Manuscript submitted to
AIMS' Journals
Volume **X**, Number **0X**, XX **200X**

Website: http://AIMsciences.org

pp. **X–XX**

ANALYSIS OF BMAP VACATION QUEUE AND ITS APPLICATION TO IEEE 802.16E SLEEP MODE

Zsolt Saffer

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

Miklós Telek

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

(Communicated by the associate editor name)

Abstract. The paper deals with the continuous-time $BMAP/G/1$ queue with multiple vacations and with its application to IEEE 802.16e sleep mode. The lengths of the vacation periods have general distribution and they depend on the number of preceding vacations (dependent multiple vacation). We give the expressions for the vector generating function of the stationary number of customers and its mean. Moreover we obtain new formulas for the vector Laplace-Stieljes transform of the stationary virtual waiting time and for its first two moments in case of First-Come First-Serve scheduling.

We apply this vacation model to the IEEE 802.16e sleep mode mechanism, and we evaluate its performance as a function of the traffic intensity and the traffic correlation parameter. We give an example for determining the best sleep mode parameters for a simple optimization criteria and we also develop a cost model for the more general case. For traffic modeling we use a two-phase Markovian Arrival Process, which is appropriate to model a fairly general correlated traffic.

1. **Introduction.** Queueing models with server vacation are effective instruments in modeling and analysis of computer and manufacturing systems as well as in analysis of telecommunication models. For more details on vacation models we refer to the excellent book of Takagi [1] and the survey of Doshi [2].

Since the introduction of batch Markovian arrival process ($BMAP$) by Lucantoni [3] many authors investigated queueing models with $BMAP$. The reason is that $BMAP$ enables more realistic and more accurate traffic modeling, since it can also capture dependency in traffic processes. Most of these works apply the standard matrix analytic-method pioneered by Neuts [4] and further extended by many others, see e.g. [5]. However only a few works are available on $BMAP$ queueing models with server vacation.

Chang and Takine [6] considered a class of $BMAP$ queues with generalized vacation and determined the vector probability generating function (vector GF) of the stationary queue length and its factorial moments for models with exhaustive discipline.

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

Key words and phrases. Queueing theory, multiple vacation model, BMAP, IEEE 802.16e sleep mode.

In our previous work [7] we considered $BMAP$ vacation queue with gated and G-limited disciplines and derived the vector GF of the stationary number of customers at an arbitrary instant and its mean.

In this paper we analyze $BMAP$ queue model with multiple vacations and exhaustive discipline, in which the vacation periods depend on the number of preceding vacations. We call this vacation strategy as *dependent multiple vacation*. To the best knowledge of the authors, no results are available for this continuous-time vacation model. This vacation system is capable of modeling sleep mode mechanism in wireless networks.

The incoming traffic has self-similar and bursty nature also in wireless networks causing correlation in inter-arrival times, which influences the performance of the system. Our motivation for using BMAP is that it can model such traffic correlation. Hence applying BMAP in the queueing model enables the traffic correlation dependent performance evaluation of the system.

The principal goal of this work is to give a general continuous-time $BMAP$ queueing model for evaluating the performance of the sleep mode mechanisms in wireless networks and for optimal tuning of the sleep mode parameters on traffic correlation dependent manner.

The IEEE 802.16 standard [8] is recommended for Wireless Metropolitan Area Networks (WMAN). It is also called as WiMAX (from "Worldwide Interoperability for Microwave Access") as it has been commercialized under this name. In the application part of this paper we focus on the IEEE 802.16e sleep mode mechanism with power saving class of type I. The three types of power saving classes in IEEE 802.16e sleep mode mechanism representing different sleep mode operations. The power saving class of type I is recommended for connections having best effort (BE) service and non-real time polling service (nrtPS).

Most of the performance analysis of the IEEE 802.16e sleep mode assumes uncorrelated traffic. In [9], [10] and [11] the incoming traffic is modeled by Poisson process while the performance evaluation in [12] is based on an appropriate discrete-time Geom/G/1 vacation queue model. Recent works on performance analysis of power saving mechanisms in IEEE 802.16e are [13] and [14], in which also Poisson process is used as traffic model.

Turck et al. [15] investigated a discrete-time $BMAP$ ($D-BMAP$) queue with multiple vacations, exhaustive discipline and First-Come First-Serve (FCFS) scheduling. Their time-slotted model also allows that the fixed lengths vacation periods depend on the number of preceding vacations. They derived the vector GF of the number of packets in the queue and the vector GF of the packet delay. They applied the so-called ON-OFF traffic model, which generates bursty, correlated traffic, to investigate the performance of the IEEE 802.16e sleep mode mechanism in terms of the burst-length factor.

Our model can be seen also as the generalized continuous-time counterpart of the model of [15], in which the lengths of the vacation periods can have a general distribution. We apply the canonical form of the two-phase MAP [16] for modeling correlated traffic and evaluate the effect of the traffic correlation on the performance measures and on the optimal parameters of the IEEE 802.16e sleep mode.

The queueing theoretic contribution of this paper is the expressions for the vector GF of the stationary number of customers at an arbitrary instant and for its mean as well as the new formulas for the vector Laplace-Stieljes transform (vector LST) of the stationary virtual waiting time and for its first two moments. The derivation

of them is based on determination of two joint transforms. The first one is the joint transform of the stationary number of customers in the system and the forward recurrence vacation time. The second one is the joint transform of the stationary number of customers in the system and the forward recurrence customer service time at an arbitrary instant in service period. In the derivation of the vector GF of the stationary number of customers at an arbitrary instant we also apply results from [7].

The application specific contribution of the paper is the capability of the considered model to predict the influence of the sleep mode parameters on the mean packet delay and the mean power savings also in presence of traffic correlation. We show an example for traffic correlation dependent determination of the optimal sleep mode parameters for the simple strategy, in which the mean power savings practically prioritized over mean packet delay (for BE and nrtPS services). It turns out that the optimal sleep mode strategy depends on the correlation parameter. We also describe how to take into account an upper bound on mean delay in optimizing the sleep mode parameters. Moreover we also introduce a cost model, which takes into account the Quality of Service (QoS) on delay constraint and the mean power savings. These optimizations facilitate the tuning of the sleep mode parameters to the requirements of the actual application scenario and thus they have potential applications in network control.

Moreover applying BMAP in the considered queueing model enables the application of correlated traffic models in the analysis of the considered sleep mode mechanism. Such data traffic models have been developed for simulation based performance analysis of the IEEE 802.16, e.g. in [17].

The rest of this paper is organized as follows. In section 2 we introduce the model and the notations. The derivation of the joint transforms follows in section 3. In section 4 the expressions for the vector GF of the stationary number of customers and its mean follow. The new formulas of vector LST of the stationary virtual waiting time and its first two moments are derived in section 5. In section 6 we determine the stationary probability vector at start of the whole vacation. The application to IEEE 802.16e sleep mode mechanism together with numerical examples are described in section 7. Our conclusion is given in section 8. Finally the Appendix with the proof of theorem 5.2 closes the paper.

2. Model and Notation.

2.1. BMAP process.

The details of $BMAP$ related definitions and notations can be found in [3]. Here we summarize only the parts, which are needed for our analysis.

The $BMAP$ batch arrival process is characterized by $\{(\Lambda(t), J(t)) ; t \geq 0\}$ bivariate continuous-time Markov chain (CTMC) on the state space $(\Lambda(t), J(t))$; where $(\Lambda(t) \in \{0, 1, \ldots\})$ denotes the number of arrivals in $(0, t]$ and $(J(t) \in \{1, 2, \ldots, L\})$ is the phase, the state of a background CTMC (phase process), at time t. The infinitesimal generator of BMAP is given as

$$\begin{pmatrix} \mathbf{D}_0 & \mathbf{D}_1 & \mathbf{D}_2 & \mathbf{D}_3 & \ldots \\ \mathbf{0} & \mathbf{D}_0 & \mathbf{D}_1 & \mathbf{D}_2 & \ldots \\ \mathbf{0} & \mathbf{0} & \mathbf{D}_0 & \mathbf{D}_1 & \ldots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{D}_0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

where $\mathbf{0}$ and $\{\mathbf{D}_k; k \geq 0\}$ are $L \times L$ matrices.

$\mathbf{D}_0$ and $\{\mathbf{D}_k; k \geq 1\}$ govern the transitions corresponding to no arrivals and to batch arrivals with size k, respectively. The irreducible infinitesimal generator of the phase process is $\mathbf{D} = \sum_{k=0}^{\infty} \mathbf{D}_k$. Let $\boldsymbol{\pi}$ be the stationary probability vector of the phase process. Then $\boldsymbol{\pi}\mathbf{D} = \mathbf{0}$ and $\boldsymbol{\pi}\mathbf{e} = 1$ uniquely determine $\boldsymbol{\pi}$, where $\mathbf{e}$ is the column vector having all elements equal to one. $\widehat{\mathbf{D}}(z)$, the matrix generating function (matrix GF) of $\mathbf{D}_k$ is defined as

$$\widehat{\mathbf{D}}(z) = \sum_{k=0}^{\infty} \mathbf{D}_k z^k, \quad |z| \leq 1. \tag{1}$$

The stationary arrival rate of the BMAP,

$$\lambda = \boldsymbol{\pi} \left. \frac{d}{dz} \widehat{\mathbf{D}}(z) \right|_{z=1} \mathbf{e} = \boldsymbol{\pi} \sum_{k=0}^{\infty} k \mathbf{D}_k \mathbf{e}, \tag{2}$$

is supposed to be positive and finite.

2.2. **The BMAP/G/1 queue with dependent multiple vacation and exhaustive service.** Batch of customers arrive to the infinite buffer queue according to a $BMAP$ process defined by $\widehat{\mathbf{D}}(z)$. The service times are independent and identically distributed. B, $B(t)$, $\widetilde{B}(s)$, b, $b^{(2)}$, $b^{(3)}$ denote the service time r.v., its cumulated distribution function, its LST and its first three moments, respectively. The mean service time is positive and finite, $0 < b < \infty$. Due to the exhaustive service the customers are served until the queue becomes empty. Then the server takes the first vacation period. If the server, upon return from the r-th ($r \geq 1$) vacation period, finds the queue empty then it immediately takes the next vacation period, whose length depends on the number of preceding vacation periods. We call the model with this vacation strategy as *dependent multiple vacation model*. We define the *total vacation period* as the sum of all vacation periods until the next service. In addition we define the *cycle time* as a service period and the total vacation period together. The server utilization is $\rho = \lambda b$.

For every $r \geq 1$ the consecutive r-th vacation periods are independent and identically distributed. Thus let V_r, $V_r(t)$, v_r denote the length of the r-th ($r \geq 1$) vacation period, its cumulated distribution function and its mean, respectively. $\widetilde{V}_r(s)$ denotes the LST of V_r, which is defined as $\widetilde{V}_r(s) = \int_{t=0}^{\infty} e^{-st} dV_r(t)$. The arrival process, the customer service times and the vacation periods are mutually independent. The service is nonpreemtive. The FCFS scheduling is applied.

Although the length of total vacation period depends only on the phase of the $BMAP$ process at the start of total vacation period, the length of an interval until an arbitrary instant in total vacation period also depends on the whole arrival process. However the length of any interval inside of the r-th vacation period ($r \geq 1$) is independent of the arrival process, as r already implicitly includes a condition on the arrival process. Therefore, in order to utilize this independency, the description of the internal structure of the total vacation period is necessary.

In the following $[Y]_{i,j}$ stands for the i,j-th element of matrix $\mathbf{Y}$. Similarly $[y]_j$ denotes the j-th element of vector $\mathbf{y}$.

We define matrix $\mathbf{A}_k$, whose (i,j)-th element denotes the conditional probability that during a customer service time the number of arrivals is k and the initial and final phases of the $BMAP$ are i and j, respectively. That is, for $k \geq 0$, $1 \leq i, j \leq L$,

$$[\mathbf{A}_k]_{i,j} = P\{\Lambda(B) = k, J(B) = j | J(0) = i\}.$$

The matrix GF $\widehat{\mathbf{A}}(z)$ is defined as $\widehat{\mathbf{A}}(z) = \sum_{k=1}^{\infty} \mathbf{A}_k z^k$. $\widehat{\mathbf{A}}(z)$ can be expressed explicitly as [3]

$$\widehat{\mathbf{A}}(z) = \int_{t=0}^{\infty} e^{\widehat{\mathbf{D}}(z)t} dB(t). \tag{3}$$

Since matrix $\widehat{\mathbf{A}}(1)$ is stochastic, we assume that $\widehat{\mathbf{A}}(z)$ can be inverted for $|z| \leq 1$.

To describe the arrivals during the r-th vacation period, for $r \geq 1$, we define matrices $\mathbf{U}_{r,k}$, whose (i,j)-th element, for $k \geq 0$, $1 \leq i,j \leq L$, is given as $[\mathbf{U}_{r,k}]_{i,j} = P\{\Lambda(V_r) = k, J(V_r) = j | J(0) = i\}$. The matrix GFs,
$\widehat{\mathbf{U}}_r(z) = \sum_{k=0}^{\infty} \mathbf{U}_{r,k} z^k$, are given as

$$\widehat{\mathbf{U}}_r(z) = \int_{t=0}^{\infty} e^{\widehat{\mathbf{D}}(z)t} dV_r(t). \tag{4}$$

Similarly to describe the arrivals during the total vacation period, we define matrices $\mathbf{U}_{(k)}$. Let V denote the length of the total vacation period. The matrices $\mathbf{U}_{(k)}$ are defined by their (i,j)-th elements, for $k \geq 1$, $1 \leq i,j \leq L$, as $[\mathbf{U}_{(k)}]_{i,j} = P\{\Lambda(V) = k, J(V) = j | J(0) = i\}$. Using them the matrix GF of the number of arriving customers during the total vacation period is defined as

$$\widehat{\mathbf{U}}(z) = \sum_{k=1}^{\infty} \mathbf{U}_{(k)} z^k. \tag{5}$$

The case when the r-th vacation period occurs is described by means of matrix $\prod_{k=1}^{r-1} \widehat{\mathbf{U}}_k(0)$, whose (i,j)-th element denotes the conditional probability that during the first $r-1$ vacation periods no arrivals occur and the phases of the $BMAP$ at the start of the first vacation and at the end of the $r-1$-th vacation are i and j, respectively. We remark here that the empty product of matrices equals the unity matrix, which is denoted by $\mathbf{I}$. The model implies that in case of exactly r vacation periods definitely there is at least one arrival in the last vacation period and it is the only vacation period having arrival. Consequently the partial matrix GF of the number of customers arriving during the total vacation period consisting of exactly r vacation periods can be expressed by $\prod_{k=1}^{r-1} \widehat{\mathbf{U}}_k(0) (\widehat{\mathbf{U}}_r(z) - \widehat{\mathbf{U}}_r(0))$. Summing up over r results in the matrix GF of the number of customers arriving during the total vacation period as

$$\widehat{\mathbf{U}}(z) = \sum_{r=1}^{\infty} \prod_{k=1}^{r-1} \widehat{\mathbf{U}}_k(0) (\widehat{\mathbf{U}}_r(z) - \widehat{\mathbf{U}}_r(0)). \tag{6}$$

We define matrix $\widetilde{\boldsymbol{\mathcal{V}}}(s)$, which is related to the LST of the last vacation period, as

$$\widetilde{\boldsymbol{\mathcal{V}}}(s) = \sum_{r=1}^{\infty} \prod_{k=1}^{r-1} \widehat{\mathbf{U}}_k(0) \left(\widetilde{V}_r(s) \mathbf{I} - \widehat{\mathbf{U}}_r(0) \right), \tag{7}$$

Note that $\widetilde{\boldsymbol{\mathcal{V}}}(0) = \mathbf{I}$, since $\widetilde{V}_r(0) = 1$ and $\prod_{k=1}^{\infty} \widehat{\mathbf{U}}_k(0) = \mathbf{0}$, because matrices $\widehat{\mathbf{U}}_k(1)$ are stochastic for $k \geq 1$.

Let t_ℓ^m denote the start of total vacation period in the ℓ-th cycle. The probability vector $\mathbf{m}$, is defined by its elements as

$$[\mathbf{m}]_j = \lim_{\ell\to\infty} P\left\{J(t_\ell^m) = j\right\}.$$

$\mathbf{m}$ is interpreted as the stationary probability vector of the phase process at starts of total vacation periods.

We define the vectors $\boldsymbol{p}_r$, $r \geq 1$, by their i-th entry, which is the probability that during a total vacation period, there are at least r vacation periods, and the phase of BMAP at the start of the r-th vacation period is i. The vectors $\boldsymbol{p}_r$ are given as

$$\boldsymbol{p}_r = \mathbf{m}\prod_{k=1}^{r-1}\widehat{\mathbf{U}}_k\left(0\right). \tag{8}$$

The mean total vacation period, which is denoted by v, can be expressed by the help of $\boldsymbol{p}_r$, $r \geq 1$ as

$$v = \sum_{r=1}^{\infty} v_r \boldsymbol{p}_r \mathbf{e} = \mathbf{m}\sum_{r=1}^{\infty} v_r \prod_{k=1}^{r-1}\widehat{\mathbf{U}}_k\left(0\right)\mathbf{e}. \tag{9}$$

The stability of the model requires that the mean cycle time is finite. This directly implies that also the mean total vacation period must be finite. This leads to

$$v = \mathbf{m}\sum_{r=1}^{\infty} v_r \prod_{k=1}^{r-1}\widehat{\mathbf{U}}_k\left(0\right)\mathbf{e} < \infty. \tag{10}$$

Under this condition the model is stable if and only if $\rho < 1$.

3. **The joint transforms.** In this section we derive expressions of joint transforms, which are needed to get the LST of the stationary virtual waiting time.

Let $N(t)$ be the number of customers in the system at time t. We introduce $F^v(t)$, which is the forward recurrence vacation time at time t in total vacation period, given that there is a virtual arrival at time t. It is defined as the interval from time t until the end of the total vacation period. The vector joint transform, $\widehat{\mathbf{q}}^v\left(z,s\right)$, is defined by its elements as

$$\begin{aligned}[\widehat{\mathbf{q}}^v\left(z,s\right)]_j = \lim_{t\to\infty}\sum_{n=0}^{\infty}\int_{\tau=0}^{\infty} e^{-s\tau} dP\big\{\ F^v\left(t\right) \leq \tau, N\left(t\right) = n,\\ J(t) = j \mid t \in t.v.p.\big\} z^n, \quad |z| \leq 1, \quad Re(s) \geq 0,\end{aligned}$$

where t.v.p. stands for total vacation period. The $\widehat{\mathbf{q}}^v\left(z,s\right)$ is interpreted as the vector joint transform of the number of customers in the system and the forward recurrence vacation time at an arbitrary instant in total vacation period.

Similarly we introduce $F^c(t)$, which is the forward recurrence customer service time at time t in a service period, given that there is a virtual arrival at time t. It is defined as the interval from time t until the end of the service of the customer, which is under service at time t. The vector joint transform, $\widehat{\mathbf{q}}^c\left(z,s\right)$, is defined by its elements as

$$[\widehat{\mathbf{q}}^{c}(z,s)]_j = \lim_{t\to\infty}\sum_{n=0}^{\infty}\int_{\tau=0}^{\infty} e^{-s\tau} dP\{\ F^{c}(t) \le \tau, N(t) = n, \\ J(t) = j \mid t \in s.p.\}z^n, \quad |z| \le 1, \quad Re(s) \ge 0,$$

where s.p. stands for service period. The $\widehat{\mathbf{q}}^{c}(z,s)$ is interpreted as the vector joint transform of the number of customers in the system and the forward recurrence customer service time at an arbitrary instant in service period.

3.1. **Joint transform in total vacation period.**

Theorem 3.1. *The vector joint transform of the number of customers in the system and the forward recurrence vacation time at an arbitrary instant in total vacation period is given as*

$$\widehat{\mathbf{q}}^{v}(z,s)\left(\widehat{\mathbf{D}}(z) + s\mathbf{I}\right) = \frac{\mathbf{m}\left(\widehat{\mathbf{U}}(z) - \widetilde{\boldsymbol{\mathcal{V}}}(s)\right)}{v}. \tag{11}$$

Proof. We introduce the vectors $\boldsymbol{p}_r^*$, $r \ge 1$, by their i-th entry, which is the probability that a random epoch in total vacation period (consisting of at least r vacation periods) belongs to r-th vacation period and the phase of BMAP at the start of r-th vacation period is i. Let us consider the Semi-Markov process in total vacation period ($t \ge 0$), whose state at time t composes from the phase of BMAP at the start of current vacation period (e.g. the r-th) and the index of the this vacation period (in that case r). Then $[\boldsymbol{p}_r]_i$ describes the probability of state (i,r) of the Markov chain embedded at starts of vacation periods and $\boldsymbol{p}_r^*$ is exactly the equilibrium distribution of the Semi-Markov process. Therefore vectors $\boldsymbol{p}_r^*$ can be expressed as

$$\boldsymbol{p}_r^* = \frac{v_r \boldsymbol{p}_r}{v} = \frac{v_r \mathbf{m}\prod_{k=1}^{r-1}\widehat{\mathbf{U}}_k(0)}{v}. \tag{12}$$

First we express $\widehat{\mathbf{q}}_r^{v}(z,s)$, which is the partial vector joint transform of the number of customers in the system and the forward recurrence vacation time at an arbitrary instant in the r-th vacation period for $r \ge 1$.

The vector GF of the stationary number of customers in the system at instant, when time τ ellapsed in r-th vacation period, is $\boldsymbol{p}_r^*\, e^{\widehat{\mathbf{D}}(z)\tau}$. The first term captures that a random epoch belongs to the r-th vacation period and the phase probability vector at the beginning of the r-th vacation period. The second term stands for the number of customers arriving in the $(0,\tau)$ interval of the r-th vacation period. The forward recurrence vacation time at instant τ equals $t - \tau$, where t is the length of the r-th vacation period. This is because the definition of forward recurrence vacation time includes a virtual arrival at time τ. To obtain the partial vector joint transform $\widehat{\mathbf{q}}_r^{v}(z,s)$ we need to take the LST of forward recurrence vacation time over the range of τ and to average the generating function of the stationary number of customers in the system over the duration of the r-th vacation period. This yields

$$\widehat{\mathbf{q}}_r^{v}(z,s) = \frac{\boldsymbol{p}_r^* \int_{t=0}^{\infty}\int_{\tau=0}^{t} e^{-s(t-\tau)} e^{\widehat{\mathbf{D}}(z)\tau} d\tau\ dV_r(t)}{v_r}. \tag{13}$$

Multiplying both sides of (13) by $\left(\widehat{\mathbf{D}}(z) + s\mathbf{I}\right)$ we have

$$\widehat{\mathbf{q}}_r^v(z,s)\ \left(\widehat{\mathbf{D}}(z)+s\mathbf{I}\right)= \\ \frac{\boldsymbol{p}_r^*}{v_r}\int_{t=0}^{\infty}e^{-st}\int_{\tau=0}^{t}e^{\left(\widehat{\mathbf{D}}(z)+s\mathbf{I}\right)\tau}\left(\widehat{\mathbf{D}}(z)+s\mathbf{I}\right)d\tau\ dV_r(t). \tag{14}$$

The internal integral term can be rewritten as

$$\begin{aligned}&\int_{\tau=0}^{t}e^{\left(\widehat{\mathbf{D}}(z)+s\mathbf{I}\right)\tau}\ \left(\widehat{\mathbf{D}}(z)+s\mathbf{I}\right)\ d\tau\\ &=\int_{\tau=0}^{t}\sum_{k=0}^{\infty}\frac{\tau^k\left(\widehat{\mathbf{D}}(z)+s\mathbf{I}\right)^k}{k!}\ \left(\widehat{\mathbf{D}}(z)+s\mathbf{I}\right)\ d\tau\\ &=\sum_{k=0}^{\infty}\int_{\tau=0}^{t}\tau^k\ d\tau\ \frac{\left(\widehat{\mathbf{D}}(z)+s\mathbf{I}\right)^{k+1}}{k!}\\ &=\sum_{k=0}^{\infty}\frac{t^{k+1}}{k+1}\ \frac{\left(\widehat{\mathbf{D}}(z)+s\mathbf{I}\right)^{k+1}}{k!}=e^{\left(\widehat{\mathbf{D}}(z)+s\mathbf{I}\right)t}-\mathbf{I}.\end{aligned} \tag{15}$$

Substituting (15) into (14), applying (4) and rearranging yields

$$\begin{aligned}\widehat{\mathbf{q}}_r^v(z,s)\ \left(\widehat{\mathbf{D}}(z)+s\mathbf{I}\right)&=\frac{\boldsymbol{p}_r^*}{v_r}\int_{t=0}^{\infty}\left(e^{\widehat{\mathbf{D}}(z)t}-e^{-st}\mathbf{I}\right)\ dV_r(t)\\ &=\frac{\boldsymbol{p}_r^*\left(\widehat{\mathbf{U}}_r(z)-\widetilde{V}_r(s)\,\mathbf{I}\right)}{v_r}.\end{aligned} \tag{16}$$

The joint transform $\widehat{\mathbf{q}}^v(z,s)$ is given as $\widehat{\mathbf{q}}^v(z,s)=\sum_{r=1}^{\infty}\widehat{\mathbf{q}}_r^v(z,s)$, from which

$$\widehat{\mathbf{q}}^v(z,s)\ \left(\widehat{\mathbf{D}}(z)+s\mathbf{I}\right)=\sum_{r=1}^{\infty}\frac{\boldsymbol{p}_r^*\left(\widehat{\mathbf{U}}_r(z)-\widetilde{V}_r(s)\,\mathbf{I}\right)}{v_r}. \tag{17}$$

Applying (12) and rearranging results in

$$\begin{aligned}&\widehat{\mathbf{q}}^v(z,s)\ \left(\widehat{\mathbf{D}}(z)+s\mathbf{I}\right)\\ &=\frac{\mathbf{m}\sum_{r=1}^{\infty}\prod_{k=1}^{r-1}\widehat{\mathbf{U}}_k(0)\left(\widehat{\mathbf{U}}_r(z)-\widetilde{V}_r(s)\,\mathbf{I}\right)}{v}\\ &=\frac{\mathbf{m}\sum_{r=1}^{\infty}\prod_{k=1}^{r-1}\widehat{\mathbf{U}}_k(0)\left(\widehat{\mathbf{U}}_r(z)-\widehat{\mathbf{U}}_r(0)\right)}{v}\\ &-\frac{\mathbf{m}\sum_{r=1}^{\infty}\prod_{k=1}^{r-1}\widehat{\mathbf{U}}_k(0)\left(\widetilde{V}_r(s)\,\mathbf{I}-\widehat{\mathbf{U}}_r(0)\right)}{v}.\end{aligned} \tag{18}$$

The statement comes by applying (6) and (7) in (18). □

3.2. **Joint transform in service period.** Let $G(\ell)$ denote the number of customer services during the ℓ-th cycle, for $\ell \geq 1$. Additionally $t^s(\ell, r)$ denotes the instants of service start of the r-th customer in the ℓ-th cycle, for $\ell \geq 1$ and $1 \leq r \leq G(\ell)$. We define the vector GF of the stationary number of customers at service start epochs $\widehat{\mathbf{q}}^s(z)$ by its elements as

$$[\widehat{\mathbf{q}}^s(z)]_j = \lim_{\ell\to\infty} \sum_{n=0}^{\infty} \frac{\sum_{r=1}^{G(\ell)} P\left\{N(t^s(\ell,r)) = n, J(t^s(\ell,r)) = j\right\}}{E[G(\ell)]} z^n, \quad |z| \leq 1.$$

Theorem 3.2. *The vector joint transform of the number of customers in the system and the forward recurrence customer service time at an arbitrary instant in service period is given as*

$$\widehat{\mathbf{q}}^c(z,s)\left(\widehat{\mathbf{D}}(z) + s\mathbf{I}\right) = \frac{\widehat{\mathbf{q}}^s(z)\left(\widehat{\mathbf{A}}(z) - \widetilde{B}(s)\,\mathbf{I}\right)}{b}. \tag{19}$$

Proof. To get the expression of $\widehat{\mathbf{q}}^c(z,s)$ the same line of argument can be applied as for obtaining (16) to express the partial vector joint transform $\widehat{\mathbf{q}}^v_r(z,s)$. We have to replace $\boldsymbol{p}^*_r$ by $\widehat{\mathbf{q}}^s(z)$, $\widehat{\mathbf{U}}_r(z)$ by $\widehat{\mathbf{A}}(z)$, $\widetilde{V}_r(s)$ by $\widetilde{B}(s)$ and v_r by b and it results in the statement. □

In the next proposition we give the expression of $\widehat{\mathbf{q}}^s(z)$, the only unknown in (19).

Proposition 1. *The vector GF of the stationary number of customers at customer service start epochs can be expressed as*

$$\lambda\widehat{\mathbf{q}}^s(z)\left(z\mathbf{I} - \widehat{\mathbf{A}}(z)\right) = (1-\rho)\,z\frac{\mathbf{m}\left(\widehat{\mathbf{U}}(z) - \mathbf{I}\right)}{v}. \tag{20}$$

Proof. Let $t^d(\ell, r)$ denote the instants at the departure of the r-th customer in the ℓ-th cycle, for $\ell \geq 1$ and $1 \leq r \leq G(\ell)$. Similar to $\widehat{\mathbf{q}}^s(z)$ we also define the vector GF of the stationary number of customers at customer departure epochs $\widehat{\mathbf{q}}^d(z)$ by its elements as

$$[\widehat{\mathbf{q}}^d(z)]_j = \lim_{\ell\to\infty} \sum_{n=0}^{\infty} \frac{\sum_{r=1}^{G(\ell)} P\left\{N(t^d(\ell,r)) = n, J(t^d(\ell,r)) = j\right\}}{E[G(\ell)]} z^n, \quad |z| \leq 1.$$

Now we relate $\widehat{\mathbf{q}}^d(z)$ to $\widehat{\mathbf{q}}^s(z)$. The number of customers just before an arbitrary departure epoch equals the number of customers at previous customer service start plus the number of customers arriving during that service. This leads to the following $BMAP$ specific relation:

$$z\widehat{\mathbf{q}}^d(z) = \widehat{\mathbf{q}}^s(z)\,\widehat{\mathbf{A}}(z). \tag{21}$$

We also define the vector GF of the stationary number of customers at an arbitrary instant $\widehat{\mathbf{q}}(z)$ by its elements as

$$[\widehat{\mathbf{q}}(z)]_j = \lim_{t\to\infty} \sum_{n=0}^{\infty} P\{N(t) = n, J(t) = j\} z^n, |z| \leq 1.$$

Takine and Takahashi proved a stationary relationship between $\widehat{\mathbf{q}}(z)$ and $\widehat{\mathbf{q}}^d(z)$ [18], which is given as

$$\widehat{\mathbf{q}}(z)\widehat{\mathbf{D}}(z) = \lambda(z-1)\widehat{\mathbf{q}}^d(z). \tag{22}$$

Finally we also need the factorization formula of Chang et al. [19], which is written as

$$\widehat{\mathbf{q}}(z)\left(z\mathbf{I} - \widehat{\mathbf{A}}(z)\right) = \widehat{\mathbf{q}}^v(z)(1-\rho)(z-1)\widehat{\mathbf{A}}(z). \tag{23}$$

Post-multiplying (23) by $z\widehat{\mathbf{D}}(z)$, utilizing that $\widehat{\mathbf{A}}(z)$ and $\widehat{\mathbf{D}}(z)$ commute as well as applying (22) and (21) leads to

$$\begin{aligned}&\lambda(z-1)\widehat{\mathbf{q}}^s(z)\left(z\mathbf{I} - \widehat{\mathbf{A}}(z)\right)\widehat{\mathbf{A}}(z) = \\ &(1-\rho)(z-1)z\widehat{\mathbf{q}}^v(z)\widehat{\mathbf{D}}(z)\widehat{\mathbf{A}}(z).\end{aligned} \tag{24}$$

Utilizing that the quantities occurring in (24) are continuous at $z = 1$ and post-multiplying both sides by $\left(\widehat{\mathbf{A}}(z)\right)^{-1}$ for $|z| \leq 1$ yields

$$\lambda\widehat{\mathbf{q}}^s(z)\left(z\mathbf{I} - \widehat{\mathbf{A}}(z)\right) = (1-\rho)z\,\widehat{\mathbf{q}}^v(z)\widehat{\mathbf{D}}(z). \tag{25}$$

Setting $s = 0$ in (11) gives $\widehat{\mathbf{q}}^v(z)$ as

$$\widehat{\mathbf{q}}^v(z)\widehat{\mathbf{D}}(z) = \frac{\mathbf{m}\left(\widehat{\mathbf{U}}(z) - \mathbf{I}\right)}{v}. \tag{26}$$

Applying (26) in (25) results in the statement. □

4. **Stationary number of customers.** In this section we provide expressions for the vector GF of the stationary number of customers at an arbitrary instant and for its mean.

Theorem 4.1. *The vector GF of the stationary number of customers at an arbitrary instant can be expressed as*

$$\widehat{\mathbf{q}}(z)\widehat{\mathbf{D}}(z)\left(z\mathbf{I} - \widehat{\mathbf{A}}(z)\right) = \frac{\mathbf{m}\left(\widehat{\mathbf{U}}(z) - \mathbf{I}\right)}{v}(1-\rho)(z-1)\widehat{\mathbf{A}}(z). \tag{27}$$

Proof. Multiplying (23) by $\widehat{\mathbf{D}}(z)$ and utilizing again that $\widehat{\mathbf{A}}(z)$ and $\widehat{\mathbf{D}}(z)$ commute leads to:

$$\widehat{\mathbf{q}}(z)\widehat{\mathbf{D}}(z)\left(z\mathbf{I} - \widehat{\mathbf{A}}(z)\right) = \widehat{\mathbf{q}}^v(z)\widehat{\mathbf{D}}(z)(1-\rho)(z-1)\widehat{\mathbf{A}}(z). \tag{28}$$

Applying (26) in (28) results in the statement. □

We introduce the notations $\mathbf{D}^{(i)}$, $\mathbf{A}^{(i)}$, $\mathbf{U}^{(i)}$ and $\mathbf{q}^{(i)}$, $i \geq 1$ for the i-th derivatives of $\widehat{\mathbf{D}}(z)$, $\widehat{\mathbf{A}}(z)$, $\widehat{\mathbf{U}}(z)$, $\widehat{\mathbf{q}}(z)$ and $\widehat{\mathbf{m}}(z)$ at $z = 1$, respectively. We also use the notations $\mathbf{D} = \widehat{\mathbf{D}}(1)$, $\mathbf{A} = \widehat{\mathbf{A}}(1)$, $\mathbf{U} = \widehat{\mathbf{U}}(1)$ and $\mathbf{q} = \widehat{\mathbf{q}}(1)$.

Theorem 4.2. *The mean of the stationary number of customers at an arbitrary instant is given by*

$$\begin{aligned}
\mathbf{q}^{(1)} &= \frac{\mathbf{m}}{\lambda v}\left(\frac{1}{2}\mathbf{U}^{(2)}\mathbf{e}\boldsymbol{\pi} + \frac{1}{2}(\mathbf{U}-\mathbf{I})\mathbf{A}^{(2)}\mathbf{e}\boldsymbol{\pi} + \mathbf{U}^{(1)}\mathbf{A}^{(1)}\mathbf{e}\boldsymbol{\pi}\right) \\
&- \frac{\mathbf{m}}{\lambda v}\left(\mathbf{U}^{(1)}\mathbf{A} + (\mathbf{U}-\mathbf{I})\mathbf{A}^{(1)}\right)(\mathbf{D}+\mathbf{e}\boldsymbol{\pi})^{-1}\mathbf{D}^{(1)}\mathbf{e}\boldsymbol{\pi} \\
&+ \frac{\mathbf{m}}{\lambda v}\left(\mathbf{U}^{(1)}\mathbf{A}\mathbf{e}\boldsymbol{\pi} + (\mathbf{U}-\mathbf{I})\mathbf{A}^{(1)}\mathbf{e}\boldsymbol{\pi}\right)\left(\frac{\mathbf{C_2}\mathbf{e}\boldsymbol{\pi}}{\lambda} + (1-\rho)\mathbf{C_1}\right) \\
&+ \frac{\mathbf{m}}{\lambda v}(\mathbf{U}-\mathbf{I})\mathbf{A}(\mathbf{D}+\mathbf{e}\boldsymbol{\pi})^{-1}\left(\lambda\mathbf{I} - \mathbf{D}^{(1)}\mathbf{e}\boldsymbol{\pi}\right)\left(\frac{\mathbf{C_2}\mathbf{e}\boldsymbol{\pi}}{\lambda} + (1-\rho)\mathbf{C_1}\right) \\
&+ \boldsymbol{\pi}\left(\frac{\mathbf{A}^{(2)}\mathbf{e}\boldsymbol{\pi}}{2(1-\rho)} - \left(\mathbf{I} - \mathbf{A}^{(1)}\right)\mathbf{C_1}\right),
\end{aligned} \tag{29}$$

where matrices $\mathbf{C_1}$ *and* $\mathbf{C_2}$ *are defined as*

$$\mathbf{C_1} = (\mathbf{I} - \mathbf{A} + \mathbf{e}\boldsymbol{\pi})^{-1}\left(\frac{\mathbf{A}^{(1)}\mathbf{e}\boldsymbol{\pi}}{(1-\rho)} + \mathbf{I}\right), \quad \mathbf{C_2} = \mathbf{D}^{(1)}(\mathbf{D}+\mathbf{e}\boldsymbol{\pi})^{-1}\mathbf{D}^{(1)} - \frac{1}{2}\mathbf{D}^{(2)}.$$

Proof. Starting from relation (28) and applying the arguments of the corresponding theorem in [7] gives the theorem. □

5. **Stationary virtual waiting time.** The virtual waiting time is the time period, which an arriving customer would experience at a time t until the start of its service. Note that there is not necessarily an arrival at time t, that is why it is called as virtual.

The virtual waiting time depends on the phase of the $BMAP$. Let $W(\tau)$ be the virtual waiting time in the system at time τ. We define the vector cumulated distribution function of the stationary virtual waiting time, $\mathbf{w}(t)$, by its elements as

$$[\mathbf{w}(t)]_j = \lim_{\tau\to\infty} P\{W(\tau) \leq t, J(\tau) = j\}.$$

The vector LST of the stationary virtual waiting time is defined as

$$\widetilde{\mathbf{w}}(s) = \int_{t=0}^{\infty} e^{-st} d\mathbf{w}(t), \quad Re(s) \geq 0.$$

5.1. **LST of stationary virtual waiting time.**

Theorem 5.1. *The vector LST of stationary virtual waiting time can be expressed as*

$$\widetilde{\mathbf{w}}(s)\left(\widehat{\mathbf{D}}(\widetilde{B}(s)) + s\mathbf{I}\right) = (1-\rho)\frac{\mathbf{m}}{v}\left(\mathbf{I} - \widetilde{\boldsymbol{\mathcal{V}}}(s)\right). \tag{30}$$

Proof. Our argument to get the LST of stationary virtual waiting time is based on the unfinished work in the system.

An arriving customer sees the system with probability ρ in service period and with probability $1-\rho$ in vacation period.

Due to FCFS scheduling the waiting time at arrival during the service period is exactly the unfinished work, i.e. it consists of the forward recurrence customer service time of the customer currently under service and the customer service times of the customers, who are already present in the queue at virtual arrival. Note that the number of those customers is one less than the number of customers in the system. Similarly the waiting time at virtual arrival during the vacation period consists of the forward recurrence vacation time and the customer service times of the customers, who are already present at virtual arrival (unfinished work). This yields

$$\widetilde{\mathbf{w}}(s) = \left(\rho \frac{\widehat{\mathbf{q}}^c(z,s)}{z} + (1-\rho)\ \widehat{\mathbf{q}}^v(z,s) \right) \Big|_{z=\widetilde{B}(s)}. \tag{31}$$

Setting $z = \widetilde{B}(s)$ in (19) and (20) gives

$$\begin{aligned} &\widehat{\mathbf{q}}^c(z,s)\left(\widehat{\mathbf{D}}(z)+s\mathbf{I}\right)\Big|_{z=\widetilde{B}(s)} = \\ &\frac{\widehat{\mathbf{q}}^s\left(\widetilde{B}(s)\right)\left(\widehat{\mathbf{A}}(\widetilde{B}(s)) - \widetilde{B}(s)\,\mathbf{I}\right)}{b}, \end{aligned} \tag{32}$$

$$\begin{aligned} &\lambda\widehat{\mathbf{q}}^s\left(\widetilde{B}(s)\right)\left(\widetilde{B}(s)\,\mathbf{I} - \widehat{\mathbf{A}}(\widetilde{B}(s))\right) = \\ &(1-\rho)\,\widetilde{B}(s)\,\frac{\mathbf{m}\left(\widehat{\mathbf{U}}(\widetilde{B}(s)) - \mathbf{I}\right)}{v}, \end{aligned} \tag{33}$$

respectively. Combining them leads to

$$\begin{aligned} &\rho\frac{\widehat{\mathbf{q}}^c(z,s)}{z}\left(\widehat{\mathbf{D}}(z)+s\mathbf{I}\right)\Big|_{z=\widetilde{B}(s)} = \\ &-(1-\rho)\,\frac{\mathbf{m}\left(\widehat{\mathbf{U}}(\widetilde{B}(s)) - \mathbf{I}\right)}{v}. \end{aligned} \tag{34}$$

Multiplying (31) by $\left(\widehat{\mathbf{D}}(\widetilde{B}(s)) + s\mathbf{I}\right)$ and applying (34) as well as (11) yields

$$\begin{aligned} &\widetilde{\mathbf{w}}(s)\left(\widehat{\mathbf{D}}(\widetilde{B}(s)) + s\mathbf{I}\right) = -(1-\rho)\,\frac{\mathbf{m}}{v}\left(\widehat{\mathbf{U}}(\widetilde{B}(s)) - \mathbf{I}\right) \\ &+(1-\rho)\,\frac{\mathbf{m}}{v}\left(\widehat{\mathbf{U}}(\widetilde{B}(s)) - \widetilde{\boldsymbol{\mathcal{V}}}(s)\right). \end{aligned} \tag{35}$$

Rearranging (35) results in the statement. □

5.2. **First two moments of stationary virtual waiting time.** Let $\mathbf{w}^{(k)}$ and $\boldsymbol{\mathcal{V}}^{(k)}$ denote the k-th ($k \geq 1$) moment of LSTs $\widetilde{\mathbf{w}}(s)$ and $\widetilde{\boldsymbol{\mathcal{V}}}(s)$, for $Re(s) \geq 0$, respectively. Thus $\mathbf{w}^{(k)} = (-1)^k \frac{d^k}{ds^k}\widetilde{\mathbf{w}}(s)\,|_{s=0}$ and $\boldsymbol{\mathcal{V}}^{(k)} = (-1)^k \frac{d^k}{ds^k}\widetilde{\boldsymbol{\mathcal{V}}}(s)\,|_{s=0}$.

Theorem 5.2. *The first two vector moments of the stationary virtual waiting time can be expressed as*

$$\mathbf{w}^{(1)} = \frac{\mathbf{m}}{v}\frac{\boldsymbol{\mathcal{V}}^{(2)}\mathbf{e}\boldsymbol{\pi}}{2} - (1-\rho)\,\frac{\mathbf{m}}{v}\boldsymbol{\mathcal{V}}^{(1)}\mathbf{C}_3 + \boldsymbol{\pi}\mathbf{C}_4, \tag{36}$$

$$\begin{aligned}\mathbf{w}^{(2)} &= \frac{\mathbf{m}}{v}\frac{\boldsymbol{\mathcal{V}}^{(3)}\mathbf{e}\boldsymbol{\pi}}{3} \\ &+ \frac{\mathbf{m}}{v}\left(\boldsymbol{\mathcal{V}}^{(2)}\mathbf{e}\boldsymbol{\pi}\mathbf{C}_4 - (1-\rho)\,\boldsymbol{\mathcal{V}}^{(2)}\mathbf{C}_3\right) \\ &- 2\,(1-\rho)\,\frac{\mathbf{m}}{v}\boldsymbol{\mathcal{V}}^{(1)}\mathbf{C}_3\mathbf{C}_4 \\ &+ \boldsymbol{\pi}\left(2\mathbf{C}_4\mathbf{C}_4 - \left(b^2\mathbf{D}^{(2)} + b^{(2)}\mathbf{D}^{(1)}\right)\mathbf{C}_3\right) \\ &+ \boldsymbol{\pi}\frac{b^3\mathbf{D}^{(3)}\mathbf{e}\boldsymbol{\pi} + 3bb^{(2)}\mathbf{D}^{(2)}\mathbf{e}\boldsymbol{\pi} + b^{(3)}\mathbf{D}^{(1)}\mathbf{e}\boldsymbol{\pi}}{3\,(1-\rho)},\end{aligned} \tag{37}$$

where matrices $\mathbf{C}_3$ *and* $\mathbf{C}_4$ *are defined as*

$$\begin{aligned}\mathbf{C}_3 &= (\mathbf{D}+\mathbf{e}\boldsymbol{\pi})^{-1}\left(\mathbf{I} - \frac{\left(\mathbf{I} - b\mathbf{D}^{(1)}\right)\mathbf{e}\boldsymbol{\pi}}{(1-\rho)}\right), \\ \mathbf{C}_4 &= \frac{\left(b^2\mathbf{D}^{(2)} + b^{(2)}\mathbf{D}^{(1)}\right)\mathbf{e}\boldsymbol{\pi}}{2\,(1-\rho)} + \left(\mathbf{I} - b\mathbf{D}^{(1)}\right)\mathbf{C}_3.\end{aligned}$$

Proof. The proof of the theorem can be found in the Appendix. □

6. **Computation of the stationary probability vector of the phase process at start of t.v.p.** In this section we give a computation method to determine the unknown $\mathbf{m}$ in (29), (36) and (37).

We define the homogenous bivariate Markov chain $\left\{(N(t_k^d), J(t_k^d)); k \in \{1,\ldots\}\right\}$ on the state space $\{0,1,\ldots\} \times \{1,2,\ldots,L\}$, where t_k^d denotes the k-th customer departure epoch for $k \geq 1$. We define matrix $\mathbf{G}$, whose (i,j)-th elements is given as the probability that starting from state $(n+1,i)$ in the Markov chain the first state visited in level n is (n,j), $n \in 0,1,2,\ldots$, $1 \leq i,j \leq L$.

Theorem 6.1. *The stationary probability vector of the phase process at starts of total vacation periods is given by*

$$\begin{aligned}\mathbf{m} &= \mathbf{e}_{\mathbf{L}}\big(\,(\mathbf{I}-\mathbf{K})\;||\;\mathbf{e}\big)^{-1}, \\ \mathbf{K} &= \sum_{r=1}^{\infty}\prod_{k=1}^{r-1}\widehat{\mathbf{U}}_k\,(0)\,\big(\widehat{\mathbf{U}}_r\,(\mathbf{G}) - \widehat{\mathbf{U}}_r\,(0)\,\big),\end{aligned} \tag{38}$$

where $\widehat{\mathbf{U}}_r\,(\mathbf{G})$ *stands for* $\sum_{k=0}^{\infty}\mathbf{U}_{r,k}\mathbf{G}^k$*, for* $r \geq 1$.

For computing matrix $\mathbf{G}$*, the only unknown in (38), the standard algorithm of Lucantoni* [3] *can be applied.*

Proof. We define matrix $\mathbf{K}$, whose (i,j)-th element is given as the probability that the Markov chain embedded at the customer departure epochs (defined above), starting from the state $(0,i)$ returns to the level 0 for the first time by hitting the state $(0,j)$.

The unknown vector $\mathbf{m}$ is the invariant probability vector of $\mathbf{K}$ and therefore it satisfies $\mathbf{mK} = \mathbf{m}$. Rearranging yields

$$\mathbf{m}\,(\mathbf{I} - \mathbf{K}) = 0. \tag{39}$$

Due to stability of the model (39) and $\mathbf{me} = 1$ uniquely determine m, which implies that matrix $(\mathbf{I} - \mathbf{K})$ has rank $L - 1$. Thus we use also the normalization condition $\mathbf{me} = 1$ to solve (39) for $\mathbf{m}$. Let $\mathbf{e_i}$ stand for the row vector, whose i-th element equals to 1 and its other elements are 0. In addition let $\mathbf{Y} \,||\, \mathbf{x}$ denote the matrix $\mathbf{Y}$ with the last column replaced by the column vector $\mathbf{x}$. Now combining normalization condition with (39) gives $\mathbf{m}$ as

$$\mathbf{m} = \mathbf{e_L}\big(\,(\mathbf{I} - \mathbf{K}) \,||\, \mathbf{e}\big)^{-1}. \tag{40}$$

Each customer arriving during the total vacation period generates a first passage described by matrix $\mathbf{G}$, and thus for $\mathbf{K}$ we get:

$$\mathbf{K} = \sum_{k=1}^{\infty} \mathbf{U}_{(k)}\mathbf{G}^k. \tag{41}$$

Applying (5), (6) and (41) results in the statement. □

7. Application to the IEEE 802.16e sleep mode mechanism.

7.1. IEEE 802.16e sleep mode mechanism.

The IEEE 802.16 standard specifies an air interface for Broadband Wireless Access (BWA). It proposes a high-speed access system supporting multimedia services and an extensive QoS guarantee. In IEEE 802.16 protocol stack the Medium Access Control (MAC) layer supports multiple Physical (PHY) layer specifications, each of them covering different operational environments.

The IEEE 802.16e sleep mode mechanism was originally specified in Corrigendum 1 [20] published along with IEEE 802.16e-2005 (amendment to IEEE 802.16-2004 Standard [21]. As a result of the recent revision of IEEE 802.16 standard, this sleep mode mechanism is incorporated into the new base standard IEEE 802.16-2009 [8], which consolidates the IEEE 802.16-2004 Standard with several amendments.

The purpose of the sleep mode mechanism is to enable a power consumption reduction at the Mobile Stations (MSs). The sleep mode mechanism utilizes the natural idle periods of the traffic, i.e. the periods without packet transmission. The MS periodically inserts *sleep intervals*, whose lengths are predetermined and negotiated with the Base Station (BS).

In the *sleep interval* the MS switches off its air interface and enters in the energy saving mode. At the end of the *sleep interval* the MS switches back for a short *listening interval* to check whether packets are waiting at BS for downlink traffic. If not then the MS enters into the next *sleep interval*. However if any packet arrived to the BS for the MS during the last *sleep interval* then the MS remains active and an *awake mode* starts (see figure 1). Thus the price for the MS power reduction

is the higher packet delay, since the packets arriving during a *sleep interval* must wait until the end of the next *listening interval*. If packets arrive to the MS for uplink during a *sleep interval*, the MS immediately interrupts the *sleep interval* and remains active until all packets are transmitted in both directions.

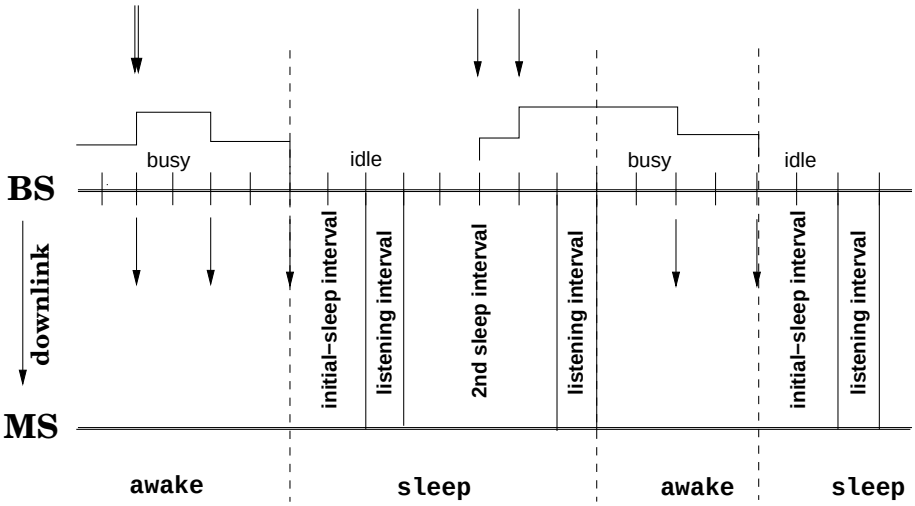

FIGURE 1. Operation of the IEEE 802.16e sleep mode mechanism.

The standard defines three types of power saving classes. In class type I starting with the *initial-sleep interval* the size of the next *sleep interval* is always doubled until reaching the *final-sleep interval*, which is then repeated. This type is recommended for the BE and nrtPS services. Class type II has fixed-length *sleep interval* and it is recommended for unsolicited grant service (UGS). Finally in class type III the *sleep interval* is negotiated only for one occasion, which is typically used for management traffic.

7.2. **Analytic model of the power saving class of type I.** We apply the presented queueing model for the power saving class of type I. This model can be applied only for the downlink traffic, so the uplink traffic is ignored. Thus the customers of the queueing model correspond to the packets sent from BS to MS. Each vacation period models the actual *sleep interval* together with the *listening interval* following it. Hence V_R, for $r \geq R$, is the sum of the fixed-length *final-sleep interval* and the fixed-length *listening interval*. However the doubling rule is relaxed on the way, which is given as:

$$\begin{aligned} V_r &= V_R, \quad r \geq R, \\ V_r &< V_R, \quad r < R. \end{aligned} \tag{42}$$

We introduce the notation $\nu = E[\text{number of vacation periods per t.v.p.}]$. Taking (42) into account the quantities $\widehat{\mathbf{U}}(z)$, $\widetilde{\boldsymbol{\mathcal{V}}}(s)$, ν and v can be expressed as

$$\widehat{\mathbf{U}}(z) = \sum_{r=1}^{R-1}\prod_{k=1}^{r-1}\widehat{\mathbf{U}}_k(0)\left(\widehat{\mathbf{U}}_r(z)-\widehat{\mathbf{U}}_r(0)\right) \tag{43}$$
$$+ \prod_{k=1}^{R-1}\widehat{\mathbf{U}}_k(0)\left(\mathbf{I}-\widehat{\mathbf{U}}_R(0)\right)^{-1}\left(\widehat{\mathbf{U}}_R(z)-\widehat{\mathbf{U}}_R(0)\right),$$

$$\widetilde{\boldsymbol{\mathcal{V}}}(s) = \sum_{r=1}^{R-1}\prod_{k=1}^{r-1}\widehat{\mathbf{U}}_k(0)\left(\widetilde{V}_r(s)\,\mathbf{I}-\widehat{\mathbf{U}}_r(0)\right) \tag{44}$$
$$+ \prod_{k=1}^{R-1}\widehat{\mathbf{U}}_k(0)\left(\mathbf{I}-\widehat{\mathbf{U}}_R(0)\right)^{-1}\left(\widetilde{V}_R(s)\,\mathbf{I}-\widehat{\mathbf{U}}_R(0)\right),$$

$$\nu = \mathbf{m}\Bigg[\sum_{r=1}^{R-1} r\prod_{k=1}^{r-1}\widehat{\mathbf{U}}_k(0)\left(\widehat{\mathbf{U}}_r(1)-\widehat{\mathbf{U}}_r(0)\right) \tag{45}$$
$$+ \prod_{k=1}^{R-1}\widehat{\mathbf{U}}_k(0)\,\widehat{\mathbf{U}}_R(0)\left(\left(\mathbf{I}-\widehat{\mathbf{U}}_R(0)\right)^{-1}\right)^2\left(\widehat{\mathbf{U}}_R(1)-\widehat{\mathbf{U}}_R(0)\right)$$
$$+ \prod_{k=1}^{R-1}\widehat{\mathbf{U}}_k(0)\,R\left(\mathbf{I}-\widehat{\mathbf{U}}_R(0)\right)^{-1}\left(\widehat{\mathbf{U}}_R(1)-\widehat{\mathbf{U}}_R(0)\right)\Bigg]\mathbf{e},$$

$$v = \mathbf{m}\Bigg[\sum_{r=1}^{R-1} v_r\prod_{k=1}^{r-1}\widehat{\mathbf{U}}_k(0) + v_R\prod_{k=1}^{R-1}\widehat{\mathbf{U}}_k(0)\left(\mathbf{I}-\widehat{\mathbf{U}}_R(0)\right)^{-1}\Bigg]\mathbf{e}. \tag{46}$$

The primary performance measure in the IEEE 802.16e sleep mode mechanism is the mean packet delay, $E[W]$, which is given by the help of $\mathbf{w}^{(1)}$ (see (36)) as

$$E[W] = \mathbf{w}^{(1)}\mathbf{e}. \tag{47}$$

Another object of interest is the savings in the energy consumption due to the *sleep intervals*. For this purpose we use the mean power savings. Let T_l stand for the length of the *listening interval*. We assume that the power consumption is the same during all active periods, i.e. during transmitting, receiving and listening. $\mathcal{P}_s$, $\mathcal{P}_a$ and $\Delta\mathcal{P}$ denote the constant power during the *sleep intervals*, the constant power during the active periods and the power savings at an arbitrary time, respectively. We also take into account the constant total extra energy needed for the switchings between the *sleep intervals* and the *listening intervals* [14], which is denoted by $\mathcal{E}_{on}$. Let β be the time fraction of the *listening intervals* in the total vacation period, which is given as

$$\beta = \frac{E[\text{number of vacation periods per t.v.p.}]\;T_l}{E[\text{length of t.v.p.}]} = \frac{\nu\,T_l}{v}.$$

The power savings arises during *sleep intervals* and it can be determined from the difference between the constant power during the active periods and the constant power during the *sleep intervals* $(\mathcal{P}_a - \mathcal{P}_s)$ as well as from a correction term due to $\mathcal{E}_{on}$. The time fraction of the *sleep intervals* equals the time fraction of the vacation $(1-\rho)$ multiplied by the time fraction of the *sleep intervals* in the total vacation period $(1-\beta)$. The mean of the total extra energy needed for

the switchings at the end of the *sleep intervals* in the total vacation period is E[number of vacation periods per t.v.p.] $\mathcal{E}_{on} = \nu\ \mathcal{E}_{on}$. Thus using the expression of β the mean power savings can be expressed as

$$\begin{aligned} E[\Delta\mathcal{P}] &= (1-\rho)\ \left((1-\beta)\ (\mathcal{P}_a - \mathcal{P}_s) - \frac{\nu\ \mathcal{E}_{on}}{v}\right) \\ &= (1-\rho)\ \left((1-\frac{\nu\ T_l}{v})\ (\mathcal{P}_a - \mathcal{P}_s) - \frac{\nu}{v}\ \mathcal{E}_{on}\right). \end{aligned} \tag{48}$$

7.3. **Modeling correlated traffic with MAP(2).** For traffic modeling we use a two-phase Markovian Arrival Process, which is referred to as $MAP(2)$. Although $MAP(2)$ is a special case of $BMAP$ it is appropriate to model a fairly general correlated traffic. Recently it was shown in [16] that every $MAP(2)$ can be transformed to a specific form, which is referred to as canonical $MAP(2)$. Hence for modeling correlated traffic with $MAP(2)$ we apply its canonical form. We summarize only selected parts from the description of the canonical form of $MAP(2)$. For more detailed description we refer to [16].

The stationary distribution of the interarrival times of $MAP(2)$ is a two states phase-type ($PH(2)$). We use the notations μ_1 and μ_2 for its first two moments. The correlation of two consecutive interarrival times are characterized by the lag-1 correlation coefficient, which is defined as

$$\mathrm{Corr}(X_0, X_1) = \frac{E[(X_0 - E[X])(X_1 - E[X])]}{\mathrm{Var}[X]} = \gamma \cdot \frac{\frac{\mu_2}{2} - \mu_1^2}{\mu_2 - \mu_1^2}, \tag{49}$$

where $-1 \le \gamma < 1$ is a correlation parameter and random variable X stands for a generic interarrival time.

Depending on the correlation characteristics of the interarrival time, there are two variants of the canonical representation. The first and the second variants of the canonical form $MAP(2)$ are given as

$$\mathbf{D}_0 = \begin{bmatrix} -\lambda_1 & (1-a^*)\lambda_1 \\ 0 & -\lambda_2 \end{bmatrix},\ \mathbf{D}_1 = \begin{bmatrix} a^*\lambda_1 & 0 \\ (1-b^*)\lambda_2 & b^*\lambda_2 \end{bmatrix}$$

and

$$\mathbf{D}_0 = \begin{bmatrix} -\lambda_1 & (1-a^*)\lambda_1 \\ 0 & -\lambda_2 \end{bmatrix},\ \mathbf{D}_1 = \begin{bmatrix} 0 & a^*\lambda_1 \\ b^*\lambda_2 & (1-b^*)\lambda_2 \end{bmatrix},$$

respectively. Here $0 < \lambda_1 \le \lambda_2$, $0 \le a^* \le 1$ and $0 \le b^* \le 1$. Additionally the parameter ranges are restricted as

- $a^*, b^* \ne 1$ in the first canonical form and
- $b^* \ne 0$ in the second canonical form and
- $\lambda_1 \ne \lambda_2$, if $a^* = 1$ in the second canonical form.

For correlated processes, a^* and b^* must be nonzero.

The correlation parameter γ and the phase probability vector at stationary arrival epochs π depend only on parameters a^* and b^* as

$$\begin{aligned} &First\ canonical\ form: && \gamma = \ a^*b^*, \quad \boldsymbol{\pi} = \left[\tfrac{1-b^*}{1-a^*b^*} \quad \tfrac{b^*-a^*b^*}{1-a^*b^*} \right], \\ &Second\ canonical\ form: && \gamma = -a^*b^*, \quad \boldsymbol{\pi} = \left[\tfrac{b^*}{1+a^*b^*} \quad 1-\tfrac{b^*}{1+a^*b^*} \right]. \end{aligned}$$

For modeling correlated traffic we set directly the parameters λ_1, λ_2, a^* and b^* by utilizing the above properties of the canonical form. To set different loads (ρ) besides the same correlation parameter (γ), the actual matrix GF of the traffic model $MAP(2)$, $\widehat{\mathbf{D}}(z) = \mathbf{D}_0 + \mathbf{D}_1 z$, is scaled as

$$\widehat{\mathbf{D}}(z) = \frac{\rho}{\lambda_{base}\, b} \widehat{\mathbf{D}}_{base}(z), \text{ where } \lambda_{base} = \boldsymbol{\pi} \left. \frac{d}{dz} \widehat{\mathbf{D}}_{base}(z) \right|_{z=1} \mathbf{e}.$$

7.4. **Examples for performance evaluation.** In this section we provide numerical examples for the performance of the IEEE 802.16e sleep mode mechanism by the help of the presented vacation model. The IEEE 802.16 system operates with slotted-time frame with length of 5ms. In the numerical examples we use normalized time, in which the time unit equals the length of the frame.

The packet service time is constant with the length of 2 frames. The length of the listening interval is set to 1 frame. The setting for the power parameters $\mathcal{P}_a$ and $\mathcal{P}_s$ as well as for the energy parameter $\mathcal{E}_{on}$ are taken from [14]. Table 1 summarizes the evaluation parameters. The parameter values imply that a *sleep interval* must be longer than 1 frame length to achieve any power savings. Therefore the length of the *initial-sleep interval* is at least 2 frames in the evaluation examples.

We use the simplified notation $(2, 4, 8)$ for the sleep mode strategy, in which the sequence of *sleep intervals* is specified by their lengthes as denoted in the bracket, measured in frames, and the last *sleep interval* is repeated. Thus in the above example the length of the *initial-sleep interval* equals 2 frames, the length of the second *sleep interval* equals 4 frames and the length of the *final-sleep interval* equals 8 frames, which is then repeated.

Parameter	Value
Frame duration	5 ms
Packet service time (constant)	2
Length of the listening interval	1
Power during active periods ($\mathcal{P}_a$)	150 mW
Power during sleep intervals ($\mathcal{P}_s$)	10 mW
Extra energy for switchings ($\mathcal{E}_{on}$)	1 mJ

TABLE 1. Evaluation parameters

For $\gamma = \pm 0.45$ the parameters of $\widehat{\mathbf{D}}_{base}(z)$ are set as $\lambda_1 = 0.07$, $\lambda_2 = 1.47$, $a^* = 0.99$ and $b^* = \frac{5}{11}$. For the uncorrelated case ($\gamma = 0$) we use $a^* = 0$ and $b^* = 0.5$.

In figure 2 we have plotted the dependency of the mean packet delay and the standard deviation of packet delay on the load for different values of the correlation parameter (γ). For this figure we used the sleep strategy $(2, 4, 8, 16, 32)$.

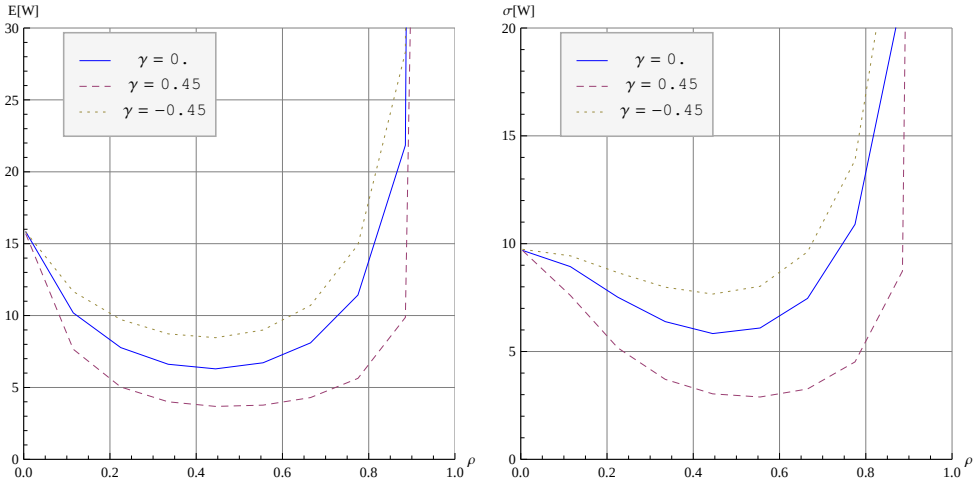

FIGURE 2. Mean and standard deviation of packet delay ($E[W]$, $\sigma[W]$) versus load (ρ) for different correlation parameters (γ).

The first observation, which is very interesting, is that the mean packet delay has a minimum and it increases as $\rho \to 0$. Similar tendency has been observed also in [15]. The reason is that in low load range as the load decreases the first arrival occurs with high probability later, i.e. in longer vacation period. Thus the observed mean packet delay is also longer. This observation is specific for the dependent multiple vacation model, in which the length of the consecutive vacation periods are non-decreasing and at least once the next vacation period is longer then the previous one. The same tendency can be observed also for the standard deviation of the packet delay.

It can be also seen on the figure that the presence of correlation has major influence on the values of both the mean and the standard deviation of the packet delay. Moreover the sign of the change in the delay value (increment or decrement) depends on the correlation parameter for both delay measures.

Figure 3 shows the dependency of the mean power savings on the load for different values of the correlation parameter (γ). For this figure we used again the sleep strategy $(2, 4, 8, 16, 32)$. It can be seen from the figure that the dependency on the load is close to linear and the presence of correlation has remarkable influence on the values of the mean power savings. Again the sign of the difference in the mean power savings values compared to the uncorrelated case depends on the value of the correlation parameter.

In the next we show an example for determining the optimal sleep mode parameters for a simple strategy. In case of the BE and nrtPS services (for which the power saving class of type I is recommended) there is no strict prescription for the packet delay. Therefore we apply a strategy, in which the mean power savings practically prioritized over the mean packet delay. This is done in two steps. In the first one we look for the best *final-sleep interval* while each of the sleep mode strategies starts with 2 frames length *initial-sleep interval*. Then in the second step we find the best *initial-sleep interval* while keeping the previously found best *final-sleep interval*.

In order to perform the first step of the selected optimization strategy, the mean power savings are plotted in figure 4 as a function of the load for different sleep mode

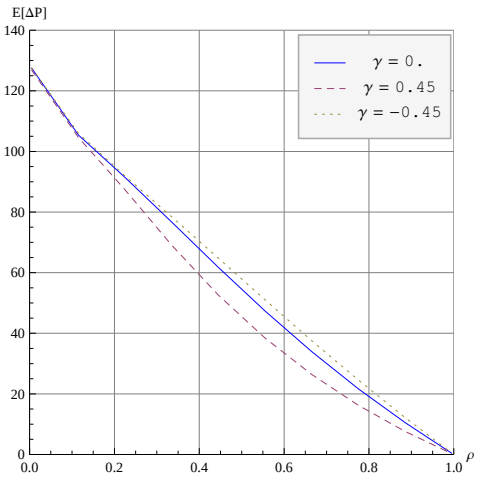

FIGURE 3. Mean power savings ($E[\Delta\mathcal{P}]$) versus load (ρ) for different correlation parameters (γ).

strategies for two different values of the correlation parameter. The length of the *final-sleep interval* is varied between $2 - 128$. In this case the practical prioritizing the mean power savings over the mean packet delay means to maximize the mean power savings on the cost of the mean packet delay as far as it results in significant plus in the mean power savings.

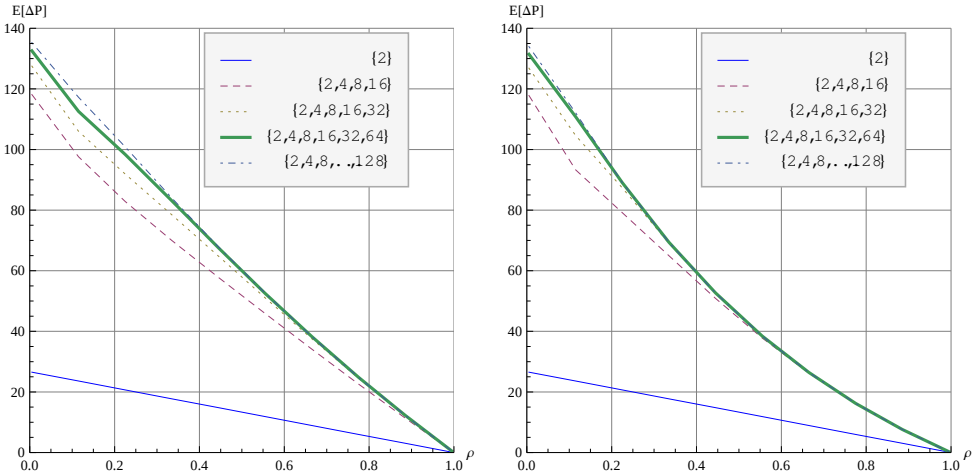

FIGURE 4. Mean power savings ($E[\Delta\mathcal{P}]$)versus load (ρ) for different sleep mode strategies for $\gamma = -0.45$ (left side) and $\gamma = 0.45$ (right side).

It can be observed from the figure that increasing the length of the *final-sleep interval* up to 32 results in essential plus in the mean power savings in case of both values of the correlation parameter. However further increase of the length of the *final-sleep interval* leads to different conclusions depending on the values of the correlation parameter. In the following we consider the practically more important

range of $\rho \geq 0.2$. For the case of $\gamma = -0.45$ the *final-sleep interval* with length of 64 yields to further considerable increment of mean power savings, while for the case of $\gamma = 0.45$ it has only marginal effect on it. Thus in case of $\gamma = -0.45$ the optimal length of the *final-sleep interval* is the one among 64 and 128 with the less mean packet delay. Similarly in the case of $\gamma = 0.45$ the optimal length of the *final-sleep interval* is the one among 32, 64 and 128 with the least mean packet delay. In the figure 5 the mean packet delays are plotted as a function of the load for for the same group of sleep mode strategies and for the same values of the correlation parameter as before.

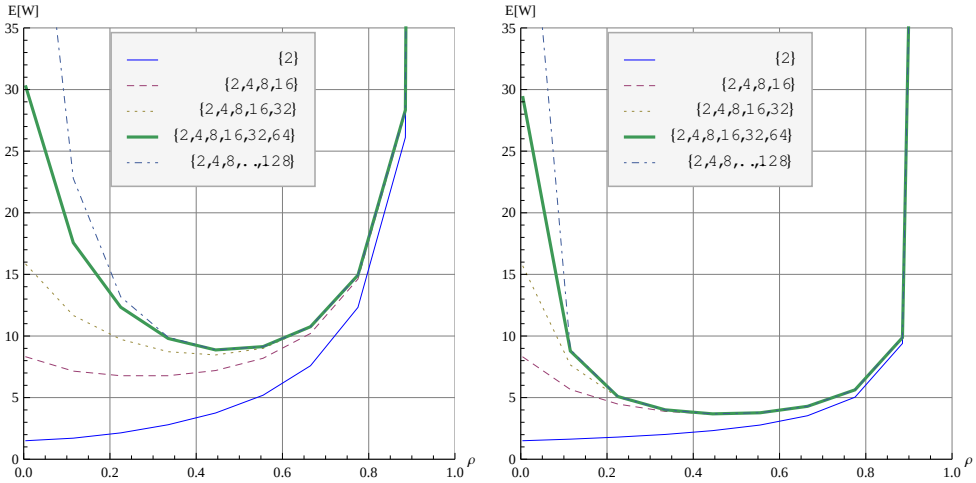

FIGURE 5. Mean packet delay ($E[W]$) versus load (ρ) for different sleep mode strategies for $\gamma = -0.45$ (left side) and $\gamma = 0.45$ (right side).

From the figure it can be seen that the optimal values of the *final-sleep interval* for the cases of $\gamma = -0.45$ and of $\gamma = 0.45$ are 64 and 32, respectively. Note that taking into account also the load range below 0.2 would results in different, but further on correlation parameter dependent optimal value for the length of the *final-sleep interval*.

Another conclusion which can be drown from this figure is that in spite of the fact that the mean packet delay values depend on the correlation parameter, the tendencies among the mean packet delay curves for the different sleep mode strategies are the same for both values of the correlation parameter. This can be explained as follows. Comparing any pair of sleep mode strategies with the doubling rule, the mean forward recurrence vacation time in the sleep mode strategy with longer *final-sleep interval* is higher than that one in the other sleep mode strategy. As far as the packet service time is small compared to the vacation periods, this effect, which does not depend on the correlation parameter, has a major impact on the tendencies between the mean packet delay curves.

Now we vary the length of *initial-sleep interval*, while keeping the length of *final-sleep interval* at the above optimal values. The figure (6) shows the mean power savings and the mean packet delay as a function of the load for sleep mode strategies with different length of *initial-sleep interval* for $\gamma = -0.45$ and $\gamma = 0.45$.

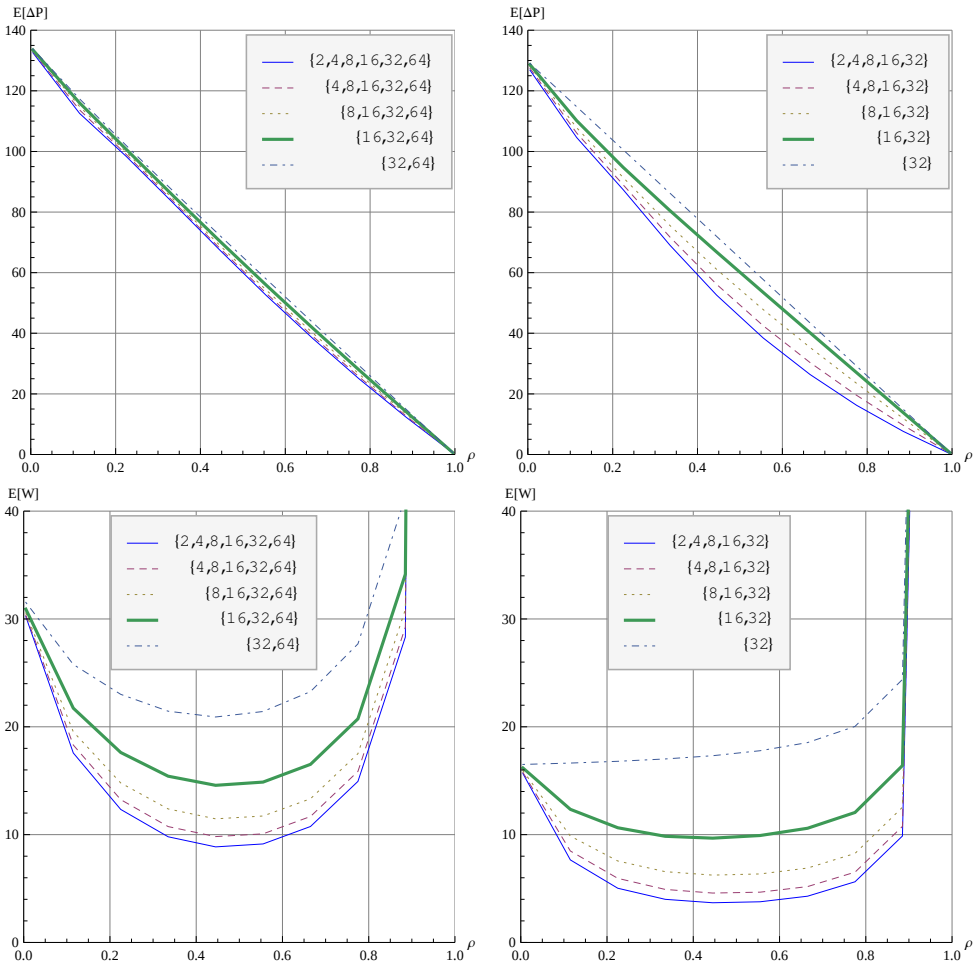

FIGURE 6. Mean power savings ($E[\Delta\mathcal{P}]$) and mean packet delay ($E[W]$) versus load (ρ) for sleep mode strategies with different length of *initial-sleep intervals* for $\gamma = -0.45$ (left side) and $\gamma = 0.45$ (right side).

The left side of the figure clearly shows that increasing the length of *initial-sleep interval* above 16 (besides the same length *final-sleep interval*) has no essential impact on the mean power savings. However it has significant influence on the mean packet delay. In this case the practical prioritizing the mean power savings over the mean packet delay means to minimize the mean packet delay among the cases with *initial-sleep interval* of 16 and 32. This results in the optimal sleep mode strategy for the case of $\gamma = -0.45$ as $(16, 32, 64)$.

In the case of $\gamma = 0.45$ the mean power savings also benefits from increasing the length of the *initial-sleep interval* from 16 to the highest possible value of 32. Therefore the sleep mode strategy (32) is considered as optimal for the case of $\gamma = 0.45$.

7.5. **Effect of different steps of *sleep intervals*.** Now we relax the doubling rule of the *sleep intervals* and study different sleep mode strategies with the optimal *initial- and final-sleep intervals* for the case of $\gamma = -0.45$. We can see from the left side of figure 6 that the mean power savings is relative insensitive to the sleep mode strategies having *final-sleep interval* with fixed length. This suggests that essential further plus in mean power savings can not be reached by applying another sequences of *sleep intervals* up to the *final-sleep interval*. Therefore we focus on minimizing the mean packet delay.

We utilize the rule that generally the sequence of shorter vacation periods results in lower mean packet delay, which can be also observed on the figure 5 and on the lower part of figure 6. Thus we gradually decrease the highest difference in the length of the consecutive *sleep intervals* from 32 up to 2. This results in the sleep mode strategies, for which the mean packet delay and the mean power savings as a function of the load are shown in figure 7.

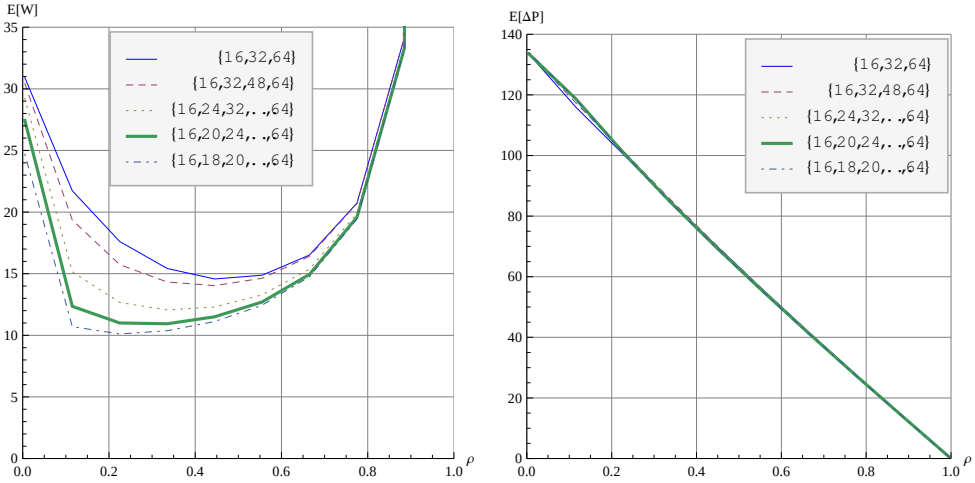

FIGURE 7. Mean packet delay ($E[W]$) and mean power savings ($E[\Delta\mathcal{P}]$) versus load (ρ) for sleep mode strategies with different steps of *sleep intervals* for $\gamma = -0.45$.

It can be seen from the figure that the mean power savings is practically the same for the considered sleep mode strategies as expected. Thus the optimal sleep mode strategy for the case of $\gamma = -0.45$ is the one with the minimal mean packet delay, i.e. $(16, 18, 20, ...64)$.

Finally in order to show the achieved performance of the out-of standard optimal sleep mode strategy $(16, 18, 20, ...64)$ for the case of $\gamma = -0.45$, in figure 8 we position its curves among the curves of several reference sleep mode strategies including the one with the lowest mean packet delay (2), the optimal one according to the standard $(16, 32, 64)$ and two other ones according to the standard with mean packet delay curves closest to the mean packet delay curve of $(16, 18, 20, ...64)$.

The figure shows that the out-of standard optimal sleep mode strategy $(16, 18, 20, ..., 64)$ results in better mean packet delay than the optimal sleep mode strategy $(16, 32, 64)$ at almost the same level of power savings.

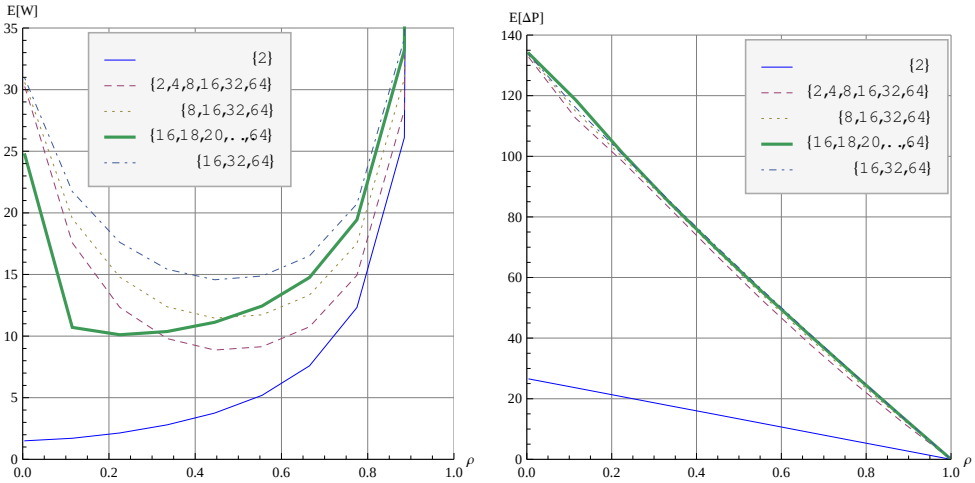

FIGURE 8. Mean packet delay ($E[W]$) and mean power savings ($E[\Delta\mathcal{P}]$) versus load (ρ) for the selected and reference sleep mode strategies for $\gamma = -0.45$.

7.6. **Enforcing an upper bound on mean delay.** The analytic model of the power saving class of type I applied in previous subsections enables also to determine the optimal sleep mode parameters while satisfying an upper bound on mean delay. Usually this can be given for the practically important range of load, which can be e.g. $0.2 \leq \rho \leq 0.7$. In a first step every sets of sleep mode parameters are determined, which satisfy the specified mean delay bound in the given load range. Afterwards in a second step the optimal set of sleep mode parameters is selected by maximizing the mean power savings over the previously determined parameter sets.

7.7. **Cost model.** In case of more general QoS requirement on delay constraint an appropriate cost model can be built up to determine the optimal sleep mode strategy. We developed a steady-state average cost function $\mathcal{F}(\varsigma)$, where the sleep mode strategy ς is the decision variable. The parameters of the cost function are defined as

$$c_1 \equiv \text{Cost of the mean packet delay,}$$
$$c_2 \equiv \text{Reward of the mean power savings.}$$

Then the optimal sleep mode strategy can be obtained by minimizing the total average system cost, which is given as

$$\mathcal{F}(\varsigma) = c_1 E[W] + \frac{c_2}{E[\Delta\mathcal{P}]}. \tag{50}$$

The minimum can be numerically determined as a function of the load and the correlation parameter by applying the expressions of the mean packet delay (47) and the mean power savings (48).

8. **Conclusion.** The considered BMAP queue with dependent multiple vacation can be also applied to model and analyze other sleep mode mechanisms in wireless systems having similar multiple vacation model.

The canonical form of $MAP(2)$ is a flexible traffic model of correlated arrival processes which is applicable to investigate the effect of the traffic parameters such as the correlation parameter in the performance evaluation of systems modeled by $BMAP$ queuing models.

Based on the numerical examples the following conclusions can be drawn for IEEE 802.16e sleep mode mechanism with power saving class of type I for BE and nrtPS services:

- The presence of correlation in the downlink traffic has considerable influence both on the packet delay and on the mean power savings. Depending on the correlation parameter both the packet delay and the mean power savings can be changed in both directions, i.e. they can be increased or decreased.
- The optimal sleep mode strategy depends on the correlation parameter (at least for the applied settings).
- The tendencies of the mean packet delay in the dependency on the applied sleep mode strategies do not show any dependency on the correlation parameter (at least for the applied settings).
- In the considered example, applying the power savings maximization strategy, in which the mean power savings practically prioritized over mean packet delay, the optimal sleep mode strategies are found as $(16, 32, 64)$ for the case of $\gamma = -0.45$ and (32) for the case of $\gamma = 0.45$.

Moreover we found that in the considered example for the case of $\gamma = -0.45$ the out-of standard sleep mode strategy $(16, 18, 20, ..., 64)$ outperforms the optimal sleep mode strategy $(16, 32, 64)$ in terms of mean packet delay, while their power savings do not differ considerable.

The presented analytic model of the power saving class of type I also enables to enforce a specified upper bound on the mean delay in a given range of load. In this case the optimal sleep mode parameters can be determined by maximizing the mean power savings over the relevant sets of parameters as described in subsection 7.6.

In case of more general QoS requirement on delay constraint, the optimal IEEE 802.16e sleep mode strategy can be determined by minimizing the cost function (50).

In order to simulate the performance of the IEEE 802.16 network numerous correlated traffic models have been elaborated for various data, voice and video traffic types [17], [23]. The data traffic models in [17] are based on the superposition of Interrupted Poisson Processes (IPP), which are special cases of MAP. Therefore they can be directly modeled by BMAP. The other traffic models can be modeled by BMAPs approximately. In theses cases appropriate fitting procedures can be applied to determine the suitable BMAPs. Hence applying BMAP in the considered queueing model opens the way for applying traffic models in the analysis of the considered sleep mode mechanism, which is left for future research.

Acknowledgement. This work is supported by the NAPA-WINE FP7-ICT (http://www.napa-wine.eu) and the OTKA K61709 projects. The authors would like to thank the anonymous reviewers whose valuable comments helped to significantly improve this manuscript.

Appendix A. **Proof of theorem 5.2.** The vector $\widetilde{\mathbf{r}}(s)$ is defined as

$$\widetilde{\mathbf{r}}(s) = \widetilde{\mathbf{w}}(s)\left(\widehat{\mathbf{D}}(\widetilde{B}(s)) + s\mathbf{I}\right). \tag{51}$$

Additionally we introduce the notation $\mathbf{r}^{(k)} = (-1)^k \frac{d^k}{ds^k}\widetilde{\mathbf{r}}(s)\,|_{s=0}$, for $k \geq 1$.
To prove the theorem we need the following lemma.

Lemma A.1. *The terms* $\mathbf{w}^{(1)}$ *and* $\mathbf{w}^{(2)}$ *can be expressed from (51) in terms of* $\mathbf{r}^{(1)}$, $\mathbf{r}^{(2)}\mathbf{e}$, $\mathbf{r}^{(2)}$ *and* $\mathbf{r}^{(3)}\mathbf{e}$ *as follows:*

$$\mathbf{w}^{(1)} = -\frac{\mathbf{r}^{(2)}\mathbf{e}\boldsymbol{\pi}}{2(1-\rho)} + \mathbf{r}^{(1)}\mathbf{C_3} + \boldsymbol{\pi}\mathbf{C_4}, \tag{52}$$

$$\begin{aligned}\mathbf{w}^{(2)} &= -\frac{\mathbf{r}^{(3)}\mathbf{e}\boldsymbol{\pi}}{3(1-\rho)} + \mathbf{r}^{(2)}\left(\mathbf{C_3} - \frac{\mathbf{e}\boldsymbol{\pi}}{1-\rho}\mathbf{C_4}\right) + 2\mathbf{r}^{(1)}\mathbf{C_3}\mathbf{C_4} \\ &+ \boldsymbol{\pi}\left(2\mathbf{C_4}\mathbf{C_4} - (b^2\mathbf{D}^{(2)} + b^{(2)}\mathbf{D}^{(1)})\mathbf{C_3}\right) \\ &+ \boldsymbol{\pi}\frac{\left(b^3\mathbf{D}^{(3)} + 3bb^{(2)}\mathbf{D}^{(2)} + b^{(3)}\mathbf{D}^{(1)}\right)\mathbf{e}\boldsymbol{\pi}}{3(1-\rho)}.\end{aligned} \tag{53}$$

Proof. Since $\left(\widehat{\mathbf{D}}(\widetilde{B}(s)) + s\mathbf{I}\right)\Big|_{s=0} = \mathbf{D}$ in (30) is singular we apply the method used by Lucantoni in [3] and Neuts in [22], which utilizes that $(\mathbf{D} + \mathbf{e}\boldsymbol{\pi})$ is nonsingular. The first three derivatives of $\left(\widehat{\mathbf{D}}(\widetilde{B}(s)) + s\mathbf{I}\right)$ at $s = 0$ can be expressed as:

$$\begin{aligned}\left.\frac{d\left(\widehat{\mathbf{D}}(\widetilde{B}(s)) + s\mathbf{I}\right)}{ds}\right|_{s=0} &= \mathbf{I} - b\mathbf{D}^{(1)}, \\ \left.\frac{d^2\left(\widehat{\mathbf{D}}(\widetilde{B}(s)) + s\mathbf{I}\right)}{ds^2}\right|_{s=0} &= b^2\mathbf{D}^{(2)} + b^{(2)}\mathbf{D}^{(1)}, \\ \left.\frac{d^3\left(\widehat{\mathbf{D}}(\widetilde{B}(s)) + s\mathbf{I}\right)}{ds^3}\right|_{s=0} &= -\left(b^3\mathbf{D}^{(3)} + 3bb^{(2)}\mathbf{D}^{(2)} + b^{(3)}\mathbf{D}^{(1)}\right).\end{aligned} \tag{54}$$

Taking the first three derivatives of (51) at $s = 0$, applying the expressions (54) and rearranging yields

$$\mathbf{w}^{(1)}\mathbf{D} = \mathbf{r}^{(1)} + \boldsymbol{\pi}(\mathbf{I} - b\mathbf{D}^{(1)}), \tag{55}$$

$$\mathbf{w}^{(2)}\mathbf{D} = \mathbf{r}^{(2)} + 2\mathbf{w}^{(1)}(\mathbf{I} - b\mathbf{D}^{(1)}) - \boldsymbol{\pi}\left(b^2\mathbf{D}^{(2)} + b^{(2)}\mathbf{D}^{(1)}\right). \tag{56}$$

$$\begin{aligned}\mathbf{w}^{(3)}\mathbf{D} &= \mathbf{r}^{(3)} + 3\mathbf{w}^{(2)}(\mathbf{I} - b\mathbf{D}^{(1)}) - 3\mathbf{w}^{(1)}\left(b^2\mathbf{D}^{(2)} + b^{(2)}\mathbf{D}^{(1)}\right) \\ &- \boldsymbol{\pi}\left(b^3\mathbf{D}^{(3)} + 3bb^{(2)}\mathbf{D}^{(2)} + b^{(3)}\mathbf{D}^{(1)}\right).\end{aligned} \tag{57}$$

Adding $\mathbf{w}^{(1)}\mathbf{e}\boldsymbol{\pi}$ to both sides of (55) and using $\boldsymbol{\pi}(\mathbf{D} + \mathbf{e}\boldsymbol{\pi})^{-1} = \boldsymbol{\pi}$ leads to

$$\mathbf{w}^{(1)} = \left(\mathbf{w}^{(1)}\mathbf{e}\right)\boldsymbol{\pi} + \left(\mathbf{r}^{(1)} + \boldsymbol{\pi}(\mathbf{I} - b\mathbf{D}^{(1)})\right)(\mathbf{D} + \mathbf{e}\boldsymbol{\pi})^{-1}. \tag{58}$$

The next step is to get the unknown term $\left(\mathbf{w}^{(1)}\mathbf{e}\right)$ in (58). Post-multiplying (56) by $\mathbf{e}$ and post-multiplying (58) by $\left(\mathbf{I} - b\mathbf{D}^{(1)}\right)\mathbf{e}$ and rearranging gives

$$\begin{aligned}
\mathbf{w}^{(1)}\left(\mathbf{I} - b\mathbf{D}^{(1)}\right)\mathbf{e} &= -\frac{1}{2}\mathbf{r}^{(2)}\mathbf{e} + \frac{1}{2}\boldsymbol{\pi}\left(b^2\mathbf{D}^{(2)} + b^{(2)}\mathbf{D}^{(1)}\right)\mathbf{e}, &(59)\\
\mathbf{w}^{(1)}\left(\mathbf{I} - b\mathbf{D}^{(1)}\right)\mathbf{e} &= \left(\mathbf{w}^{(1)}\mathbf{e}\right)\boldsymbol{\pi}(\mathbf{I} - b\mathbf{D}^{(1)})\mathbf{e} &(60)\\
&+ \left(\mathbf{r}^{(1)} + \boldsymbol{\pi}(\mathbf{I} - b\mathbf{D}^{(1)})\right)(\mathbf{D} + \mathbf{e}\boldsymbol{\pi})^{-1}(\mathbf{I} - b\mathbf{D}^{(1)})\mathbf{e},
\end{aligned}$$

respectively. Combining (59) and (60) and applying $\boldsymbol{\pi}\left(\mathbf{I} - b\mathbf{D}^{(1)}\right)\mathbf{e} = 1 - \rho$ results in the expression of the required term:

$$\begin{aligned}
\mathbf{w}^{(1)}\mathbf{e} &= \frac{1}{2(1-\rho)}\left(-\mathbf{r}^{(2)}\mathbf{e} + \boldsymbol{\pi}(b^2\mathbf{D}^{(2)} + b^{(2)}\mathbf{D}^{(1)})\mathbf{e}\right) &(61)\\
&- \frac{1}{(1-\rho)}\left(\mathbf{r}^{(1)} + \boldsymbol{\pi}(\mathbf{I} - b\mathbf{D}^{(1)})\right)(\mathbf{D} + \mathbf{e}\boldsymbol{\pi})^{-1}(\mathbf{I} - b\mathbf{D}^{(1)})\mathbf{e}.
\end{aligned}$$

Substituting (61) into (58) leads to

$$\begin{aligned}
\mathbf{w}^{(1)} &= -\frac{\mathbf{r}^{(2)}\mathbf{e}\boldsymbol{\pi}}{2(1-\rho)} + \mathbf{r}^{(1)}(\mathbf{D} + \mathbf{e}\boldsymbol{\pi})^{-1}\left(\mathbf{I} - \frac{(\mathbf{I} - b\mathbf{D}^{(1)})\mathbf{e}\boldsymbol{\pi}}{1-\rho}\right) &(62)\\
&+ \boldsymbol{\pi}\left(\frac{(b^2\mathbf{D}^{(2)} + b^{(2)}\mathbf{D}^{(1)})\mathbf{e}\boldsymbol{\pi}}{2(1-\rho)}\right)\\
&+ \boldsymbol{\pi}(\mathbf{I} - b\mathbf{D}^{(1)})(\mathbf{D} + \mathbf{e}\boldsymbol{\pi})^{-1}\left(\mathbf{I} - \frac{(\mathbf{I} - b\mathbf{D}^{(1)})\mathbf{e}\boldsymbol{\pi}}{1-\rho}\right).
\end{aligned}$$

Substituting matrices $\mathbf{C_3}$ and $\mathbf{C_4}$ into (62) results in the first statement.

Now we add $\mathbf{w}^{(2)}\mathbf{e}\boldsymbol{\pi}$ to both sides of (56). Using $\boldsymbol{\pi}(\mathbf{D} + \mathbf{e}\boldsymbol{\pi})^{-1} = \boldsymbol{\pi}$ leads to

$$\mathbf{w}^{(2)} = \left(\mathbf{w}^{(2)}\mathbf{e}\right)\boldsymbol{\pi} + \left(\mathbf{r}^{(2)} + 2\mathbf{w}^{(1)}(\mathbf{I} - b\mathbf{D}^{(1)}) - \boldsymbol{\pi}(b^2\mathbf{D}^{(2)} + b^{(2)}\mathbf{D}^{(1)})\right)(\mathbf{D} + \mathbf{e}\boldsymbol{\pi})^{-1}. \tag{63}$$

The next step is again to determine the unknown term $\left(\mathbf{w}^{(2)}\mathbf{e}\right)$ in (63). Post-multiplying (57) by $\mathbf{e}$ and post-multiplying (63) by $\left(\mathbf{I} - b\mathbf{D}^{(1)}\right)\mathbf{e}$ and rearranging gives

$$\begin{aligned}
\mathbf{w}^{(2)}\left(\mathbf{I} - b\mathbf{D}^{(1)}\right)\mathbf{e} &= -\frac{1}{3}\mathbf{r}^{(3)}\mathbf{e} + \mathbf{w}^{(1)}\left(b^2\mathbf{D}^{(2)}\mathbf{e} + b^{(2)}\mathbf{D}^{(1)}\right)\mathbf{e} &(64)\\
&+ \frac{1}{3}\boldsymbol{\pi}\left(b^3\mathbf{D}^{(3)} + 3bb^{(2)}\mathbf{D}^{(2)} + b^{(3)}\mathbf{D}^{(1)}\right)\mathbf{e},\\
\mathbf{w}^{(2)}\left(\mathbf{I} - b\mathbf{D}^{(1)}\right)\mathbf{e} &= \left(\mathbf{w}^{(2)}\mathbf{e}\right)\boldsymbol{\pi}(\mathbf{I} - b\mathbf{D}^{(1)})\mathbf{e} &(65)\\
&+ \left(\mathbf{r}^{(2)} + 2\mathbf{w}^{(1)}(\mathbf{I} - b\mathbf{D}^{(1)})\right)(\mathbf{D} + \mathbf{e}\boldsymbol{\pi})^{-1}(\mathbf{I} - b\mathbf{D}^{(1)})\mathbf{e},\\
&- \left(\boldsymbol{\pi}(b^2\mathbf{D}^{(2)} + b^{(2)}\mathbf{D}^{(1)})\right)(\mathbf{D} + \mathbf{e}\boldsymbol{\pi})^{-1}(\mathbf{I} - b\mathbf{D}^{(1)})\mathbf{e}.
\end{aligned}$$

respectively. Combining (64) and (65) and applying $\boldsymbol{\pi}\left(\mathbf{I}-b\mathbf{D}^{(1)}\right)\mathbf{e}=1-\rho$ results in the expression of the required term:

$$\begin{aligned}\mathbf{w}^{(2)}\mathbf{e} &= -\frac{1}{3\left(1-\rho\right)}\left(\mathbf{r}^{(3)}\mathbf{e}-\boldsymbol{\pi}(b^3\mathbf{D}^{(3)}+3bb^{(2)}\mathbf{D}^{(2)}+b^{(3)}\mathbf{D}^{(1)})\mathbf{e}\right) \\ &- \frac{1}{(1-\rho)}\left(\mathbf{r}^{(2)}+2\mathbf{w}^{(1)}(\mathbf{I}-b\mathbf{D}^{(1)})\right)(\mathbf{D}+\mathbf{e}\boldsymbol{\pi})^{-1}(\mathbf{I}-b\mathbf{D}^{(1)})\mathbf{e} \\ &+ \frac{1}{(1-\rho)}\mathbf{w}^{(1)}\left(b^2\mathbf{D}^{(2)}\mathbf{e}+b^{(2)}\mathbf{D}^{(1)}\right)\mathbf{e}, \\ &+ \frac{1}{(1-\rho)}\left(\boldsymbol{\pi}(b^2\mathbf{D}^{(2)}+b^{(2)}\mathbf{D}^{(1)})\right)(\mathbf{D}+\mathbf{e}\boldsymbol{\pi})^{-1}(\mathbf{I}-b\mathbf{D}^{(1)})\mathbf{e}.\end{aligned} \tag{66}$$

Applying (66) in (63) leads to:

$$\begin{aligned}&\mathbf{w}^{(2)}=-\frac{\mathbf{r}^{(3)}\mathbf{e}\boldsymbol{\pi}}{3\left(1-\rho\right)}+\mathbf{r}^{(2)}\left(\mathbf{D}+\mathbf{e}\boldsymbol{\pi}\right)^{-1}\left(\mathbf{I}-\frac{(\mathbf{I}-b\mathbf{D}^{(1)})\mathbf{e}\boldsymbol{\pi}}{1-\rho}\right) \\ &+2\mathbf{w}^{(1)}\left(\left(\frac{(b^2\mathbf{D}^{(2)}+b^{(2)}\mathbf{D}^{(1)})\mathbf{e}\boldsymbol{\pi}}{2\left(1-\rho\right)}\right)+\left(\mathbf{I}-b\mathbf{D}^{(1)}\right)\left(\mathbf{D}+\mathbf{e}\boldsymbol{\pi}\right)^{-1}\left(\mathbf{I}-\frac{(\mathbf{I}-b\mathbf{D}^{(1)})\mathbf{e}\boldsymbol{\pi}}{1-\rho}\right)\right) \\ &-\boldsymbol{\pi}(b^2\mathbf{D}^{(2)}+b^{(2)}\mathbf{D}^{(1)})\left(\mathbf{D}+\mathbf{e}\boldsymbol{\pi}\right)^{-1}\left(\mathbf{I}-\frac{(\mathbf{I}-b\mathbf{D}^{(1)})\mathbf{e}\boldsymbol{\pi}}{1-\rho}\right) \\ &+\boldsymbol{\pi}\frac{\left(b^3\mathbf{D}^{(3)}+3bb^{(2)}\mathbf{D}^{(2)}+b^{(3)}\mathbf{D}^{(1)}\right)\mathbf{e}\boldsymbol{\pi}}{3\left(1-\rho\right)}.\end{aligned} \tag{67}$$

Substituting matrices $\mathbf{C_3}$ and $\mathbf{C_4}$ into (67) leads to

$$\begin{aligned}\mathbf{w}^{(2)} &= -\frac{\mathbf{r}^{(3)}\mathbf{e}\boldsymbol{\pi}}{3\left(1-\rho\right)}+\mathbf{r}^{(2)}\mathbf{C_3}+2\mathbf{w}^{(1)}\mathbf{C_4} \\ &- \boldsymbol{\pi}(b^2\mathbf{D}^{(2)}+b^{(2)}\mathbf{D}^{(1)})\mathbf{C_3} \\ &+ \boldsymbol{\pi}\frac{\left(b^3\mathbf{D}^{(3)}+3bb^{(2)}\mathbf{D}^{(2)}+b^{(3)}\mathbf{D}^{(1)}\right)\mathbf{e}\boldsymbol{\pi}}{3\left(1-\rho\right)}.\end{aligned} \tag{68}$$

Applying (52) in (68) gives the second statement.

□

Substituting (30) into (51) yields

$$\widetilde{\mathbf{r}}\left(s\right)=\left(1-\rho\right)\frac{\mathbf{m}}{v}\left(\mathbf{I}-\widetilde{\boldsymbol{\mathcal{V}}}\left(s\right)\right). \tag{69}$$

Taking the first three derivatives of $\widetilde{\mathbf{r}}\left(s\right)$ at $s=0$ gives

$\mathbf{r}^{(k)}$ for $k=1,2,3$ can be computed by means of taking the first three derivatives of $\widetilde{\mathbf{r}}\left(s\right)$ at $s=0$, which results in

$$\mathbf{r}^{(1)} = -\left(1-\rho\right)\frac{\mathbf{m}}{v}\mathcal{V}^{(1)}, \tag{70}$$

$$\mathbf{r}^{(2)} = -\left(1-\rho\right)\frac{\mathbf{m}}{v}\mathcal{V}^{(2)}, \tag{71}$$

$$\mathbf{r}^{(3)} = -\left(1-\rho\right)\frac{\mathbf{m}}{v}\mathcal{V}^{(3)}. \tag{72}$$

Substituting (70), (71), (72) into (52) and (53) gives the theorem.

REFERENCES

[1] H. Takagi, "Queueing Analysis - A Foundation of Performance Evaluation, Vacation and Prority Systems, vol.1," North-Holland, New York, 1991.

[2] B. T. Doshi. *Queueing systems with vacations - a survey*, Queueing Systems, **1** (1986), 29–66.

[3] D. L. Lucantoni. *New results on the single server queue with a batch markovian arrival process*, Stochastic Models, **7** (1991), 1–46.

[4] M. F. Neuts. "Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach," The John Hopkins University Press, Baltimore, 1981.

[5] D. L. Lucantoni, *The BMAP/G/1 queue: A tutorial*, in "Models and Techniques for Performance Evaluation of Computer and Communications Systems" (eds. L. Donatiello and R. Nelson), Springer Verlag, (1993), 330-358.

[6] S. H. Chang and T. Takine. *Factorization and stochastic decomposition properties in bulk queues with generalized vacations*, Queueing Systems, **50** (2005), 165–183.

[7] Zs. Saffer and M. Telek. *Analysis of BMAP/G/1 vacation model of non-M/G/1-type*, in "Computer Performance Engineering - EPEW 2008. LNCS, Vol. 5261" (eds. N. Thomas and C. Juiz), Springer Verlag, (2008), 212–226.

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

[9] Y. Xiao. *Energy saving mechanism in the IEEE 802.16e wireless MAN*, IEEE Communications Letters, **9/7** (2005), 595-597.

[10] K. Han and S. Choi. *Performance analysis of sleep mode operation in IEEE 802.16e mobile broadband wireless access systems*, In Proceedings of the IEEE 63rd Vehicular Technology Conference, VTC2006-Spring (Melbourne), **3** (2006), 1141-1145.

[11] J.-B. Seo, S.-Q. Lee, N.-H. Park, H.-W. Lee and C.-H. Cho. *Performance analysis of sleep mode operation in IEEE 802.16e*, In Proceedings of the 60th Vehicular Technology Conference, VTC2004-Fall (Los Angeles), **2** (2004), 1169-1173.

[12] Z. Huo, W. Yue, N. Tian and S. Jin. *Performance evaluation for the sleep mode in the 802.16e based on a queueing model with close-down time and multiple vacations*, Journal of Industrial and Management Optimization, **5/3** (2009), 511–524.

[13] E. Hwang, Y. H. Lee, K. J. Kim, J. J. Son, and B. D. Choi. *Performance analysis of power saving mechanism employing both sleep mode and idle mode in IEEE 802.16e*, IEICE Transactions on Communications, **E92-B/9** (2009), 2809-2822.

[14] E. Hwang, K. J. Kim, J. J. Son, and B. D. Choi. *The power saving mechanism with binary exponential traffic indications in the IEEE 802.16e/m*, Queueing Systems, **62** (2009), 197-227.

[15] K. D. Turck, S. D. Vuyst, D. Fiems, and S. Wittevrongel. *Performance analysis of the IEEE 802.16e sleep mode for correlated downlink traffic*, Telecommunication systems, **39** (2008), 145–156.

[16] L. Bodrog, A. Heindl, G. Horvath and M. Telek. *A Markovian canonical form of second-order matrix-exponential processes*, European Journal of Operational Research, **190(2)** (2008), 459–477.

[17] IEEE 802.16.3c-01/30r1 " Traffic Model for 802.16 TG3 MAC/PHY Simulations," March 2001.

[18] T. Takine and Y. Takahashi. *On the relationship between queue lengths at a random instant and at a departure in the stationary queue with bmap arrivals*, Stochastic Models, **14** (1998), 601–610.

[19] S. H. Chang, T. Takine, K. C. Chae, and H. W. Lee. *A unified queue length formula for BMAP/G/1 queue with generalized vacations*, Stochastic Models, **18** (2002), 369–386.

[20] IEEE 802.16e-2005, Part 16: " Air Interface for Fixed and Mobile Broadband Wireless Access Systems - Amendment for Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands, - Corrigendum 1," February 2006.

[21] IEEE 802.16-2004, Part 16: " Air Interface for Fixed Broadband Wireless Access Systems, Standard for Local and Metropolitan Area Networks," October 2004.

[22] M. F. Neuts. "Structured Stochastic Matrices of $M/G/1$ type and their Applications," Marcel Dekker, New York, 1989.

[23] WiMAX Forum " WiMAX System Evaluation Methodology V2.0," December 2007.

Received xxxx 20xx; revised xxxx 20xx.

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