Analysis of globally gated Markovian limited cyclic polling model and its application to IEEE 802.16 network

Zsolt Saffer*
Department of Telecommunications
Budapest University of Technology and
Economics, Budapest, Hungary
safferzs@hit.bme.hu

Miklós Telek
Department of Telecommunications
Budapest University of Technology and
Economics, Budapest, Hungary
telek@hit.bme.hu

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. At each station customers arrive with Poisson process and the customer service time is constant. Moreover the cycle time is a fixed integer multiple of the customer service time. The model enables asymmetric 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 demonstrated.

*Corresponding author.

*QTNA*2010, July 24-26, 2010, Beijing, China.
Copyright © 2010 ACM 978-1-4503-0212-8/ ...$10.00.

Categories and Subject Descriptors

G.3 [**Mathematics of Computing**]: Probability and Statistics—*Queueing theory*

General Terms

Theory, Performance

Keywords

Queueing theory, polling model, waiting time, capacity allocation, IEEE 802.16

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 [1].

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. 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 [2].

Related works on delay analysis of IEEE 802.16 network are [3], [4] and [5]. In [6] 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 [7] 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 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 [1]. 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. b_i, $b_i^{(2)}$ and $\widetilde{B}_i(s)$ stand for its mean, its second moment and its LST, respectively. Additionally b_i is the same for every $i = 1, \ldots, N$ and thus it is also 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. As a consequence of the above model definition the cycle time is also an integer multiple of the constant customer service time. Let c denote its length 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 [9]) 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-customers 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.

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 $\{Y_i(t_0^f(m));\ m \in \{1, \ldots\}\}$ 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 $K_i(m)$ 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

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

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

where $\lfloor d \rfloor$ stands for the integral part of d.

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

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

where $K_i = \lim^d{}_{m \to \infty} K_i(\ m)$ and $Y_j = \lim^d{}_{m \to \infty} Y_j(\ m)$ and $\lim^d$ stands for the convergence in distribution.

Y_i can represent a reserved capacity at station i from the total capacity c, where the capacity is in the number of constant customer service times. Thus $c - \sum_{j=1}^{N} Y_j$ is the total remaining capacity in the system, from which station i gets K_i according to its priority weight ζ_i.

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 given as

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

Let S_i denote the length of the i-station time in the number of constant customer service times. The switchover time R_i is explicitly given as

$$\begin{aligned} &R_i = K_i - S_i, \quad i = 1, \ldots, N-1, \\ &R_N \geq K_N - S_N. \end{aligned} \tag{3}$$

Thus the cycle setup time at the begin of the cycle can represent the total reserved capacity. Additionally for every $i = 1, \ldots, N$ the random limit K_i is the remaining capacity allocated to station i according to its priority weight, while the switchover time R_i is the unused part of it. Note that besides of the unused capacity at station N the last switchover part (R_N) can incorporate also an additional interval. The globally gated Markovian limited cyclic polling model is illustrated on Fig. 1.

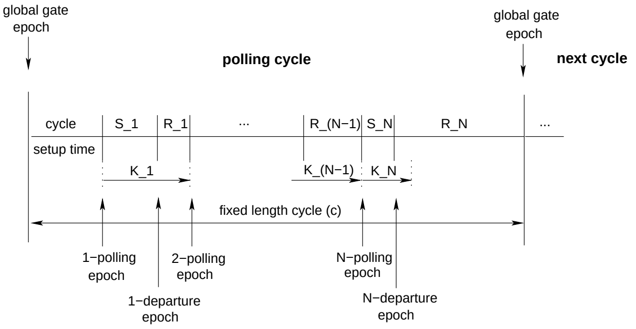

Figure 1: Globally gated Markovian limited cyclic model

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

A.1 For every $i = 1, \ldots, N$ 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 $i = 1, \ldots, N$.

A.4 The following relation holds for the model

$$c - N\omega_{max} \geq N,$$

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, which implies that the traffic at the 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

For $m \geq 1$ let $A_i(\ m)$ and $G_i(\ m)$ stand for the number of arriving and the number of served i-customers during the m-th cycle, respectively. In addition we define $a_i = \lim_{\ m \to \infty} E[A_i(\ m)]$ and $g_i = \lim_{\ m \to \infty} E[G_i(\ m)]$. a_i and g_i can be interpreted as the mean stationary number of i-customers arrivals and services during a cycle, respectively. 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

$$\mathbf{Z}(t_0^f(\ m)) = \big(N_1(t_0^f(\ m)), \ldots, N_N(t_0^f(\ m)),\\ Y_1(t_0^f(\ m)), \ldots, Y_N(t_0^f(\ m))\big). \qquad (4)$$

$\mathbf{Z}(t_0^f(\ m))$ 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. The state space of this Markov chain consists only of finite valued and countable infinite spaces. This ensures that the stability analysis and results in [10] 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

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

Applying (2) in (5) leads to the condition of the whole stability as

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

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

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

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. Thus it is enough to establish relations among their joint probabilities 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

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

THEOREM 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*

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

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 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 $\frac{(\lambda_i c)^{n_i - k_i + s_i}}{(n_i - k_i + s_i)!} e^{-\lambda_i c}$. 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

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

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

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

3.2 The stationary joint probabilities at polling and departure epochs

For $m \geq 1$ let $t_i^f(\ m)$ and $t_i^m(\ m)$ 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

$$\begin{aligned} p_i^f(n_i, y_1, \ldots, y_N) = & \lim_{m\to\infty} Pr\{N_i(t_i^f(\ m)) = n_i, \\ & Y_1(t_i^f(\ m)) = y_1, \ldots, Y_N(t_i^f(\ m)) = y_N\}, \\ & n_i \in \{0, 1, \ldots\},\ \ y_1, \ldots, y_N \in \Omega,\ \ i = 1, \ldots, N. \end{aligned} \tag{9}$$

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

$$\begin{aligned} p_i^m(n_i, y_1, \ldots, y_N) = & \lim_{m\to\infty} Pr\{N_i(t_i^m(\ m)) = n_i, \\ & Y_1(t_i^m(\ m)) = y_1, \ldots, Y_N(t_i^m(\ m)) = y_N\}, \\ & n_i \in \{0, 1, \ldots\},\ \ y_1, \ldots, y_N \in \Omega,\ \ i = 1, \ldots, N. \end{aligned} \tag{10}$$

THEOREM 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*

$$\begin{aligned} & p_i^f(n_i, y_1, \ldots, y_N) = \sum_{k_i=0}^{n_i} p_{0,i}^f(k_i, y_1, \ldots, y_N) \\ & \frac{(\lambda_i(R_0 + \sum_{\ell=1}^{i-1} K_\ell))^{n_i-k_i}}{(n_i - k_i)!} e^{-\lambda_i(R_0+\sum_{\ell=1}^{i-1} K_\ell)}, \\ & p_i^m(n_i, y_1, \ldots, y_N) = \sum_{k_i=0}^{n_i+K_i} p_{0,i}^f(k_i, y_1, \ldots, y_N) \\ & \frac{(\lambda_i(R_0 + \sum_{\ell=1}^{i-1} K_\ell + s_i))^{n_i-k_i+s_i}}{(n_i - k_i + s_i)!} e^{-\lambda_i(R_0+\sum_{\ell=1}^{i-1} K_\ell+s_i)}, \\ & R_0 = \sum_{\ell=1}^{N} y_\ell,\ \ K_i = \left\lfloor \zeta_i(c - \sum_{\ell=1}^{N} y_\ell) \right\rfloor,\ \ s_i = min(k_i, K_i) \\ & n_i \in \{0, 1, \ldots\},\ \ y_1, \ldots, y_N \in \Omega,\ \ i = 1, \ldots, N. \end{aligned} \tag{11}$$

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 $\frac{(\lambda_i(R_0+\sum_{\ell=1}^{i-1} K_\ell))^{n_i-k_i}}{(n_i-k_i)!} e^{-\lambda_i(R_0+\sum_{\ell=1}^{i-1} K_\ell)}$. 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 $\frac{(\lambda_i(R_0+\sum_{\ell=1}^{i-1} K_\ell+s_i))^{n_i-k_i+s_i}}{(n_i-k_i+s_i)!} e^{-\lambda_i(R_0+\sum_{\ell=1}^{i-1} K_\ell+s_i)}$. 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

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

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

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

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

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

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

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

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 [7] can be extended to the globally gated Markovian limited cyclic polling model. The statement (13) is proven in [7] for the classical cyclic polling model. 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 4. *The LST of the stationary waiting time of an i-customer is given by*

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

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 [7] that under this 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*

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

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 $n_i \leq X$ 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, at which the probabilities $p_{0,i}^f(n_i, y_1, \ldots, y_N)$ for $n_i > X$ can be neglected. This probabilities, $p_{0,i}^f(n_i, y_1, \ldots, y_N)$ for $n_i > X$, are set 0. This leads to

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

Similarly setting the same upper limit $n_i \leq X$ 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

$$\begin{aligned}&p_i^f(n_i, y_1, \ldots, y_N) = \sum_{k_i=0}^{min(n_i,X)} p_{0,i}^f(k_i, y_1, \ldots, y_N) \\ &\frac{(\lambda_i(R_0 + \sum_{\ell=1}^{i-1} K_\ell))^{n_i-k_i}}{(n_i-k_i)!} e^{-\lambda_i(R_0+\sum_{\ell=1}^{i-1} K_\ell)}, \\ &p_i^m(n_i, y_1, \ldots, y_N) = \sum_{k_i=0}^{min(n_i+K_i,X)} p_{0,i}^f(k_i, y_1, \ldots, y_N) \\ &\frac{(\lambda_i(R_0 + \sum_{\ell=1}^{i-1} K_\ell + s_i))^{n_i-k_i+s_i}}{(n_i-k_i+s_i)!} e^{-\lambda_i(R_0+\sum_{\ell=1}^{i-1} K_\ell+s_i)}, \\ &R_0 = \sum_{\ell=1}^{N} y_\ell, \quad K_i = \left\lfloor \zeta_i(c - \sum_{\ell=1}^{N} y_\ell) \right\rfloor, \quad s_i = min(k_i, K_i) \\ &n_i \in \{0, 1, \ldots X\}, \quad y_1, \ldots, y_N \in \Omega, \quad i = 1, \ldots, N.\end{aligned} \tag{17}$$

6.2 The steps of the numerical procedure

The computation of the first moment of the stationary waiting time of an i-customers 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 $\mathbf{e}_\ell^{X+1} = (0, \ldots, 0, 1, 0, \ldots, 0)$ denote the $1 \times (X+1)$ vector with 1 at the ℓ-th position. Furthermore $\otimes$ stands for the Kronecker product. We define the $1 \times L^N(X+1)$ vector $\boldsymbol{\theta}_i$, representing the unknowns of the above system of linear equations as

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

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 $L^N(X+1)\times L^N(X+1)$ matrix $\boldsymbol{\Upsilon}_i$ representing the coefficients on the right-hand side (r.h.s.) of the equation (16). It is defined as

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

In this matrix the values of k_i, j_1, …, j_N and the values of n_i, y_1, …, y_N specify the row and the column indices of the corresponding coefficient. The following relation holds for the Kronecker product of the probability transition matrices $\boldsymbol{\Pi}_1,\ldots,\boldsymbol{\Pi}_N$:

$$\begin{aligned}&\sum_{j_1\in\Omega}\cdots\sum_{j_N\in\Omega}\sum_{y_1\in\Omega}\cdots\sum_{y_N\in\Omega}[\boldsymbol{\Pi}_1]_{j_1,y_1}\ldots[\boldsymbol{\Pi}_N]_{j_N,y_N}\\ &\left(\mathbf{e}^L_{\mathcal{I}(j_1)}\ldots\otimes\mathbf{e}^L_{\mathcal{I}(j_N)}\right)^T\left(\mathbf{e}^L_{\mathcal{I}(y_1)}\ldots\otimes\mathbf{e}^L_{\mathcal{I}(y_N)}\right)\\ &=\boldsymbol{\Pi}_1\otimes\ldots\otimes\boldsymbol{\Pi}_N.\end{aligned}\tag{20}$$

By using (20) the relation (19) can be rearranged as

$$\begin{aligned}\boldsymbol{\Upsilon}_i =&\boldsymbol{\Pi}_1\otimes\ldots\otimes\boldsymbol{\Pi}_N\otimes\Bigg(\sum_{k_i=0}^{X}\sum_{n_i=0}^{X}\frac{(\lambda_i c)^{n_i-k_i+s_i}}{(n_i-k_i+s_i)!}e^{-\lambda_i c}\\ &\left(\mathbf{e}^{X+1}_{k_i+1}\right)^T\left(\mathbf{e}^{X+1}_{n_i+1}\right)\Bigg)\quad i=1,\ldots,N.\end{aligned}\tag{21}$$

Using definitions (18) and (21) the matrix form of the system of linear equation (16) can be given as

$$\boldsymbol{\theta}_i=\boldsymbol{\theta}_i\boldsymbol{\Upsilon}_i,\quad i=1,\ldots,N.\tag{22}$$

The sums on the r.h.s. of (16) realize the product of vector $\boldsymbol{\theta}_i$ by matrix $\boldsymbol{\Upsilon}_i$.

Matrix $\boldsymbol{\Upsilon}_i$ relates the probabilities of vector $\boldsymbol{\theta}_i$ and hence it can be interpreted as transition probability matrix. It follows that $\boldsymbol{\Upsilon}_i$ is stochastic and as it is shown in [8] for an equation with such a form, $rank\,(\mathbf{I}-\boldsymbol{\Upsilon}_i)$ is one less than the dimension of $\boldsymbol{\Upsilon}_i$. Therefore (22) does not determine $\boldsymbol{\theta}_i$ uniquely. To make the system of linear equations complete we add the normalization condition as

$$\boldsymbol{\theta}_i\mathbf{e}^{L^N(X+1)}=1,\quad i=1,\ldots,N,\tag{23}$$

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

The joint probabilities $p^f_{0,i}(n_i,y_1,\ldots,y_N)$ for $n_i\in\{0,1,\ldots X\}$, $y_1,\ldots,y_N\in\Omega$, $i=1,\ldots,N$ can be uniquely determined from the system of linear equation (22) and (23).

2. Solving the matrix form system of linear equation (22) and (23) 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 $p^f_i(n_i)$ for $n_i\in\{0,1,\ldots X\}$ 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, $f^{(2)}_i$, m_i, $m^{(2)}_i$) from the probabilities $p^f_i(n_i)$ and $p^f_i(n_i)$ for $n\in\{0,1,\ldots X\}$ 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, $f^{(2)}_i$, m_i, $m^{(2)}_i$ 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 (22) and (23). The number of equations and unknowns in these system of linear equations for all stations is $NL^N(X+1)$. Therefore the total number of operations required by the whole numerical procedure is in the magnitude of $NL^{3N}(X+1)^3$.

7. APPLICATION TO THE IEEE 802.16 NETWORK

7.1 Analytic model of the uplink nrtPS traffic in the IEEE 802.16 network

The presented model can be applied to model the uplink Non Real-Time Polling Service (nrtPS) traffic in the IEEE 802.16 network. The operational mode is point-to-multipoint (PMP) and Time Division Duplex (TDD)/TDMA channel allocation scheme is used. Piggybacking is not used. The Subscriber Stations (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 UpLink (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 Reservation Interval (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.

We assume that the BS knows the number of Real-Time Polling Service (rtPS) and Extended Real-Time Variable Rate (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 Base Station (BS) first assigns capacity for the uplink Unsolicited Grant Service (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 $1, \ldots, N$. Thus the cycle setup time R_0 is 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 $1, \ldots, N$. The capacity, which is not used by the nrtPS traffic of any SS, is allowed to be used for the Best Effort (BE) service flow of that SS. Thus R_i is the available capacity for the BE service flow of SS i, for $1, \ldots, N-1$. 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. Therefore 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 $1, \ldots, N$). 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

$$E[W_i^p] = E[W_i] + c, \quad 1, \ldots, N. \tag{24}$$

7.2 Performance modeling

The application of the presented analytical model to the uplink nrtPS traffic in the IEEE 802.16 network enables to investigate the packet delay of the nrtPS service flow as a function of the parameters of the real-time traffics. This includes the mean and the maximum of the reserved capacity for the UGS, rtPS and ertPS service flows ($E[Y_i]$, ω_{max}) as well as the correlation of two consecutive random capacity reservation for the real-time traffic flows (γ_i).

This modeling can be also used to enforcing a specified upper bound on mean packet delay in a specified range of load. In this case the the mean or the maximum of the reserved capacity for the real-time traffic flows ($E[Y_i]$, ω_{max}) is maximized over a restricted parameter set, which is determined by the specified upper bound on mean packet delay and by the specified range of load.

7.3 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 $\boldsymbol{\omega}$ 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

$$\mathcal{F}(\boldsymbol{\omega}) = \sum_{i=1}^{N} \left(\varpi_i E[W_i^p] + \frac{\vartheta_i}{E[Y_i]} \right). \tag{25}$$

The minimum can be numerically determined as a function of the load and the correlations γ_i, for $i = 1, \ldots, N$, by applying the expressions of the mean i-packet delay (24).

8. REFERENCES

[1] H. Takagi. *Analysis of Polling Systems.* MIT Press, 1986.

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

[3] R. Iyengar, P. Iyer, and B. Sikdar. Delay Analysis of 802.16 based Last Mile Wireless Networks. In *IEEE Global Telecommunications Conference (GLOBECOM)*, volume 5, pages 3123–3127, 2005.

[4] 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. In *IEEE Global Telecomm. Conf. (GLOBECOM)*, volume 1, pages 1–6, 2006.

[5] A. Vinel, Y. Zhang, Q. Ni, and A. Lyakhov. Efficient Request Mechanism Usage in IEEE 802.16. In *IEEE Global Telecommunications Conference (GLOBECOM)*, volume 1, pages 1–5, 2006.

[6] S. Andreev, Zs. Saffer, and A. Anisimov. Overall delay analysis of IEEE 802.16 network. In *Int. Workshop on Multiple Access Comm. (MACOM)*, 2009.

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

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

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

[10] Zs. Saffer and M. Telek. Stability of periodic polling system with BMAP arrivals. *European Journal of Operational Research*, 197(1):188–195.