On an exact solution of the rate matrix of G / M / 1 −type Markov process with small number of phases

Article *in* Journal of Parallel and Distributed Computing · May 2018

DOI: 10.1016/j.jpdc.2018.04.013

CITATIONS
8

READS
143

2 authors:

Rama Murthy Garimella
Mahindra Ecole Centrale, Hyderabad, India
189 PUBLICATIONS **810** CITATIONS

SEE PROFILE

Alexander Rumyantsev
Karelian Research Centre of the Russian Academy of Sciences
64 PUBLICATIONS **218** CITATIONS

SEE PROFILE

On an exact solution of the rate matrix of $G/M/1$-type Markov process with small number of phases

Rama Murthy Garimella

International Institute of Information Technology, Hyderabad, India

Rumyantsev Alexander*

Institute of Applied Mathematical Research of the Karelian Research Centre RAS; Petrozavodsk State University, Petrozavodsk, Russia

Abstract

In this research paper we consider the matrix polynomial equation arising naturally in the equilibrium analysis of a structured $G/M/1$-type Markov process. We obtain an explicit expression for the unknown rate matrix R being 2×2 matrix. The method is based on symbolic solution of the determinental polynomial equation. Using Cayley–Hamilton theorem, the matrix polynomial equation for the matrix R is reduced to the system of linear equations. Motivated by applications in Edge Computing by means of Internet of Things devices having tight constraints in energy consumption, we demonstrate the applicability of the method by a novel approach to energy efficiency of a single-server computing system. A new randomized regime switching scheme is proposed, which, as it is shown by means of numeri-

*Corresponding author

Email addresses: rammurthy@iiit.ac.in (Rama Murthy Garimella), ar0@krc.karelia.ru (Rumyantsev Alexander)

cal experiment, provides significant decrease of energy consumption of the system under study.

Keywords: $G/M/1$-type Markov Process, Matrix Polynomial Equation, Matrix-Analytic Method, Explicit Solution, Energy Efficiency

2000 MSC: 60K25, 60J28

---

1. Introduction

Matrix analytic method continues to prove its applicability to a wide variety of models in the field of modern computing and communication systems. Extensively developed by M. Neuts in the celebrated work [1], the powerful method allows to study sophisticated objects: stability [2, 3] and performance [4] of stochastic model of a high-performance cluster, power management in datacenters [5], redundancy in high-performance systems [6], cloud computing [7], to name a few.

The system under study is usually modelled as a continuous time Markov process $\{(X(t), Y(t)), t \geqslant 0\}$ with countable state space $E := \{(0, j), j = 1, \ldots, m_0; (i, j), i \geqslant 1, j = 1, \ldots, m\}$, where the *phase* variable $Y(t)$ may take one of m (or m_0 for boundary states) values and *level* variable $X(t)$ may be increased/decreased at each transition. The state space E can be partitioned into *levels* with level $n \geqslant 1$ being the subset $\{(n, j), j = 1, \ldots, m\} \subset E$. In many fields of interest, it is assumed that the level is increased by at most one, and decreased by at most $N - 1$ (we focus on the case $N < \infty$) units at each transition epoch. These models belong to the so-called structured $G/M/1$-type Markov processes, extensively studied in [1], with the natural example of such a process being the queue length process, embedded at arrival epochs.

The infinitesimal generator matrix of a structured $G/M/1$-type process has the following block-multidiagonal representation

$$Q = \begin{pmatrix} A^{0,0} & A^{0,1} & 0 & 0 & \dots \\ A^{1,0} & A^{1,1} & A^{(0)} & 0 & \dots \\ A^{2,0} & A^{(2)} & A^{(1)} & A^{(0)} & \dots \\ A^{3,0} & A^{(3)} & A^{(2)} & A^{(1)} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \\ \mathbf{0} & A^{(N+1)} & A^{(N)} & A^{(N-1)} & \dots \\ \mathbf{0} & \mathbf{0} & A^{(N+1)} & A^{(N)} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{1}$$

where $A^{(i)}, i = 0, \dots, N+1$ are square matrices of order m, satisfying the balance equation

$$A\mathbf{1} = \mathbf{0}, \quad \text{where } A := \sum_{i=0}^{N+1} A^{(i)}, \tag{2}$$

$\mathbf{1}$ ($\mathbf{0}$) is the vector of ones (zeroes) of corresponding dimension, $A^{0,0}$ is a square matrix of order m_0 and $A^{i,0}, A^{0,1}$ are possibly rectangular matrices. (Recall that for these type of processes the off-diagonal elements of matrix Q, i.e. the rates of transitions of the process, are nonnegative.)

The key component of the method is to obtain the steady-state probability vector $\pi = (\pi_{i,j}), i, j \in E$ of the system states in the level-wise matrix-geometric form [1] (for more details on the method see [8, 9])

$$\pi_k = \pi_{k-1} R, \quad k \geqslant 1, \tag{3}$$

where $\pi_k = (\pi_{k,1}, \dots, \pi_{k,m})$ and the matrix R is the minimal nonnegative

solution of a matrix polynomial equation

$$P(R) := \sum_{i=0}^{N} R^i A^{(i)} = \mathbf{0}. \tag{4}$$

Thus, the analysis is essentially reduced to obtaining the matrix R. However, in general, the rate matrix R is obtained by means of converging iterative procedures [1, 10, 11] (see also the comparison of iterative procedures [12]), and, to the best of our knowledge, the explicit solution in general is available only for a number of special cases (in particular, when matrix $A^{(0)}$ is a rank-one matrix [13], or $A^{(2)}$ is a rank-one matrix in a system with $N = 2$ [10, 14], see also [15]).

To avoid the numerical procedure of obtaining the matrix R, some alternative methods are developed. The spectral decomposition-based methods were suggested in [16, 17] (which required eigenanalysis of a matrix polynomial), some of them utilizing special structure of the model to decrease the computational complexity [18]. In [19], the Spectral Analysis method based on Jordan canonical form is suggested to obtain a closed form analytic solution of (4). In [20] an extensive study of the Jordan form representation of R (i.e computing eigenvalues and generalized eigenvectors of R) is performed, and the finite memory recursions are suggested to decrease memory requirements of the algorythm. A more general discussion of Jordan canonical form method is provided in [21]. Alternatively, new method of steady-state analysis and obtaining the closed form solution by finding the roots of some (complex-variable) polynomial is developed in [7].

A particular case of $G/M/1$ process with $N = 2$ is the so-called Quasi-Birth-Death (QBD) process, when the level variable is allowed to increase

or decrease by at most one at a time. In this case the matrix polynomial equation (4) is reduced to the matrix quadratic equation

$$R^2 A^{(2)} + R A^{(1)} + A^{(0)} = 0, \tag{5}$$

first used to find a solution of a QBD process (by means of Complex Analysis-based method) in late 60's [22, 23].

Recently a method of obtaining explicit solution for the rate matrix R of a QBD process with a small number of phases was proposed [24], where, in particular, the 2×2 matrices $A^{(0)}, A^{(2)}$ may be of full rank (we also note a closely related work [25], where both 2×2 matrices were full rank, however, upper-triangular). In the present paper we extend the results of [24] to a more general structured $G/M/1$-type Markov process with $N > 1$ and two phases at a level.

As an example application of the proposed solution, we elaborate on the new method of randomized switching for power saving (first proposed in [24]) which we feel may be implemented in Edge Computing systems based on Internet of Things devices. By numerical experiments, we investigate the optimal configuration of the switching parameters that minimize the average energy consumption and guarantee the required quality of service. We note that the proposed randomized switching approach may be applied to systems, where cost effectiveness and service elasticity is important, such as high-performance and cloud-based computing systems [26], as well as teletraffic systems (e.g. on-demand content servers), where the operational cost (e.g. energy cost, or cloud service cost), as well as the system speed, is to be adopted to the working conditions. The approach is suitable for heavy load conditions, since the switching is performed autonomously, without any

knowledge on the system state, except the type of the current event.

This research paper is organized as follows. First, we briefly describe the results of [24]. Then we present an algebraic approach on obtaining the matrix R. Next, we apply this approach to solve the optimization problem related to Energy-Performance tradeoff, and illustrate the approach with simulation results.

2. Explicit Rate Matrix of a QBD Process with 2 Phases at a Level

Below we briefly recall the results introduced in [24], where a QBD process (with $N = 2$) was considered. In this case, the rate matrix R is a solution of matrix quadratic equation (5).

We assume that the Neuts ergodicity condition holds [9],

$$\rho := \alpha A^{(0)}\mathbf{1}/\alpha A^{(2)}\mathbf{1} < 1, \tag{6}$$

where the stochastic vector α is the solution of the following system

$$\begin{cases} \alpha A & = 0 \\ \alpha \mathbf{1} & = 1. \end{cases} \tag{7}$$

First, the following *factorization lemma* (proven by direct expansion of the r.h.s. of (8) and substitution of (5)) holds true [27, 20]:

Lemma 1. *The polynomial matrix* $A(\xi) := A^{(0)} + \xi A^{(1)} + \xi^2 A^{(2)}$ *allows the following factorization:*

$$A(\xi) \equiv (\xi I - R)(\xi A^{(2)} + RA^{(2)} + A^{(1)}). \tag{8}$$

Then, the *determinantal polynomial* $\det A(\xi)$ is constructed, and it is easy to see that the polynomial is of degree four. Note also that $\xi = 1$ is always root of $\det A(\xi)$ by the balance condition (2). Rewrite

$$\det A(\xi)/(\xi - 1) = a_3\xi^3 + a_2\xi^2 + a_1\xi + a_0. \tag{9}$$

Denote by $\xi_i, i = 1, 2, 3$ the zeroes of the polynomial (9). It follows from Lemma 1 that $\det A(\xi) = \det(\xi I - R)\det(\xi A^{(2)} + RA^{(2)} + A^{(1)})$. Thus, the eigenvalues of R are zeroes of the determinantal polynomial $\det A(\xi)$. Note that there are exactly $m = 2$ zeroes (denote them ξ_1, ξ_2) of the determinantal polynomial $\det A(\xi)$ which are strictly inside unit circle and these are the eigenvalues of R (provided $\eta := \mathrm{sp}(R) < 1$), see e.g. [20]. Hence the zeroes of $\det(\xi A^{(2)} + RA^{(2)} + A^{(1)})$ are outside the unit circle.

Consider now the *characteristic polynomial* of the matrix R in a monic form

$$\det(\xi I - R) = (\xi - \xi_1)(\xi - \xi_2) = \xi^2 + b_1\xi + b_0, \tag{10}$$

where $b_0 = \xi_1\xi_2 = \det R$, $b_1 = -(\xi_1 + \xi_2) = -\mathrm{Trace}(R)$ are (real) scalars, since R is a nonnegative matrix. By Perron–Frobenius theorem, $\eta = \mathrm{sp}(R)$ is simple, real eigenvalue, and $\eta \in (0, 1)$, provided (6) holds. W.o.l.o.g. let $\xi_1 = \eta$. Then ξ_2 is also real, which makes ξ_3 also real. Given that, it is easy to obtain

$$b_1 = \frac{a_2}{a_3} + \xi_3, \tag{11}$$

$$b_0 = -\frac{a_0}{a_3\xi_3}. \tag{12}$$

The value ξ_3 (which is the real zero outside the unit disk) is obtained following the celebrated trigonometric solution of the cubic equation (9) by the

following substitutions:

$$p = \frac{3a_3a_1 - a_2^2}{9a_3^2},\ q = \frac{2a_2^3 - 9a_3a_2a_1 + 27a_3^2a_4}{27a_3^3}. \tag{13}$$

Then [28]

$$\xi_3 = -\frac{a_2}{3a_3} + 2\sqrt{-p}\cos\left(\frac{1}{3}\cos^{-1}\left(\frac{q}{2p\sqrt{-p}}\right)\right). \tag{14}$$

By Cayley-Hamilton theorem, we obtain

$$R^2 = -b_1R - b_0I, \tag{15}$$

which, by substitution into (5), leads to the following system of linear equations:

$$R\left[A^{(1)} - b_1A^{(2)}\right] - b_0A^{(2)} + A^{(0)} = \mathbf{0}. \tag{16}$$

Thus, if $A^{(1)} - b_1A^{(2)}$ is invertible, then

$$R = \left[b_0A^{(2)} - A^{(0)}\right]\left[A^{(1)} - b_1A^{(2)}\right]^{-1}. \tag{17}$$

It remains to note that the invertibility of $A^{(1)} - b_1A^{(2)}$ was proven by diagonal dominance.

We finalize the section with a procedure to obtain the rate matrix R in an explicit form.

Step 1. Obtain the maximal eigenvalue ξ_3 by (14).

Step 2. Obtain b_1 by (11) and b_0 by (12).

Step 3. Obtain R by (17).

It remains to note that the initial vectors π_0, π_1 are obtained by the following linear system of equations

$$(\pi_0, \pi_1)\begin{pmatrix} A^{0,0} & A^{0,1} \\ A^{1,0} & A^{1,1} + RA^{(2)} \end{pmatrix} = 0, \tag{18}$$

$$\pi_0\mathbf{1} + \pi_1(I - R)^{-1}\mathbf{1} = 1. \tag{19}$$

3. Explicit Rate Matrix of a $G/M/1$-type Process with 2 Phases at a Level

We are interested in matrix polynomial equation (4), provided the stability condition holds [1]

$$\alpha A^{(0)}\mathbf{1} < \alpha \sum_{k=2}^{N}(k-1)A^{(k)}\mathbf{1}, \tag{20}$$

where, recall, the stochastic vector α is the solution of the system (7). We define the corresponding generator function

$$G(\xi) = \sum_{i=0}^{N} \xi^i A^{(i)},$$

and it is easy to show that

$$G(\xi) = (\xi I - R)J(\xi, R), \tag{21}$$

where

$$J(\xi, R) := \sum_{i=0}^{N-1} \xi^i E_i, \tag{22}$$

and

$$E_i = \sum_{j=i+1}^{N} R^{j-i-1} A^{(j)}.$$

The derivation of (21) follows the Residual Theorem [20].

Following Cayley–Hamilton theorem, if the matrix R is 2×2 matrix, then

$$R^2 = (\eta + \mu)R - \eta\mu I, \tag{23}$$

where $\eta > \mu$ are the distinct real eigenvalues of R (recall also that $\eta = \mathrm{sp}(R)$). Thus, if we define the (column) vector

$$\begin{pmatrix} a_1 \\ b_1 \end{pmatrix} := \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

then from (23) by induction it is easy to obtain

$$R^j = a_j R + b_j I, \tag{24}$$

where the values a_j, b_j are defined recursively by the following linear transform

$$\begin{pmatrix} a_j \\ b_j \end{pmatrix} = \begin{pmatrix} \eta+\mu & 1 \\ -\eta\mu & 0 \end{pmatrix} \begin{pmatrix} a_{j-1} \\ b_{j-1} \end{pmatrix}, \quad j \geqslant 2. \tag{25}$$

Thus, returning back to the polynomial equation, we obtain that $P(R)$ is equivalent to a linear polynomial

$$P(R) = R\sum_{j=1}^{N} a_j A^{(j)} + \sum_{j=2}^{n} b_j A^{(j)} + A^{(0)}. \tag{26}$$

It follows that the matrix R may be obtained as a solution of the linear equation $P(R) = \mathbf{0}$. However, to obtain the solution, we may need to define the coefficients explicitly. This is done in the following Theorem.

Theorem 1. *The coefficients in* (26) *are as follows*

$$a_j = \sum_{i=0}^{j-1} \eta^i \mu^{j-i-1}, \tag{27}$$

$$b_j = -\sum_{i=1}^{j-1} \eta^i \mu^{j-i}. \tag{28}$$

Proof: Indeed, for $j = 1$ the equalities (27), (28) hold (with convention $b_1 = 0$), and it is straightforward to check from (25) that they hold for $j = 2$, with $a_2 = \eta + \mu, b_2 = -\eta\mu$. Now we consider that (27) holds for $j \leqslant k-1$ and prove the induction step. Indeed,

$$\sum_{i=0}^{k-1} \eta^i \mu^{k-i-1} = \mu \sum_{i=0}^{k-2} \eta^i \mu^{k-i-2} + \eta^{k-1} = \mu a_{k-1} + \eta^{k-1}. \tag{29}$$

However, by (25),

$$a_k = (\eta+\mu)a_{k-1} + b_{k-1} = (\eta+\mu)a_{k-1} - \eta\mu a_{k-2} = \mu a_{k-1} + \eta(a_{k-1} - \mu a_{k-2}).$$

Thus, showing that the l.h.s. of (29) equals a_k is equivalent to showing that η^{k-2} equals $a_{k-1} - \mu a_{k-2}$. Indeed, by induction,

$$\begin{aligned} a_{k-1} - \mu a_{k-2} &= \sum_{i=0}^{k-2} \eta^i \mu^{k-i-2} - \mu \sum_{i=0}^{k-3} \eta^i \mu^{k-i-3} = \\ &= \sum_{i=0}^{k-2} \eta^i \mu^{k-i-2} - \sum_{i=0}^{k-3} \eta^i \mu^{k-i-2} = \eta^{k-2}. \end{aligned}$$

The result (28) follows directly from (25), since

$$b_j = -\eta\mu a_{j-1}.$$

This completes the proof. ■

Remark 1. *Note that* (27) *originates in the following proposition:*

$$J(\eta, \mu I) = \sum_{j=1}^{N} a_j A^{(j)}.$$

Indeed, it follows from (22) *that*

$$J(\eta, \mu I) = \sum_{i=1}^{N-1} \eta^i \sum_{j=1}^{N} \mu^{j-i-1} A^{(j)} = \sum_{j=1}^{N} A^{(j)} \sum_{i=0}^{j-1} \eta^i \mu^{j-i-1}, \tag{30}$$

where the last equation follows by changing the order of sums.

We note that the equations (10)–(14) hold true, since the localization of the eigenvalues η, μ inside the unit disk, and ξ_3 outside the unit disk is valid,

as proven in [21]. Thus, we only need to change the equation (17) w.r.t. (26) and (30) as follows

$$R = -\left(\sum_{j=2}^{n} b_j A^{(j)} + A^{(0)}\right) [J(\eta, \mu I)]^{-1}. \tag{31}$$

It remains to show that $J(\eta, \mu I)$ is invertible, which we do by the following lemma, generalizing the corresponding result in [24].

Lemma 2. *The matrix $J(\eta, \mu I)$ is invertible.*

Proof: It was shown in [21] that $J(\xi, R)$ is nonsingular and diagonally dominant for arbitrary (complex) $\xi : |\xi| < 1$. In particular, it follows that $J(\eta, R)$ is a diagonally dominant matrix. Note that by the properties of the generator matrix Q, only the diagonal elements of matrix $A^{(1)}$ are nonpositive, while other matrices are nonnegative. It follows from (22) that

$$J(\eta, R) = A^{(1)} + \sum_{j=2}^{N} R^{j-1} A^{(j)} + \sum_{i=1}^{N-1} \eta^i E_i,$$

where $E_i, i \geqslant 1$ do not contain the matrix $A^{(1)}$. Then the diagonal dominance guarantees that

$$|A^{(1)}_{i,i}| > \left| \left[\sum_{j=2}^{N} R^{j-1} A^{(j)} + \sum_{i=1}^{N-1} \eta^i E_i \right] + \sum_{j=1, j\neq i}^{m} A^{(1)}_{i,j} \right|, \tag{32}$$

Since the eigenvalues are distinct, then R is diagonalizable, hence R has a spectral representation [21]

$$R = \eta U_1 + \mu U_2, \tag{33}$$

where U_1, U_2 are the so-called residue matrices, with $U_1 + U_2 = I$. Note that by Perron–Frobenius theorem, $U_1 \geqslant \mathbf{0}$ componentwise [10], which provides a

componentwise inequality

$$R \geqslant \mu U_1 + \mu U_2 = \mu I. \tag{34}$$

Note that since the matrices $A^{(j)}, j \geqslant 2$ as well as $E_i, i \geqslant 1$ are nonnegative, then replacement of the matrix R by μI in (32) keeps the inequality. Thus, the matrix $J(\eta, \mu I)$ is strictly diagonally dominant matrix and hence is nonsingular. ■

Thus, the corresponding algorithm to obtain the explicit solution for the rate matrix R of a $G/M/1$-type Markov process is as follows.

Step 1. Obtain the maximal eigenvalue ξ_3 by (14).

Step 2. Obtain b_1 by (11) and b_0 by (12).

Step 3. Obtain R by (31).

It remains to note that the initial vectors π_0, π_1 are obtained by the following linear system of equations following from [13] and (24)

$$(\pi_0, \pi_1) \begin{pmatrix} A^{0,0} & A^{0,1} \\ \sum_{j=1}^{N}(a_{j-1}R + b_{j-1}I)A^{j,0} & A^{1,1} + \sum_{j=1}^{N}(a_jR + b_jI)A^{(j+1)} \end{pmatrix} = 0,$$

with normalizing condition

$$\pi_0 \mathbf{1} + \pi_1 (I - R)^{-1} \mathbf{1} = 1.$$

4. Random Switching for Energy Efficiency

In this section we elaborate more on the novel method of energy efficiency improvement by randomized switching of low and high power consuming regimes, presented in [24]. This method is a randomized extension of the well-known power throttling method used in high-performance and distributed

computing [29, 30]. We briefly recall the system under study. Consider a queueing system with a renewal input flow of customers arriving into an (unbounded) First-Come-First-Served queue. The i.i.d. interarrival times are exponentially distributed with rate $\lambda > 0$. Each customer requires an exponentially distributed (with unit rate) amount of work to be done (say, information to be transmitted). The single server operates two speed modes (call them high and low), with rates $\mu_2 > \mu_1 > 0$. The server may switch the speed only at the arrival/departure epochs (asynchronously). Denote $c_0 < c_1 < c_2$ the energy consumption per unit time in idle (no customers in the system)/low/high modes. In order to preserve energy, the server implements the following random switching policy:

- at the task arrival epoch, given the current mode is low, switch to high mode with probability (w.p.) p_1, or remain low w.p. $1 - p_1$;
- at the task departure epoch, given the current mode is high, switch to low mode w.p. p_2, or remain high w.p. $1 - p_2$.

Let $\nu(t) \in \{0, 1 \dots\}$ be the number of customers, and $m(t) \in \{1, 2\}$ be the mode of the system at time $t \geqslant 0$. Then the following Markov process

$$\{(\nu(t), m(t)) \in \{0, 1, \dots\} \times \{1, 2\},\ t \geqslant 0\} \tag{35}$$

is a continuous-time QBD process, with $\nu(t)$ being the *level*, and $m(t)$ being the *phase* at time t.

The infinitesimal generator matrix of the process (35) has the form (1),

where we define the matrices explicitly:

$$A^{(0)} = \begin{pmatrix} (1-p_1)\lambda & p_1\lambda \\ 0 & \lambda \end{pmatrix}, \tag{36}$$

$$A^{(1)} = \begin{pmatrix} -\lambda-\mu_1 & 0 \\ 0 & -\lambda-\mu_2 \end{pmatrix}, \tag{37}$$

$$A^{(2)} = \begin{pmatrix} \mu_1 & 0 \\ p_2\mu_2 & (1-p_2)\mu_2 \end{pmatrix}, \tag{38}$$

$$A^{0,0} = -\lambda I, \quad A^{0,1} = A^{(0)}, \tag{39}$$

$$A^{1,1} = A^{(1)}, \quad A^{1,0} = A^{(2)}. \tag{40}$$

The stability criterion of the process (35), established in [24], is as follows

$$\lambda p_1(\lambda-\mu_2) + \mu_2 p_2(\lambda-\mu_1) < 0. \tag{41}$$

Intuitively, the condition (41) indicates a negative drift of the service process of the system under heavy load, with respect to the mode switching intensity.

Note that $p_1 = 1$ and $p_2 = 0$ corresponds to a classical $M/M/1$ service system working at the speed μ_2 (referred below as classical system). Let $\mathrm{E}\nu_0$ be the average number of customers, and $\mathrm{E}\mathcal{E}_0$ be the average energy consumption per unit time in the classical system in stationary regime. It can be readily seen that

$$\mathrm{E}\nu_0 = \frac{\rho}{1-\rho}, \tag{42}$$

$$\mathrm{E}\mathcal{E}_0 = c_0(1-\rho) + c_2\rho, \tag{43}$$

where, recall, $1-\rho$ is the stationary idle probability of the classical system (for details on the aforementioned classical results see [31]). We may consider $\mathrm{E}\nu_0$ as the QoS parameter of the classical system.

Now we turn to the original two-mode system, i.e. we consider the non-trivial case $p_1, p_2 > 0$. Provided (41) holds, we define the matrix R following the steps of the suggested procedure of exact computation, where the coefficients in the polynomial (9) are obtained as follows:

$$\begin{aligned} a_3 &= \mu_1\mu_2(1-p_2), \\ a_2 &= -\mu_1(\lambda+\mu_2) - \lambda\mu_2(1-p_2), \\ a_1 &= \lambda(\lambda+\mu_1) + \lambda\mu_2(1-p_1), \\ a_0 &= -\lambda^2(1-p_1). \end{aligned}$$

After obtaining R from the equation (17), the equations (3)–(19) provide the stationary system state probabilities. Straightforward manipulation leads to the following system for π_1:

$$\begin{cases} \pi_1 \left(\frac{1}{\lambda}A^{(2)} - R^{-1}\right) A^{(0)}\mathbf{1} & = 0, \\ \pi_1 \left(\frac{1}{\lambda}A^{(2)} + (I-R)^{-1}\right)\mathbf{1} & = 1. \end{cases} \tag{44}$$

Then the value π_0 is obtained as follows:

$$\pi_0 = \frac{1}{\lambda}\pi_1 A^{(2)}. \tag{45}$$

The obtained solution allows to evaluate the average number of customers in the system as the QoS measure:

$$\mathrm{E}\nu_{p_1,p_2} = \pi_1(I-R)^{-2}\mathbf{1}, \tag{46}$$

where by notation ν_{p_1,p_2} we stress the dependence on the mode switching probabilities. The average energy consumption may be obtained as follows:

$$\mathrm{E}\mathcal{E}_{p_1,p_2} = \pi_0 c_0 \mathbf{1} + \pi_1(I-R)^{-1}(c_1,c_2)^T, \tag{47}$$

where the transposed vector (c_1, c_2) is the column-vector of energy consumption in each mode. Then the following optimization problem can be solved: minimization of the average energy consumption in stationary regime, provided the controlled QoS decrease:

$$\begin{aligned} &\min_{p_1,p_2>0} \mathrm{E}\mathcal{E}_{p_1,p_2}, \\ &\text{s.t. } \mathrm{E}\nu_{p_1,p_2} \leqslant (1+\varepsilon)\mathrm{E}\nu_0, \end{aligned} \tag{48}$$

for some $\varepsilon > 0$.

By numerical experiment we study the dependence of optimal regime switching strategy (p_1, p_2) on the intensity of arrivals λ. We set

$$\begin{aligned} &\mu_1 = 1/2.8,\ \mu_2 = 1/1.3, \\ &c_0 = 2300,\ c_1 = 3200,\ c_2 = 5400, \end{aligned}$$

which correspond to the frequencies and (roughly measured) power consumption of a single-CPU laptop we used as a testbed. Then, we vary $\lambda = 0.3, 0.4, 0.5, 0.6$ such that at first the system is stable at lower frequency, but with increasing λ the low-frequency regime becomes unstable.

Following the procedure of obtaining the exact solution, we numerically solve the optimization problem (48) and obtain approximate optimal values p_1, p_2, as well as approximate optimal energy consumption. We vary the QoS degradation $1 + \varepsilon \in (1.1, 30)$ and plot the obtained values. We perform simulation with R language [32].

It is easy to see from Fig. 1 that, as expected, if the low regime is stable ($\lambda = 0.3$), the system tends to an M/M/1 system working at low speed, with the appropriate growth of QoS degradation. When the low regime becomes

unstable ($\lambda \geqslant 0.4$), the system is forced to switch to high regime so as to guarantee stability, however, it still is (starting from some tolerated QoS decrease) switching to the low regime at each departure ($p_2 \to 1$). With an increasing input rate, the switching to lower regime at departures becomes less probable ($p_2 < 1$). Finally, when the system in a high regime becomes heavily loaded ($\lambda = 0.6$), the optimal strategy is such that switching to low regime at departures becomes less probable, than switching to high regime at arrivals.

5. Conclusion

We have presented the algebraic approach for obtaining the rate matrix R of the structured $G/M/1$-type process explicitly, when the process has only two phases at each level. Note that, albeit being simplistic, such a process may be used to model quite a number of recent applications, such as the modern communication devices, Internet of Things devices, high-performance and cloud-based servers (where processors use the Dynamic Voltage and Frequency Scaling technology), telecommunications (with two types of service high and low priority customers [33]). In such fields of application, it is crucial to have an easy and asynchronous (to guarantee robustness at high loads) scheme for optimizing the cost (e.g. average energy consumption per unit time). We proposed such a randomized scheme which may be implemented e.g. in the Edge Computing systems based on Internet of Things to optimize the cost under controlled QoS degradation, when possible, and evaluated it in a numerical experiment. We leave the extended field test of the proposed scheme for future research.

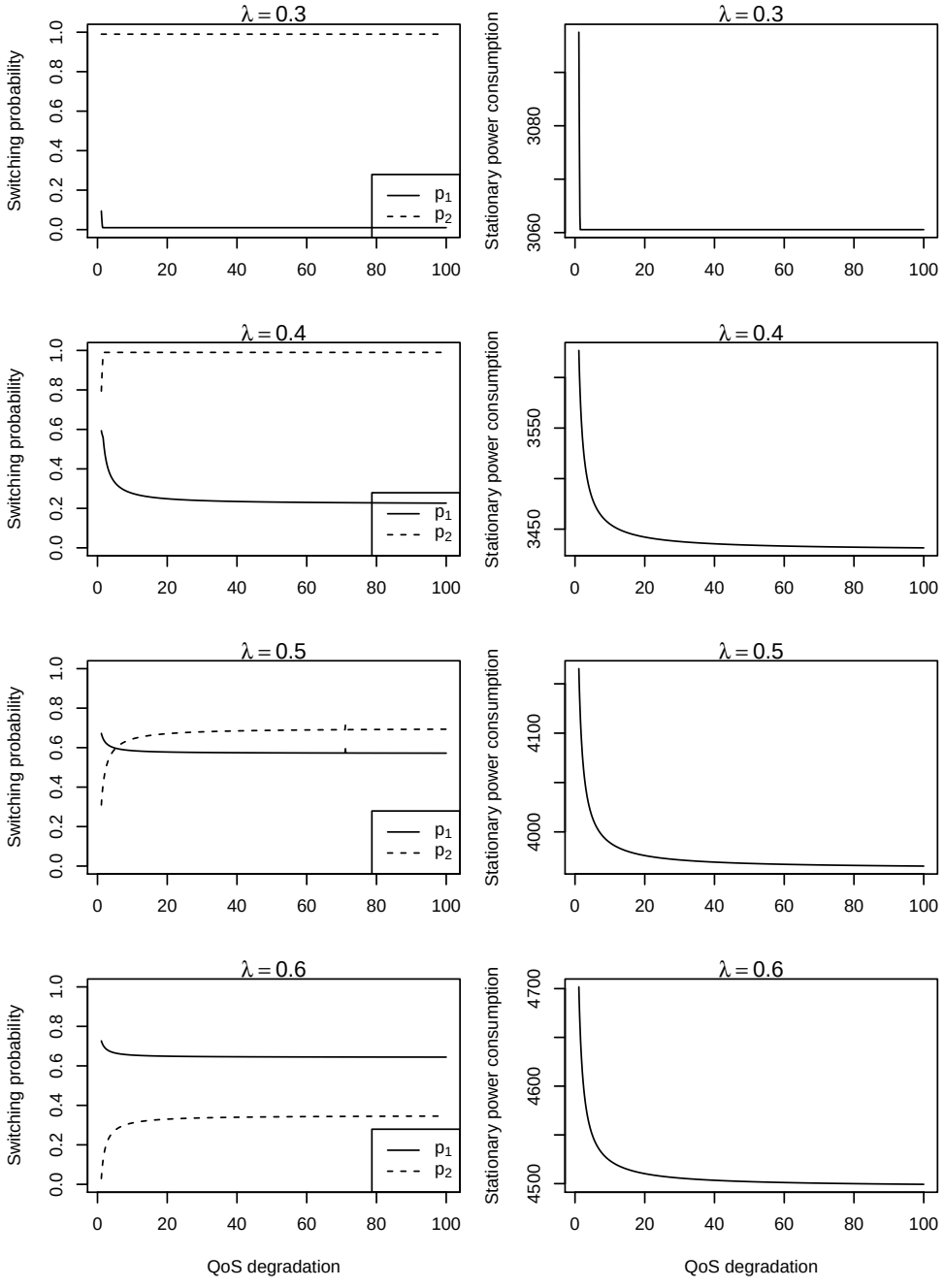

Figure 1: Switching probability vs. QoS degradation (left); average power consumption per unit time vs. QoS degradation (right) in a system for various input rate λ.

Note that the proposed randomized energy efficiency management scheme may be extended to multiserver models, which may be relevant to web servers working under heavy and/or bursty load. However, in this case analytical solution for the rate matrix R is not available, and numerical methods should be applied. When working with systems of large scale, the state space of the system dramatically increases, which motivates the need to apply High-Performance Computing for system performance evaluation.

6. Acknowledgements

The research of AR was carried out under state order to the Karelian Research Centre of the Russian Academy of Sciences (Institute of Applied Mathematical Research KRC RAS), and partially supported by RF President's grant MK-1641.2017.1 and RFBR, projects 16-07-00622, 18-07-00147, 18-07-00156.

[1] M. F. Neuts, Matrix-Geometric Solutions in Stochastic Models, Johns Hopkins University Press, Baltimore, 1981.

[2] A. Rumyantsev, E. Morozov, Stability criterion of a multiserver model with simultaneous service, Annals of Operations Research 252 (1) (2017) 29–39. doi:10.1007/s10479-015-1917-2.
URL https://doi.org/10.1007/s10479-015-1917-2

[3] E. Morozov, A. Rumyantsev, Stability Analysis of a MAP/M/s Cluster Model by Matrix-Analytic Method, in: D. Fiems, M. Paolieri, N. A. Platis (Eds.), Computer Performance Engineering: 13th European Workshop, EPEW 2016, Chios, Greece, October 5-7, 2016, Pro-

ceedings, Springer International Publishing, Cham, 2016, pp. 63–76. doi:10.1007/978-3-319-46433-6_5.
URL http://dx.doi.org/10.1007/978-3-319-46433-6_5

[4] S. R. Chakravarthy, H. D. Karatza, Two-server parallel system with pure space sharing and Markovian arrivals, Computers & Operations Research 40 (1) (2013) 510 – 519.

[5] S. Doroudi, B. Fralix, M. Harchol-Balter, Clearing analysis on phases: Exact limiting probabilities for skip-free, unidirectional, quasi-birth-death processes, Stoch. Syst. 6 (2) (2016) 420–458. doi:10.1214/15-SSY183.

[6] P. M. Fiorini, L. Lipsky, Exact analysis of some split-merge queues, ACM SIGMETRICS Performance Evaluation Review 43 (2) (2015) 51–53.
URL http://dl.acm.org/citation.cfm?id=2825257

[7] M. L. Chaudhry, A. D. Banik, A. Pacheco, S. Ghosh, A simple analysis of system characteristics in the batch service queue with infinite-buffer and Markovian service process using the roots method: $GI \ / \ C$ - $MS \ P^{\ (\ a,b\ )} \ / \ 1 \ / \ \infty$, RAIRO - Operations Research 50 (3) (2016) 519–551. doi:10.1051/ro/2015035.
URL http://www.rairo-ro.org/10.1051/ro/2015035

[8] M. Bladt, B. F. Nielsen, Matrix-Exponential Distributions in Applied Probability, Vol. 81 of Probability Theory and Stochastic Modelling,

Springer US, Boston, MA, 2017, dOI: 10.1007/978-1-4939-7049-0.
URL http://link.springer.com/10.1007/978-1-4939-7049-0

[9] Q.-M. He, Fundamentals of Matrix-Analytic Methods, Springer New York, 2014.

[10] G. Latouche, V. Ramaswami, Introduction to Matrix Analytic Methods in Stochastic Modeling, ASASIAM, Philadelphia, 1999.

[11] D. A. Bini, G. Latouche, B. Meini, Solving matrix polynomial equations arising in queueing problems, Linear Algebra and its Applications 340 (1) (2002) 225–244. doi:10.1016/S0024-3795(01)00426-8.

[12] T. T. Hung, T. V. Do, Computational aspects for steady state analysis of QBD processes, Periodica Polytech. Ser. Electr. Eng 44 (2) (2001) 179–200.

[13] D. Liu, Y. Q. Zhao, Determination of explicit solution for a general class of Markov processes, Matrix-Analytic Methods in Stochastic Models. Lecture Notes Pure Appl. Math. 183 (1996) 343–357.

[14] F. Gillent, G. Latouche, Semi-explicit solutions for M/PH/1-like queuing systems, European Journal of Operational Research 13 (2) (1983) 151–160. doi:10.1016/0377-2217(83)90077-2.

[15] J. van Leeuwaarden, E. Winands, Quasi-birth-and-death processes with an explicit rate matrix, Stochastic Models 22 (1) (2006) 77–98. doi:10.1080/15326340500481747.

[16] J. N. Daigle, D. M. Lucantoni, Queueing systems having phase-dependent arrival and service rates, Numerical Solution of Markov Chains (1991) 161–202.

[17] I. Mitrani, R. Chakka, Spectral expansion solution for a class of Markov models: Application and comparison with the matrix-geometric method, Performance Evaluation 23 (3) (1995) 241–260.

[18] T. V. Do, R. Chakka, An efficient method to compute the rate matrix for retrial queues with large number of servers, Applied Mathematics Letters 23 (5) (2010) 638–643. doi:10.1016/j.aml.2010.02.003.

[19] H. R. Gail, S. L. Hantler, B. A. Taylor, Spectral Analysis of M/G/1 and G/M/1 Type Markov Chains, Advances in Applied Probability 28 (1) (1996) 114. doi:10.2307/1427915.

[20] G. R. Murthy, Transient and equilibrium analysis of computer networks: Finite memory and matrix geometric recursions, Ph.D. thesis, Purdue University, West Lafayette (1989).

[21] G. R. Murthy, M. Kim, E. J. Coyle, Equilibrium analysis of skip free markov chains: Nonlinear matrix equations, Communications in Statistics: Stochastic Models 7 (1991) 547–571.

[22] V. L. Wallace, The Solution of Quasi Birth and Death Processes Arising from Multiple Access Computer Systems, Ph.D. thesis, University of Michigan (1969).

[23] R. V. Evans, Geometric Distribution in Some Two-Dimensional

Queuing Systems, Operations Research 15 (5) (1967) 830–846. doi:10.1287/opre.15.5.830.

[24] G. R. Murthy, A. Rumyantsev, On an Exact Solution of the Rate Matrix of Quasi-Birth-Death Process With Small Number of Phases, in: Proceedings: 31st European Conference on Modelling and Simulation ECMS 2017, May 23rd - May 26th, 2017, Budapest, Hungary, 2017, pp. 713–719. doi:10.7148/2017-0713.
URL https://doi.org/10.7148/2017-0713

[25] A. Krishnamoorthy, C. Sreenivasan, An M/M/2 Queueing System with Heterogeneous Servers Including One with Working Vacation, International Journal of Stochastic Analysis 2012 (2012) 1–16. doi:10.1155/2012/145867.
URL http://www.hindawi.com/journals/ijsa/2012/145867/

[26] D. Mukherjee, S. Dhara, S. Borst, J. S. H. van Leeuwaarden, Optimal Service Elasticity in Large-Scale Distributed Systems, ArXiv e-prints 1703.08373.

[27] F. Gantmacher, Theory of matrices, AMS Chelsea publishing, 1959.

[28] D. Zwillinger, CRC Standard Mathematical Tables and Formulae, 31st Edition, CRC, Boca Raton, 2003.

[29] A. Gandhi, M. Harchol-Balter, R. Das, J. O. Kephart, C. Lefurgy, Power capping via forced idleness, in: Proceedings of Workshop on Energy Efficient Design, 2009, pp. 1–6.
URL http://repository.cmu.edu/compsci/868/

[30] P. Hanappe, Fine-grained CPU Throttling to Reduce the Energy Footprint of Volunteer Computing, Tech. rep., Sony Computer Science Laboratory Paris (2012).
URL http://low-energy-boinc.cslparis.fr/info/images/f/fd/Hanappe-12a.pdf

[31] S. Asmussen, Applied probability and queues, Springer, New York, 2003.

[32] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2018).
URL https://www.R-project.org

[33] R. Razumchik, M. Telek, Delay analysis of a queue with re-sequencing buffer and markov environment, Queueing Systems 82 (1) (2016) 7–28. doi:10.1007/s11134-015-9444-z.
URL https://doi.org/10.1007/s11134-015-9444-z