{0,i}(n_i, y_1, \dots, y_N) = \sum_{j_1 \in \Omega} \cdots \sum_{j_N \in \Omega} [\mathbf{\Pi}_1]_{j_1,y_1} \dots [\mathbf{\Pi}_N]_{j_N,y_N} \\
&\sum_{k_i=0}^{min(n_i+K_i,X_i)} p^f_{0,i}(k_i, j_1, \dots, j_N) \frac{(\lambda_i bc)^{n_i-k_i+s_i}}{(n_i-k_i+s_i)!} e^{-\lambda_i bc}, \\
&s_i = min(k_i, K_i) \quad \text{and} \quad K_i = \left\lfloor \zeta_i(u - \sum_{\ell=1}^{N} j_\ell) \right\rfloor, \\
&n_i \in \{0, 1, \dots X_i\}, \quad y_1, \dots, y_N \in \Omega, \quad i = 1, \dots, N.
\end{aligned} \tag{16}$$

Similarly setting the same upper limit $n_i \leq X_i$ on the number of i-customers in equations (11) leads to the computation of the joint probabilities at i-polling and i-departure epoch as

$$\begin{aligned}
&p^f_i(n_i, y_1, \dots, y_N) = \sum_{k_i=0}^{min(n_i,X_i)} p^f_{0,i}(k_i, y_1, \dots, y_N) \\
&\frac{(\lambda_i b(R_0 + \sum_{\ell=1}^{i-1} K_\ell))^{n_i-k_i}}{(n_i-k_i)!} e^{-\lambda_i b(R_0+\sum_{\ell=1}^{i-1} K_\ell)}, \\
&p^m_i(n_i, y_1, \dots, y_N) = \sum_{k_i=0}^{min(n_i+K_i,X_i)} p^f_{0,i}(k_i, y_1, \dots, y_N) \\
&\frac{(\lambda_i b(R_0 + \sum_{\ell=1}^{i-1} K_\ell + s_i))^{n_i-k_i+s_i}}{(n_i-k_i+s_i)!} e^{-\lambda_i b(R_0+\sum_{\ell=1}^{i-1} K_\ell+s_i)}, \\
&R_0 = T + \sum_{\ell=1}^{N} y_\ell, \quad K_i = \left\lfloor \zeta_i(u - \sum_{\ell=1}^{N} y_\ell) \right\rfloor, \quad s_i = min(k_i, K_i) \\
&n_i \in \{0, 1, \dots X_i\}, \quad y_1, \dots, y_N \in \Omega, \quad i = 1, \dots, N.
\end{aligned} \tag{17}$$

6.2. The steps of the numerical procedure.

The computation of the first moment of the stationary waiting time of an i-customer consists of several steps.

1. Build up a matrix form system of linear equations for computation of the joint probabilities at global gate epoch.

   The system of linear equation (16) is rearranged into a matrix form. Let $\mathbf{e}^{X_i+1}_\ell = (0, \dots, 0, 1, 0, \dots, 0)$ denote the $1 \times (X_i + 1)$ vector with 1 at the ℓ-th position. Furthermore $\otimes$ stands for the Kronecker product. We define the

$1 \times L^N(X_i+1)$ vector $\boldsymbol{\theta}_i$, representing the unknowns of the above system of linear equations as

$$\begin{aligned}\boldsymbol{\theta}_i = & \sum_{y_1 \in \Omega} \ldots \sum_{y_N \in \Omega} \sum_{n_i=0}^{X_i} p^f_{0,i}(n_i, y_1, \ldots, y_N) \\ & \mathbf{e}^L_{\mathcal{I}(y_1)} \ldots \otimes \mathbf{e}^L_{\mathcal{I}(y_N)} \otimes \mathbf{e}^{X_i+1}_{n_i+1} \quad i = 1, \ldots, N,\end{aligned} \tag{18}$$

where $\mathcal{I}(y_1)$ denotes the index of y_1 in the set Ω, which can be $1, \ldots, L$. Note that each element of $\boldsymbol{\theta}_i$ is a probability. We also introduce the $L^N(X_i+1) \times L^N(X_i+1)$ matrix $\boldsymbol{\Upsilon}_i$ representing the coefficients on the right-hand side (r.h.s.) of the equation (16). It is defined as

$$\begin{aligned}\boldsymbol{\Upsilon}_i = & \sum_{j_1 \in \Omega} \ldots \sum_{j_N \in \Omega} \sum_{k_i=0}^{X_i} \sum_{y_1 \in \Omega} \ldots \sum_{y_N \in \Omega} \sum_{n_i=0}^{X_i} \\ & I_{(k_i \leq min(n_i+K_i, X_i))} [\boldsymbol{\Pi}_1]_{j_1, y_1} \ldots [\boldsymbol{\Pi}_N]_{j_N, y_N} \frac{(\lambda_i bc)^{n_i - k_i + s_i}}{(n_i - k_i + s_i)!} e^{-\lambda_i bc} \\ & \left(\mathbf{e}^L_{\mathcal{I}(j_1)} \ldots \otimes \mathbf{e}^L_{\mathcal{I}(j_N)} \otimes \mathbf{e}^{X_i+1}_{k_i+1} \right)^T \\ & \left(\mathbf{e}^L_{\mathcal{I}(y_1)} \ldots \otimes \mathbf{e}^L_{\mathcal{I}(y_N)} \otimes \mathbf{e}^{X_i+1}_{n_i+1} \right) \; i = 1, \ldots, N,\end{aligned} \tag{19}$$

where $I_{(\text{con})}$ denote the indicator of condition "con". In this matrix the values of j_1, …, j_N, k_i and the values of y_1, …, y_N, n_i specify the row and the column indices of the corresponding coefficient. Using definitions (18) and (19) the matrix form of the system of linear equation (16) can be given as

$$\boldsymbol{\theta}_i = \boldsymbol{\theta}_i \boldsymbol{\Upsilon}_i, \quad i = 1, \ldots, N. \tag{20}$$

The sums on the r.h.s. of (16) realize the product of vector $\boldsymbol{\theta}_i$ by matrix $\boldsymbol{\Upsilon}_i$ in (20). The normalization condition for the probabilities in vector $\boldsymbol{\theta}_i$ can be given as

$$\boldsymbol{\theta}_i \mathbf{e}^{L^N(X_i+1)} = 1, \quad i = 1, \ldots, N, \tag{21}$$

where $\mathbf{e}^{L^N(X_i+1)}$ denotes the $(L^N(X_i+1)) \times 1$ column vector having all elements equal to one.

Matrix $\boldsymbol{\Upsilon}_i$ in (20) relates the probabilities of vector $\boldsymbol{\theta}_i$ and hence it can be interpreted as the transition probability matrix of the $\{(N_i(t^f_0(\ m)), Y_1(t^f_0(\ m)), \ldots, Y_N(t^f_0(\ m)));\ m \in \{1, \ldots\}\}$ embedded Markov chain, whose unique limiting distribution exists due to the stability of the model. It follows that the joint probabilities $p^f_{0,i}(n_i, y_1, \ldots, y_N)$ for $n_i \in \{0, 1, \ldots X_i\}, \;\; y_1, \ldots, y_N \in \Omega, \;\; i = 1, \ldots, N$ can be uniquely determined from the system of linear equation (20) and (21).

2. Solving the matrix form system of linear equation (20) and (21) for the joint probabilities at global gate epoch for every $i = 1, \ldots, N$.
3. Calculation of the joint probabilities at i-polling and i-departure epochs from the joint probabilities at global gate epoch by using equations (17).
4. Computation of the probabilities $p^f_i(n_i)$ and $p^m_i(n_i)$ for $n_i \in \{0, 1, \ldots X_i\}$ by using (12) for every $i = 1, \ldots, N$.

5. Calculation of the factorial moments of the stationary number of i-customers at i-polling and i-departure epochs (f_i, $f_i^{(2)}$, m_i, $m_i^{(2)}$) from the probabilities $p_i^f(n_i)$ and $p_i^m(n_i)$ for $n \in \{0, 1, \ldots X_i\}$ on elementary way for every $i = 1, \ldots, N$.
6. Computation of first moment of the stationary waiting time of an i-customers from the factorial moments f_i, $f_i^{(2)}$, m_i, $m_i^{(2)}$ by applying formula (15) for every $i = 1, \ldots, N$.

6.3. **Numerical complexity.** The most computational intensive parts of the procedure is the solution of the system of linear equations (20) and (21). The number of equations and unknowns in these system of linear equations for all stations is $L^N \sum_{i=1}^{N}(X_i + 1)$. Thus the total number of operations required by the whole numerical procedure is in the magnitude of $L^{3N} \sum_{i=1}^{N}(X_i + 1)^3$.

Therefore the total number of required elementary computational steps increases with the number of stations (N), with the number of states of the background Markov chains (L) and with X_i-s.

7. Application to the IEEE 802.16 network.

7.1. **Overview of IEEE 802.16.** In this subsection we give a brief summary on the basic characteristics of the IEEE 802.16.

7.1.1. *Point-to-multipoint operational mode.* The IEEE 802.16 standard supports the mandatory point-to-multipoint (PMP) and the optional mesh mode. In the centralized PMP IEEE 802.16 architecture there are one Base Station (BS) and one or more Subscriber Stations (SSs). The packets are exchanged between BS and SSs via separate channels. A DownLink (DL) channel is used for the traffic from the BS to the SSs and the UpLink (UL) channel is used in the reverse direction.

7.1.2. *Channel allocation schemes.* The standard defines two mechanisms of multiplexing DL and UL channels: Time Division Duplex (TDD) and Frequency Division Duplex (FDD). In FDD the the DL and the UL channels are assigned to different sub-band frequencies. In TDD mode the channels are differentiated by assigning different time intervals to them, i.e. MAC frame is divided into the DL sub-frame and the UL sub-frame. The border between these parts may change dynamically depending on the SSs bandwidth requirements. The SSs access the UL channel by means of TDMA. The structure of the MAC frame in TDD/TDMA mode is shown in Figure 2.

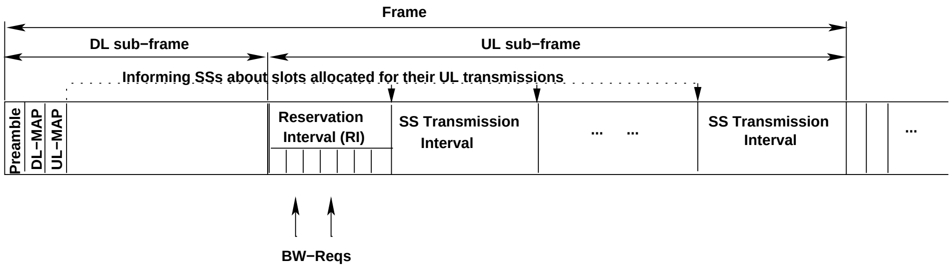

Figure 2. IEEE 802.16 MAC frame structure in TDD/TDMA mode.

In the DL channel the BS - as the only sending station - broadcasts the packets to all the SSs. Together with the data packets, the BS also transmits service information about the slots which are allocated for each of the SSs in the UL channel. This information is incorporated in the UL-MAP message and is used by the SSs for scheduling their data packets in the UL channel of the next MAC frame. The access procedure to the UL channel is the subject to one of the specified multiple access protocols.

7.1.3. *Bandwidth reservation.* The BS also specifies a portion of channel resources as the Reservation Interval (RI), which is used by the SSs for transmitting their bandwidth requests (BW-Req), which are then processed by the BS. The access procedure of the SSs to the RI could be either contention-free or contention-based. The former is referred to as unicast polling and corresponds to the case when BS assigns to each of the SSs a transmission opportunity for its bandwidth request. The latter consists of two mechanisms, namely, multicast and broadcast polling. When broadcast polling is enabled all the SSs are expected to send their bandwidth requests by choosing one of all the transmission opportunities uniformly. During the access to the RI collisions may occur, which may be subject to a subsequent resolution. The specified collision resolution algorithm is the truncated binary exponential backoff. In case of multicast polling the SSs are polled in groups and within a group the rules of broadcast polling are applied. Furthermore, IEEE 802.16 enables piggybacking for sending BW-Reqs attached to data packets.

7.1.4. *Service flow types.* To ensure QoS requirements for a variety of traffic types the IEEE 802.16 standard defines the following five service flow types:

1. Unsolicited Grant Service (UGS) is planned for the real-time traffic where fixed-size data packets are generated periodically such as T1/E1. No bandwidth reservation mechanism is used.
2. Real-Time Polling Service (rtPS) supports real-time traffic where variable-size data packets are generated periodically such as MPEG video. Unicast polling is used for bandwidth reservation.
3. Non Real-Time Polling Service (nrtPS) is offered for non real-time traffic where variable-size data packets are generated on a regular basis such as high bandwidth FTP. Both contention-free unicast polling and contention-based multicast and broadcast polling are allowed.
4. Best Effort (BE) is suitable for applications where no throughput or delay guarantee is provided, since it utilizes the remaining bandwidth after bandwidth allocation for the other service flows. Both reservation mechanisms are allowed.
5. Extended Real-Time Variable Rate (ERT-VR) is like rtPS but with more strict delay requirement (guaranteed jitter) to support real-time applications like VoIP with silence suppression. This class is often referred to as Extended Real-Time Polling Service (ertPS). It supports both contention-free polling and contention-based random access reservation mechanisms.

7.2. **Analytic model of the uplink nrtPS traffic in the IEEE 802.16 network.** The presented model can be applied to model the uplink nrtPS traffic in the IEEE 802.16 network. The operational mode is PMP and TDD/TDMA channel allocation scheme is used. Piggybacking is not used. The SSs are the stations of the model. The nrtPS packets arriving to SS i are the i-customers. Thus we call them i-packets. b is the packet length in seconds, which is the integer multiple of the length of the time slot, τ. The polling cycle of the model corresponds to the interval from the start of UL sub-frame until the start of the UL sub-frame in the next frame. Thus c equals the frame length in number of packets.

The SSs apply unicast polling for bandwidth reservation for their nrtPS service flows. It is assumed that there are N polling slots in the RI, among which each of them is dedicated to a SS. Thus each SS has a bandwidth request opportunity in each frame. The uplink bandwidth needs of the nrtPS packets arriving to SS i until the start of the RI are incorporated in the next bandwidth request, which is sent in the dedicated polling slot of SS i in the actual RI. Hence the global gate epoch is the start of the RI and the reserved capacity T of the model corresponds to the RI. u is the total uplink capacity and thus the length of the UL sub-frame is $T+u$ and the length of the DL sub-frame is D.

We assume that the BS knows the number of rtPS and ertPS packets of SS i in each frame and thus it can take them into account at reserving the capacity for the real-time traffic. In the course of scheduling, the BS first assigns capacity for the uplink UGS, rtPS and ertPS transmissions. This reserved capacity for SS i is represented by the actual value of the i-th background Markov chain, Y_i, for each $1,\dots,N$. Thus the cycle setup time R_0 is the RI plus the reserved capacity for these real-time service flows. The remaining capacity is shared among the SSs for their nrtPS traffic so that the available capacity for the nrtPS service flow at SS i is K_i, for each $1,\dots,N$. The capacity, which is not used by the nrtPS traffic of any SS, is allowed to be used for the BE service flow of that SS. Thus R_i is the available capacity for the BE service flow of SS i, for $1,\dots,N-1$. Besides of the BE service flow of SS N, the last switchover time R_N includes also the DL sub-frame of the next frame. Hence the characteristics of this scheduling mechanism can be given as

- The capacity requirements of the UGS, rtPS and ertPS service flows are always ensured.
- The capacity allocation enables priorities for the nrtPS service flows (ζ_i at SS i for $1,\dots,N$). This realizes a weighted round-robin scheduling of the dynamically variable capacity, which remains available after the reservation for the real-time traffic flows.
- The scheduling mechanism ensures an efficient capacity utilizing, since the BE service flows utilize the capacity, which is not used by the nrtPS traffic flows.

The i-packet scheduled for transmission at BS gets service first only in the next frame after informing SS i about the allocated time slots for their uplink transmission. This causes an extra delay with length of one frame for every i-packets. Taking it into account the mean i-packet delay, $E[W_i^p]$, can be given as

$$E[W_i^p] = E[W_i] + cb, \quad 1,\dots,N. \tag{22}$$

7.3. **Modeling correlated real-time traffic.**

7.3.1. *Real-time capacity characteristics.* Let $\boldsymbol{\pi}_i$ be the stationary probability vector of the i-th background Markov chain for $i=1,\dots,N$. Then $\boldsymbol{\pi}_i\boldsymbol{\Pi}_i = \boldsymbol{\pi}_i$ and $\boldsymbol{\pi}_i\mathbf{e} = 1$ uniquely determine $\boldsymbol{\pi}_i$, where $\mathbf{e}$ is the column vector having all elements equal to one.

Let $\boldsymbol{\omega}_i$ be the $L\times 1$ column vector representing the possible values of Y_i in increasing order, i.e. if $j>k$ and $j,k\in\{1,\dots L\}$ then $[\boldsymbol{\omega}_i]_j \geq [\boldsymbol{\omega}_i]_k$. The mean capacity reservation for the real-time traffic flows is given as

$$E[Y_i] = \boldsymbol{\pi}_i \boldsymbol{\omega}_i, \quad i = 1, \ldots, N. \tag{23}$$

Let $\boldsymbol{\Xi}_i$ be a diagonal matrix defined by its elements as $[\boldsymbol{\Xi}_i]_{j,j} = [\boldsymbol{\pi}_i]_j$ for $i = 1, \ldots, N$ and $j = 1, \ldots, L$. Using it the variance of Y_i can be expressed as

$$\begin{aligned} \mathrm{Var}[Y_i] &= E[(Y_i)^2] - (E[Y_i])^2 \\ &= (\boldsymbol{\omega}_i)^T \ \boldsymbol{\Xi}_i \ \omega_i - (\boldsymbol{\pi}_i \boldsymbol{\omega}_i)^2, \quad 1, \ldots, N. \end{aligned} \tag{24}$$

The correlation of two consecutive random capacity reservation for the real-time traffic flows at station i, for $1, \ldots, N$, is defined as

$$\begin{aligned} \gamma_i &= \lim_{m \to \infty} \mathrm{Corr}(Y_i(t_0^f(\ m)), Y_i(t_0^f(\ m+1))) \\ &= \frac{\lim_{m \to \infty} E[Y_i(t_0^f(\ m)) \ Y_i(t_0^f(\ m+1))] - (E[Y_i])^2}{\mathrm{Var}[Y_i]}. \end{aligned} \tag{25}$$

The term $\lim_{m \to \infty} E[Y_i(t_0^f(\ m)) \quad Y_i(t_0^f(\ m+1))]$ can be expressed as $(\boldsymbol{\omega}_i)^T \ \boldsymbol{\Xi}_i \ \boldsymbol{\Pi}_i \ \boldsymbol{\omega}_i$. Using it and applying (23) and (24) in (25) yields

$$\gamma_i = \frac{(\boldsymbol{\omega}_i)^T \ \boldsymbol{\Xi}_i \ \boldsymbol{\Pi}_i \ \boldsymbol{\omega}_i - (\boldsymbol{\pi}_i \boldsymbol{\omega}_i)^2}{(\boldsymbol{\omega}_i)^T \ \boldsymbol{\Xi}_i \ \boldsymbol{\omega}_i - (\boldsymbol{\pi}_i \boldsymbol{\omega}_i)^2}. \tag{26}$$

If $\boldsymbol{\Pi}_i = \mathbf{e}\boldsymbol{\pi}_i$, then $(\boldsymbol{\omega}_i)^T \ \boldsymbol{\Xi}_i \ \boldsymbol{\Pi}_i \ \boldsymbol{\omega}_i = (\boldsymbol{\omega}_i)^T \ \boldsymbol{\Xi}_i \ \mathbf{e} \ \boldsymbol{\pi}_i \ \boldsymbol{\omega}_i = (\boldsymbol{\omega}_i)^T \ (\boldsymbol{\pi}_i)^T \ (\boldsymbol{\pi}_i \ \boldsymbol{\omega}_i) = (\boldsymbol{\pi}_i \boldsymbol{\omega}_i)^2$. It follows that in this case the correlation of two consecutive capacity reservation for the real-time traffic flows at station i is 0.

7.3.2. *Characterizing real-time traffic.* Let ω_{min} is the smallest value among the values $\omega_1, \ldots, \omega_L$. The capacity requirement of the (e)rtPS service flow is usually specified by the required minimum and maximum bandwidth, i.e. by ω_{min} and ω_{max}. Thus increasing the capacity demand of the same type of (e)rtPS traffic implies a *multiplication* on the values of ω_{min} and ω_{max}. On the other hand the capacity requirement of the UGS service flow is usually specified as a fixed amount of bandwidth. It follows that increasing the capacity demand of the same type of UGS traffic means a *shift* on the values of ω_{min} and ω_{max}. Hence the amount independent characterization of the real time traffic can be given by such *normalized quantities* which are invariant for the above mentioned linear operations.

The normalized mean capacity for the real-time traffic flows at station i is defined as

$$E[\mathcal{Y}_i] = \frac{E[Y_i] - \omega_{min}}{\omega_{max} - \omega_{min}}, \quad i = 1, \ldots, N. \tag{27}$$

Similarly the normalized variance of the capacity for the real-time traffic flows at station i is defined as

$$\mathrm{V}ar[\mathcal{Y}_i] = \frac{\mathrm{Var}[Y_i]}{(\omega_{max} - \omega_{min})^2}, \quad i = 1, \ldots, N. \tag{28}$$

The definitions of the above normalized quantities imply that they are linear invariant. Similarly it can be seen from the definition of the correlation γ_i that it is already linear invariant and hence there is no need to apply any normalization on it.

Summarizing all these the characterization of the real time traffic (the UGS and the (e)rtPS service flows) can be separated into *amount independent characterization* and *amount specifying characterization.* The amount independent characterization is represented by the quantities $E[\mathcal{Y}_i]$, $\mathrm{V}ar[\mathcal{Y}_i]$ and γ_i, while the amount specifying characterization is given by the values of ω_{min} and ω_{max}.

7.3.3. *Computation of the transition probability matrix.* For modeling the real time capacity we use a two-state Markov chain, i.e. $L = 2$. In spite of having only two states it is appropriate to model a correlated traffic.

The stationary probability vector of the i-th background Markov chain, $\boldsymbol{\pi}_i$ is determined from $E[\mathcal{Y}_i]$. Using (23) and $\boldsymbol{\pi}_i\mathbf{e} = 1$ after some rearranging leads to

$$\boldsymbol{\pi}_i = (1 - E[\mathcal{Y}_i], E[\mathcal{Y}_i]). \tag{29}$$

The 2×2 transition probability matrix at station i has the form

$$\boldsymbol{\Pi}_i = \begin{pmatrix} 1 - p_{i,12} & p_{i,12} \\ p_{i,21} & 1 - p_{i,21}, \end{pmatrix} \tag{30}$$

where we already utilized that matrix $\boldsymbol{\Pi}_i$ is stochastic. Using the relation $\boldsymbol{\pi}_i\boldsymbol{\Pi}_i = \boldsymbol{\pi}_i$ and (26) yields 2 equations for the unknowns $p_{i,12}$ and $p_{i,21}$. Solving this system of linear equations results in the expressions of the elements of $\boldsymbol{\Pi}_i$ as

$$p_{i,12} = \frac{\mathrm{V}ar[\mathcal{Y}_i](1-\gamma_i)}{1 - E[\mathcal{Y}_i]}, \quad p_{i,21} = \frac{\mathrm{V}ar[\mathcal{Y}_i](1-\gamma_i)}{E[\mathcal{Y}_i]}. \tag{31}$$

$p_{i,12} \geq 0$ implies $\gamma_i \leq 1$. For this case of two states Markov chain $\mathrm{V}ar[\mathcal{Y}_i]$ can be expressed by $E[\mathcal{Y}_i]$ as

$$\mathrm{V}ar[\mathcal{Y}_i] = E[\mathcal{Y}_i](1 - E[\mathcal{Y}_i]). \tag{32}$$

Using $p_{i,12} \leq 1$, $p_{i,21} \leq 1$ and (32) after some rearrangement we get

$$\gamma_i \geq 1 - \min(\frac{1}{E[\mathcal{Y}_i]}, \frac{1}{1 - E[\mathcal{Y}_i]}). \tag{33}$$

The minimum value of right-hand side of (33) is -1 at $E[\mathcal{Y}_i] = 0.5$. Thus $-1 \leq \gamma_i \leq 1$ as expected and the actual minimum value of γ_i depends on $E[\mathcal{Y}_i]$ according to the relation (33).

7.4. **Examples for performance evaluation.** In this section we provide examples for the performance evaluation of the IEEE 802.16 uplink nrtPS service flow by applying the presented polling model. For the numerical computations we assume 10 MHz TDD system with 5 *ms* frame duration, in which the UL sub-frame comprises 175 slots and the IEEE 802.16-2009 system transmits 16 bytes per UL slot. These parameters are taken from [9]. For the sake of simplicity we set the length of DL sub-frame to 0. We model the nrtPS traffic with Peer-To-Peer (P2P) workload. According to [4] P2P workload is one of the major data source on the internet and the packets size is fixed 128 bytes for one of the dominant P2P application.

Parameter	Value
Frame duration (c)	5 ms
Packet service time (b)	1/21.875 $frames$
DL sub-frame length (D)	0 $packets$
RI length (T)	0.875 $packet$
Total uplink capacity (u)	21 $packets$
$E[\mathcal{Y}_i]$, $i = 1, 2$	0.5
γ_i, $i = 1, 2$	0.3
ω_{min}	1 $packets$
ω_{max}	7 $packets$
ζ_i, $i = 1, 2$	0.5

TABLE 1. Evaluation parameters

Hence the packet service time is constant with the length of 8 slots and the frame duration has a length of 21.875 packets. The length of the RI is 0.875 packet service time (7 slots) and thus the total uplink capacity consists of 21 packet service times. In the numerical examples we use normalized time, in which the time unit equals the length of the frame. The number of SSs are 2 and both the real time traffic and nrtPS traffic parameter setting is symmetric. Table 1 summarizes the evaluation parameters.

In figure 3 we have plotted the dependency of the mean packet delay on the load for different values of the normalized mean real time traffic capacity ($E[\mathcal{Y}_i]$) and of the correlation parameter (γ_i).

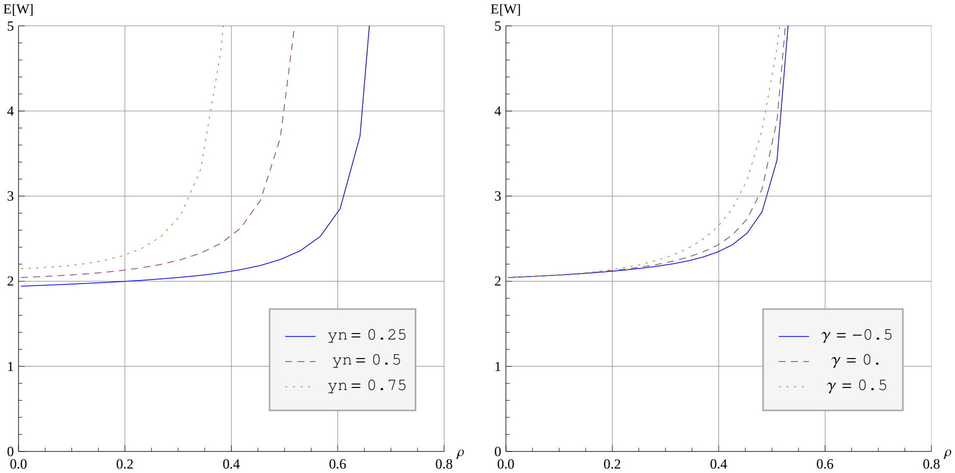

FIGURE 3. Mean packet delay ($E[W]$) versus load (ρ) for different normalized mean real time traffic capacity - $E[\mathcal{Y}_i]$=yn (left side) and correlation parameters - γ_i=γ (right side).

It can be seen on the left side of the figure that increasing $E[\mathcal{Y}_i]$ leads to higher mean nrtPS packet delay. It is due to the effect that increasing $E[\mathcal{Y}_i]$ at fixed values of ω_{min} and ω_{max} yields higher mean real time capacity $E[Y_i]$. Another effect is that increasing $E[\mathcal{Y}_i]$ restricts the maximum allowed value of the offered nrtPS load. This is because higher mean real time capacity implies smaller stability boundary for the offered nrtPS traffic (6). Another conclusion which can be drawn from the right side of this figure is that the effect of changing the correlation parameter (γ_i) is not so crucial for the mean nrtPS packet delay. The positive correlation increases the mean nrtPS packet delay.

Figure 4 shows the dependency of the mean packet delay on the load for different values of ω_{min} and ω_{max}, i.e. for different amount of (e)rtPS traffic and UGS traffic. In general it can be seen from the figure that the effects of increasing both the (e)rtPS traffic and UGS traffic are similar. The mean nrtPS packet delay becomes higher in both cases and the maximum allowed value of the offered nrtPS load decreases. The reason for the second effect is again that the higher mean real time capacity implies smaller stability boundary for the offered nrtPS traffic (6). The concrete differences in the runs of the curves are rather the implications of the applied parameter settings than possible consequences of any other modeled effects.

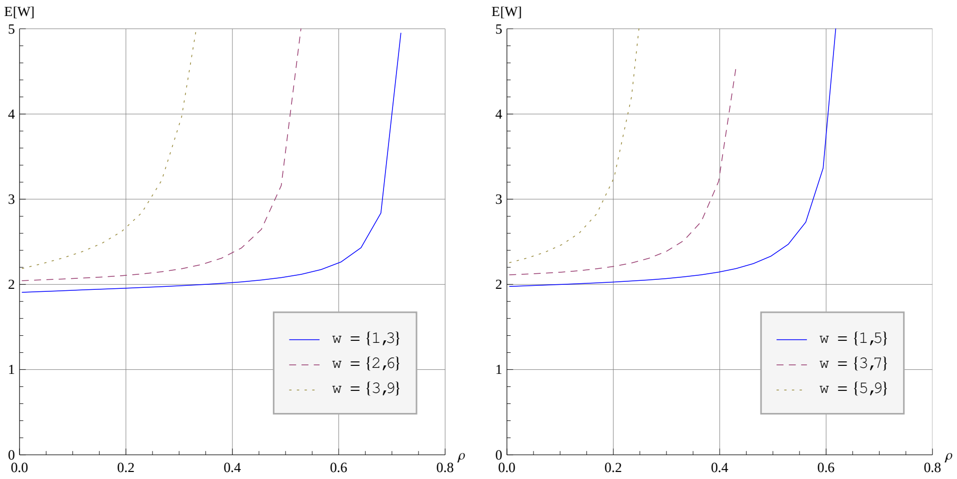

FIGURE 4. Mean packet delay $(E[W])$ versus load (ρ) for different amount of (e)rtPS traffic - $\boldsymbol{\omega}$ (left side) and UGS traffic - $\boldsymbol{\omega}$ (right side).

7.5. **Enforcing an upper bound on mean delay.** This modeling can be also used to enforcing specified upper bounds on mean nrtPS packet delays at every SSs in a specified range of load. These bounds can be different for the individual SSs. The characteristics of the real time traffic ($E[\mathcal{Y}_i]$ and γ_i) as well as the priority weight for the nrtPS service flow (ζ_i) at SS i are given for every $i = 1, \ldots, N$. In this case the amount of the reserved capacity for the real-time traffic flows (in terms of ω_{min} and ω_{max}) is maximized over a restricted parameter set, which is determined by the specified upper bounds on mean nrtPS packet delays and by the specified range of load. Additionally a further relation is needed among ω_{min} and ω_{max}, which can be established from the (e)rtPS and UGS bandwidth requirements.

7.6. **Cost model.** In case of more general QoS requirement on delay constraint an appropriate cost model can be built to determine the optimal parameters of the real-time traffic flows. We developed a steady-state average cost function $\mathcal{F}(\boldsymbol{\omega})$, where the real-time capacity range $\boldsymbol{\omega} = (\omega_{min}, \omega_{max})$ is the decision variable. The parameters of the cost function for $i = 1, \ldots, N$ are defined as

$$\varpi_i \equiv \text{Cost of the mean packet delay at station i},$$
$$\vartheta_i \equiv \text{Reward of the mean real-time capacity at station i}.$$

Then the optimal parameters of the real-time traffic flows can be obtained by minimizing the total average system cost, which is given as

$$\mathcal{F}(\boldsymbol{\omega}) = \sum_{i=1}^{N} \left(\varpi_i E[W_i^p] + \frac{\vartheta_i}{E[Y_i]} \right). \tag{34}$$

The minimum can be numerically determined as a function of the load, the normalized mean real time traffic capacities ($E[\mathcal{Y}_i]$), the correlation parameters (γ_i) and the priority weights for the nrtPS service flows (ζ_i), for $i = 1, \ldots, N$, by applying the expressions (22), (27), (31) and (32).

REFERENCES

[1] S. Andreev, Zs. Saffer and A. Anisimov, "Overall Delay Analysis of IEEE 802.16 Network," Int. Workshop on Multiple Access Comm. (MACOM), 2009.

[2] O. J. Boxma, H. Levy and U. Yechiali, *Cyclic reservation schemes for efficient operation of multiple-queue single-server systems,* Annals of Operations Research, **35** (1992), 187–208.

[3] Y.-J. Chang, F.-T. Chien and C.-C. J. Kuo, *Delay analysis and comparison of OFDM-TDMA and OFDMA under IEEE 802.16 QoS framework,* IEEE Global Telecomm. Conf. (GLOBECOM), **1** (2006), 1–6.

[4] S. Forconi, G. Iazeolla, P. Kritzinger and P. Pillegi, "Modelling Internet Workloads for IEEE 802.16," Technical Report CS08-03-00, Department of Computer Science, University of Cape Town, 2008.

[5] R. Iyengar, P. Iyer and B. Sikdar, *Delay analysis of 802.16 based last mile wireless networks,* IEEE Global Telecommunications Conference (GLOBECOM), **5** (2005), 3123–3127.

[6] Zs. Saffer, *An introduction to classical cyclic polling model*, Proc. of the 14th Int. Conf. on Analytical and Stochastic Modelling Techniques and Applications (ASMTA'07), (2007), 59–64.

[7] (MR2502344) Zs. Saffer and M. Telek, *Stability of periodic polling system with BMAP arrivals,* European Journal of Operational Research, **197** (2009), 188–195.

[8] (MR2584755) Zs. Saffer and M. Telek, *Unified analysis of $BMAP/G/1$ cyclic polling models,* Queueing Systems, **64** (2010), 69–102.

[9] C. So-In, R. Jain and A.-K. Tamimi, *Capacity evaluation for IEEE 802.16e mobile WiMAX,* Journal of Computer Systems, Networks and Communications, **2010** (2010), 1–12.

[10] Standard IEEE 802.16-2009, Part 16: Air Interface for Broadband Wireless Access Systems, Standard for Local and Metropolitan Area Networks, May 2009.

[11] H. Takagi, "Analysis of Polling Systems," MIT Press, 1986.

[12] A. Vinel, Y. Zhang, Q. Ni and A. Lyakhov, *Efficient request mechanism usage in IEEE 802.16,* IEEE Global Telecommunications Conference (GLOBECOM), **1** (2006), 1–5.

Received September 2010; 1st revision December 2010; final revision May 2011.

E-mail address: safferzs@hit.bme.hu
E-mail address: telek@hit.bme.hu