Regenerative Estimation of a Simultaneous Service Multiserver System with Speed Scaling

Ruslana Nekrasova, Alexander Rumyantsev
Institite of Applied Mathematical Research KarRC RAS, Petrozavodsk State University
Petrozavodsk, Russia
ruslana.nekrasova@mail.ru, ar0@krc.karelia.ru

***Abstract*—The paper deals with simultaneous service multiserver system, where customer occupies a random number of servers. The system also admits speed scaling mechanism by switching speed regimes at arrival/departure instants according to the corresponding transition probability matrices. Stability conditions for Markovian case of such a model were derived. Results of regenerative confidence estimation for the system in steady-state regime are presented.**

I. Introduction

Modern computing systems are essentially parallel. Ranging from smaller autonomous sensors in the Internet of Things appliances [1], as well as wearable devices, up to large-scale High-Performance Computing systems, all of these possess multiple computational units (cores, processors, servers) scaling up to millions at the top. As usual, high computational power needs to be compromised with reasonable power consumption, which is done by leveraging various *energy efficiency* techniques. The latter require appropriate parameter tuning so as to obtain optimum in the energy-performance tradeoff.

Random nature of the events appearing in parallel computing systems makes multiserver queueing models best candidates for stochastic modeling of the former. Though being relatively simple, these models may have unwanted features that complicate the analysis, e.g. the so-called non-work-conserving property, when idling system resources cannot be used for serving customers awaiting in the queue, and thus, are wasted. These features raise the need to use sophisticated mathematical tools to establish the model stability and performance.

Regenerative method is a powerful instrument in stochastic simulation and performance analysis. It allows to obtain confidence estimates for system performance parameters even in complicated cases, where traditional estimates (e.g. simple average) are not applicable (in particular, when independence assumption does not hold). Detailed description of regenerative estimation is well presented in [2], [3]. Moreover, regenerative approach allows to establish stochastic stability of the system under weak assumptions (e.g. return of the process to some state/set with positive probability) [4], [5], [6].

In this paper, we apply the regenerative approach to multiserver queueing model with a distinctive feature which we call simultaneous service. In such a system, a pool of homogeneous servers are serving customers, each of the latter occupying a random number of servers simultaneously, for a (same at each occupied server) random amount of time. The customers are waiting in a single unbounded queue in the order of arrival, and enter service only when the requested number of servers are available. The simultaneous service feature reflects not only the key property of a High-Performance Cluster (where computing tasks are executed on many cores/servers to gain more computing power), but also addresses the mechanism used in conventional multicore systems to scale an application to the available number cores (e.g. in video processing). To balance the performance with energy consumption, the model we consider utilizes the speed scaling technique as the less intrusive one among available. In such a system, we establish confidence estimates for performance by means of simulation. This research is considered a first step towards large scale applications, and thus we start with a simple small scale model, which though allows to obtain some explicit analytical results.

The structure of the paper is as follows. First, in Section II we introduce the model of simultaneous service multiserver system and discuss its properties. We establish the stability condition in matrix form in Section III. We present the regenerative structure for the model and briefly discuss the regenerative method of confidence estimation in section IV. Some results of regenerative simulation are presented and discussed in Section V. Finally, conclusions are stated in Section VI.

II. Model Description

We consider a multi-server queueing system with m equivalent servers. Customers arrive into the system at the epochs $\{t_n;\, n = 0, 1, \dots\}$ of a Poisson process of rate λ, and thus inter-arrival times $\tau_n = t_n - t_{n-1}, n = 1, 2, \dots,$ are independent and exponentially distributed (i.i.d.) random variables (r.v.) with a generic element τ such that $\mathsf{E}\tau = 1/\lambda$. A customer n, arriving at instant t_n, is characterized by a pair of parameters: amount of work $S_n > 0$ and number of required servers $C_n \leq m$. The customer n demands exactly C_n servers simultaneously, and if the resources are not available, s/he waits until service is possible in a single queue operating on a First-Come-First-Served basis. The C_n servers requested by the customer n are seized and released simultaneously, after completion of the required amount of work, S_n, at each of the servers requested. Note that the considered system is *non-work-conserving*, as its properties allow to observe non-zero queue together with idle servers.

The sequences $\{S_n;\, n \geq 1\}$ is i.i.d. and independent of an i.i.d. sequence $\{C_n;\, n \geq 1\}$. We define the corresponding generic elements of such sequences by S and C. (Note that in a real-world system S_n and C_n might be dependent, e.g.

ISSN 2305-7254

in terms of distribution, see [7]). Generic number of required servers, C, is a discrete r.v. that takes values $k = 1, \dots, m$ with corresponding probability p_k from a given distribution $\mathbf{p} = (p_1, \dots, p_m)$. More detailed description and stability criterion of such a model are presented in [8].

The main feature of the system considered in this paper is the *speed scaling* technique, i.e. the servers can process customers at L distinct speeds, $\mu_1 < \dots < \mu_L$, and we assume that the speed switching simultaneously affects all m servers. (Such an assumption is motivated by simultaneous service feature, since a customer served at several servers with distinct speeds will experience the speed of the slowest server.) More precisely, if the servers work at rate μ_i, $i = 1, \dots, L$, then a generic work S can be completed in S/μ_i amount of time. It follows that the work S_n of customer n will require computing time in the interval $[S_n/\mu_L, S_n/\mu_1]$.

The motivation of speed scaling implementation in multiserver systems is mainly related to energy efficiency improvement under quality of service restrictions. Thus, inspired by high-performance and IoT applications, we study the model with asynchronous blind randomized switching policy, such that the (system) speed is altered only at customer arrival/departure epochs according to (corresponding) Markov chain. This means that at an arrival (departure) epoch, speed μ_i may be switched to speed μ_j with probabilities $a_{i,j}$ (or $d_{i,j}$, respectively), where $a_{i,j}$ ($d_{i,j}$) is a corresponding element of a square stochastic matrix A (D) of order L. Such a speed scaling approach was used in a single-server model to mimic the asynchronous switching at arrival/departure epoch in wireless transmission devices [9]. Note, that such a policy can be used to obtain the optimal performance/energy tradeoff without the need to keep track of the system state.

A. $M/G/2$-type System

We concentrate on a particular case of two-server system, that is, $m = 2$. Thus, customers can seize one or both servers according to a given distribution $\mathbf{p} = (p_1,\, 1 - p_1)$.

Note that the matrices A and D may define rather complicated speed scaling rules, e.g. of ladder type (when speeds are sequentially increased at arrivals and decreased at departures). However, to simplify comprehension, we focus on the two-speed case, that is, $L = 2$. Moreover, we adopt the transition matrices from [9]:

$$A = \begin{bmatrix} 1-a & a \\ 0 & 1 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 0 \\ d & 1-d \end{bmatrix}, \tag{1}$$

where $a,\, d \in [0,\, 1]$. As such, at arrival instants, the speed may be switched from lower to higher with probability a, while at departure instants it may be switched from higher to lower with probability d. Conventionally, the speed of an empty system at customer arrival epoch is chosen randomly.

The presented two-server system of $M/G/2$-type could be uniquely described by a set of parameters $\lambda, p_1, a, d, \mu_1, \mu_2$ and the distribution function of generic work amount, S.

III. Stability analysis

In this section we establish the stability condition for a $M/M/2$-type system, that is, the generic work amount S is exponentially distributed. Thus, we extend the matrix analytic model of a simultaneous service multiserver system [8] in a two-server case. Let $X(t)$ be the number of customers in the system at time $t \geqslant 0$ and two dimensional vector $N(t) = (n_1, n_2) \in \{1, 2\}^2$ define numbers of servers required by the two oldest customers in the system (if any). Finally, let $J(t) \in \{1, 2\}$ be the number of system speed at time t (that is, at time t system works at speed $\mu_{J(t)}$). The process $\{X(t), N(t), J(t), t \geqslant 0\}$ is the so-called Quasi-Birth-Death (QBD) process, that is, $X(t)$ is changed by at most 1, at the process transition epochs. The infinitesimal generator of a QBD process has block-tridiagonal form:

$$\begin{bmatrix} B^{0,0} & B^{0,1} & \mathbf{0} & \mathbf{0} & \dots \\ B^{1,0} & B^{1,1} & B_0 & \mathbf{0} & \dots \\ \mathbf{0} & B_2 & B_1 & B_0 & \dots \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{bmatrix}.$$

It remains to define the block matrices that constitute the infinitesimal generator. Since a change in the component $X(t)$ triggers a change in $N(t)$ and (independent) change in $J(t)$, the matrices governing these changes may be obtained by Kronecker product $\otimes$, of the corresponding matrices governing a reduced process $\{X(t), N(t), t \geqslant 0\}$ considered in [8], and matrices A, D governing the changes in $J(t)$. Namely, denote

$$P_N = \begin{bmatrix} p_1 & 1-p_1 & 0 & 0 \\ 0 & 0 & p_1 & 1-p_1 \\ p_1 & 1-p_1 & 0 & 0 \\ 0 & 0 & p_1 & 1-p_1 \end{bmatrix},$$

and it is easy to see that P_N is the matrix of transitions of phase $N(t)$ at departure epochs. Construct the following diagonal matrices $D_\sigma = \mathrm{diag}(2, 1, 1, 1)$, related to the number of customers simultaneously served at each phase, and $M = \mathrm{diag}(\mu_1, \mu_2)$. Now define

$$B_0 = I \otimes \lambda A,\ B_1 = -\lambda I - D_\sigma \otimes M,\ B_2 = (D_\sigma P_N) \otimes (MD).$$

We omit the matrices $B^{i,j}, i, j \in \{0, 1\}$ to save space, since their contents is not useful for stability analysis, which is based on the celebrated Neuts ergodicity criterion [10]

$$\alpha B_2 \mathbf{1} > \alpha B_0 \mathbf{1},$$

where $\mathbf{1}$ is the vector of ones, and α is the stochastic vector solving the equation

$$\alpha(B_0 + B_1 + B_2) = \mathbf{0}, \quad \alpha\mathbf{1} = 1. \tag{2}$$

Observe that (2) it is a linear system of equations that has unique solution which can be easily obtained. In that sense, α being the solution of a (well defined) linear system, is explicit.

We note that $\theta(a, d) = \alpha B_2 \mathbf{1}$ defines the *maximal throughput* of the system (by notation we stress the dependence of θ on scaling policy parameters a, d). To study the sensitivity of θ on the parameters a, d, we perform a numerical study. We vary $a_0 = 2^i, i = -10, \dots, 0$, and for each such a_0 plot the curve $(d, \theta(a_0, d))$ for $d \in [2^{-25}, 1]$. The results of numerical study are depicted on Fig. 1. It can be seen that the maximal throughput is bounded by the values of maximal throughput in a system without speed scaling, working at speed μ_2 and μ_1, respectively. Recall that these are available from [8] (see also [11]): $\theta(0, 1) = 2\mu_1/(2 - p_1^2)$ and $\theta(1, 0) = 2\mu_2/(2 - p_1^2)$, respectively. Moreover, the slope of the curve depends on the

value a, that is, if a is sufficiently small, the throughput is close to the lower bound, apart from the cases when $d \ll a$.

The stability condition obtained is useful for M/G-type model validation with an M/M-type one. This approach can also be used to obtain the system performance explicitly in an $M/M/2$-type model, however, this discussion is left for future research.

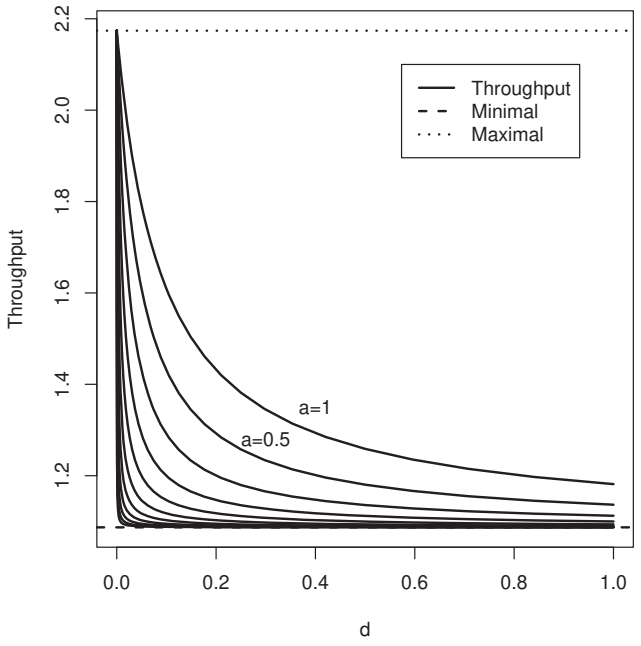

Fig. 1. Sensitivity analysis of the maximal throughput, $\theta(a, d)$, on the parameters $a = 2^i, i = -10, \ldots, 0$ and $d \in (0, 1]$

IV. Regenerative structure

Define by $\nu(t)$ and $Q(t)$ the number of customers at service and in the queue at time instant $t \geq 0$, respectively. Thus, the basic process X, associated the total number of customers, is presented as

$$X(t) = \{\nu(t) + Q(t);\, t \geq 0\}.$$

Consider $X_n = X(t_n^-)$, $n \geq 0$ – the discrete analogue of the process $X(t)$, which defines number of customers just before n-th arrival. Assume that the system has *zero initial state*: $X(t_0) = 0$ ($t_0 := 0$ for simplicity) and consider the following sequence:

$$\beta_n = \min_k\{k > \beta_{n-1} : X_k = 0\}, \quad n \geq 1,\, \beta_0 = 0.$$

Thus, β_n denotes the numbers of customers which arrive into empty system. At each such a moment the system starts over in stochastic sense, or *regenerates*.

Denote $\alpha_n = \beta_n - \beta_{n-1}$, $n \geq 1$. The process $\{X_n\}$ (and its continuous analogue $\{X(t)\}$) is called a *regenerative* process and has independent stochastically equivalent cycles with i.i.d. lengths $\{\alpha_n\}$ (denote a generic length by α). The sequence $\{\beta_n\}$ defines *regeneration points* of a process $\{X_n\}$ in discrete time. Regeneration points (instants) in continuous time $\{T_n\}$ (with generic element T) are related with β_n as follows:

$$T_n = t_{\beta_n}, \quad n \geq 1,\, T_0 = 0.$$

Thus, $\{T_n\}$ denotes such arrival instants, that customers enter an empty system. If the mean cycle length is finite, the process $\{X(t), t \geq 0\}$ is called *positive recurrent* [4], [12]. In general, positive recurrence also includes the condition for the first cycle:

$$T_1 < \infty, \quad \text{w.p. 1 (with probability 1)},$$

which automatically holds, since the process was considered without delay, $X(0) = 0$ (the first cycle is stochastically equal to a generic one).

If the condition $\mathsf{E}T < \infty$ holds good, then the regenerative method, which is a strong instrument in stochastic simulation and stability analysis, can be applied. More detailed description of regeneration approach is well-presented in [4], [13], [14], [15].

A. Regenerative estimation

Consider some measurable function f of a regenerative process $\{X(t), t \geq 0\}$, which represents a QoS characteristics of the considered system, and construct

$$r(t) = \frac{1}{t}\int_0^t f[X(u)]du,$$

which represents an "average" value of the process $f[X]$ at $[0,\, t]$. The following limit (if exists)

$$\lim_{t\to\infty} r(t) = r, \tag{3}$$

is a steady-state performance measure. Conditions of existence for the limit (3) are rather simple. Define the sequence, $\{Y_n\}$, of i.i.d. accumulated values of the characteristics over the segments of the process $\{X(t), t \geq 0\}$

$$\mathcal{Y}_n = \int_{T_{n-1}}^{T_n} f[X(u)]du, \quad n \geq 1.$$

If mean cycle length is finite ($\mathsf{E}T < \infty$), first cycle is finite $T_1 < \infty$ w. p. 1 and $\int_0^T |f[X(u)]|du < \infty$, then

$$\frac{1}{t}\int_0^t f[X(u)]du \to \frac{\mathsf{E}\mathcal{Y}}{\mathsf{E}T}. \tag{4}$$

Moreover, if the generic length T is non-lattice, the following weak convergence takes place

$$X(t) \Rightarrow X, \quad t \to \infty.$$

where X is a random variable (limit distribution). Thus,

$$\frac{\mathsf{E}\mathcal{Y}}{\mathsf{E}T} \equiv \mathsf{E}f[X].$$

Consider the corresponding values in discrete time

$$Y_n = \sum_{i=\beta_{n-1}}^{\beta_n - 1} f[X_i], \quad n \geq 1.$$

In case of positive recurrence, which means that the following conditions hold good,

$$\mathsf{E}Y < \infty, \quad \mathsf{E}\alpha < \infty,$$

the estimator

$$r_n := \frac{\sum_{k=0}^n Y_k}{\sum_{k=0}^n \alpha_k} = \frac{\sum_{i=0}^{\beta_n - 1} f[X_i]}{\beta_n}$$

converges w. p. 1 to a discrete analogue of (4). Namely,

$$r_n \to r = \frac{\mathsf{E}Y}{\mathsf{E}\alpha} = \frac{\sum_{i=0}^{\alpha-1} f[X_i]}{\mathsf{E}\alpha}, \quad n \to \infty. \tag{5}$$

Consider the sequence of i.i.d. variables

$$Z_n = Y_n - r\alpha_n, \quad n \geq 1$$

with a generic element Z. Obviously, $\mathsf{E}Z = 0$ and its variance is defined as follows

$$\mathsf{D}[Z] = \mathsf{D}[Y] - 2r\mathsf{Cov}(Y, \alpha) + r^2\mathsf{D}[\alpha].$$

By Central Limit Theorem, under the condition $\mathsf{D}[Z] < \infty$, we obtain a weak convergence

$$\frac{\sqrt{n}[r_n - r]}{\sqrt{D[Z]}/\mathsf{E}\alpha} \Rightarrow \mathcal{N}(0, 1), \quad n \to \infty, \tag{6}$$

where $\mathcal{N}(0, 1)$ denotes a standard normal distribution.

The convergence (6) allows to build an interval estimator for parameter r. In this way, the unknown constant $\sqrt{D[Z]}/\mathsf{E}\alpha$ is replaced by its strongly consistent estimator.

Obviously, that mean average $\overline{\alpha}_n := \frac{1}{n}\sum_{k=1}^{n}\alpha_n$ converges to $\mathsf{E}\alpha$ w. p. 1. Then we construct the following unbiased estimators

$$\begin{aligned}
\overline{\mathsf{D}}_n[\alpha] &= \frac{1}{n-1}\sum_{k=1}^{n}[\alpha_k - \overline{\alpha}_n]^2, \\
\overline{\mathsf{D}}_n[Y] &= \frac{1}{n-1}\sum_{k=1}^{n}[Y_k - \overline{Y}_n]^2, \\
\overline{\mathsf{Cov}}_n[\alpha, Y] &= \frac{1}{n-1}\sum_{k=1}^{n}\Big[(\alpha_k - \overline{\alpha}_n)(Y_k - \overline{Y}_n)\Big],
\end{aligned}$$

which converge w.p. 1 to $\mathsf{D}[\alpha]$, $\mathsf{D}[Y]$, $\mathsf{Cov}(\alpha, Y)$, respectively. Note, that $\overline{Y}_n$ is mean average among n independent replications of Y. Thus, having in mind (5), we obtain an estimator

$$s_n := \Big[\overline{\mathsf{D}}_n[Y] - 2r_n\overline{\mathsf{Cov}}_n[\alpha, Y] + (r_n)^2\overline{\mathsf{D}}_n[\alpha]\Big]^{\frac{1}{2}}, \tag{7}$$

which converges w.p. 1 to $\sqrt{D[Z]}$, as $n \to \infty$.

Hence, (6) evaluates to

$$\frac{\sqrt{n}[r_n - r]}{s_n/\overline{\alpha}_n} \Rightarrow \mathcal{N}(0, 1), \quad n \to \infty,$$

which implies the following $100(1-\gamma)\%$ confidence interval:

$$r \in \Big[r_n - \frac{z_\gamma s_n}{\overline{\alpha}_n\sqrt{n}}, r_n + \frac{z_\gamma s_n}{\overline{\alpha}_n\sqrt{n}}\Big].$$

Note, that γ is a given reliability, and z_γ is obtained from the statement

$$\Phi(z_\gamma) = \frac{1-\gamma}{2}, \tag{8}$$

where Φ is the Laplace function.

B. Estimation of the mean queue

Next, our goal is to apply presented regenerative method to the considered queueing system with simultaneous service and speed scaling. We illustrate confidence estimation of mean queue size (mean number of customers in the queue).

First, we present the conditions of applicability for regenerative method. Note, that the queue process $Q(t)$ regenerates on the same instants $\{T_n\}_{n\geq 0}$ (has same regeneration points $\{\beta_n\}_{n\geq 0}$ defined in discrete time) as the basic process X. The positive recurrence of Q is ($\mathsf{E}T < \infty$), which is equivalent to stability of the considered system.

Construct a single server (working at unit speed) model $M/G/1$ with the same arrival instants as in the original system and assume that service times are independent and distributed as S_{max}, where

$$S_{max} = S/\mu_1,$$

recall that μ_1 is the smallest speed scaling coefficient. Such a new system is stochastically *majorant system*, as its workload is greater or equal to the original model with simultaneous service and speed scaling (see [16]). Note, that the stability condition of the majorant system

$$\lambda\mathsf{E}S_{max} < 1. \tag{9}$$

automatically implies the stability (positive recurrence) of the original system, which means $\mathsf{E}T < \infty$.

Let $Q_n = Q(t_n^-)$ define a queue size just before the n-th arrival and construct i.i.d. sequence

$$Y_n = \sum_{i=\beta_{n-1}}^{\beta_n - 1} Q_i, \quad n \geq 1.$$

Note that $Q_{\beta_{n-1}} = 0$. Since n-th regenerative cycle consists of α_n arrivals,

$$Q_i < \alpha_n \text{ w.p. } 1, \quad i = \beta_{n-1} + 1, \cdots, \beta_n - 1,$$

which implies

$$Y_n < \sum_{i=1}^{\alpha_n}\alpha_n \text{ w.p. } 1. \tag{10}$$

Then construct the following estimator of the mean queue:

$$r_n := \frac{\sum_{i=0}^{\beta_n - 1} Q_i}{\beta_n}.$$

By Wald's identity [13] $\mathsf{E}T = \mathsf{E}\tau\mathsf{E}\alpha$. Note that the condition (9) provides finite mean cycle length in discrete time: $\mathsf{E}\alpha < \infty$. Next, from (10) and by Wald's identity:

$$\mathsf{E}Y < [\mathsf{E}\alpha]^2.$$

Under these conditions:

$$r_n \to \frac{\mathsf{E}Y}{\mathsf{E}\alpha}, \quad n \to \infty.$$

Moreover, assume that

$$\mathsf{P}(\tau > S_{max}) > 0. \tag{11}$$

Thus, with a positive probability, there exists a regeneration cycle having only one arrival, $\mathsf{P}(\alpha = 1) > 0$. Hence, cycle

length α is aperiodic (discrete analog of non-lattice property) and the queue process $\{Q_n;\, n \geq 1\}$ converges to its stationary distribution (r.v. Q).

$$Q_n \Rightarrow Q, \quad n \to \infty,$$

hence,

$$r_n \to \mathsf{E}Q, \quad n \to \infty.$$

Note, that condition (11) automatically holds for the models with Poisson input.

Since we consider the model in steady state, we expect that queue process does not go to infinity, and for any arbitrary time, the system becomes empty in a finite time with a positive probability. Thus, assuming the condition $\mathsf{D}[Y - \mathsf{E}Q \cdot \alpha] < \infty$, it is possible to build the following $100(1-\gamma)\%$ confidence interval for $\mathsf{E}Q$:

$$\mathsf{E}Q \in \Big[r_n - \Delta_n,\, r_n + \Delta_n\Big], \tag{12}$$

where

$$\Delta_n = \frac{z_\gamma s_n}{\overline{\alpha}_n \sqrt{n}}, \tag{13}$$

z_γ is obtained from (8), $\overline{\alpha}_n$ defines mean average cycle length, and s_n is an unbiased strongly consistent estimator of $\Big[\mathsf{D}(Y - \mathsf{E}Q \cdot \alpha)\Big]^{1/2}$, constructed similarly as in (7).

V. Simulation

In this section we present results of confidence estimation of mean queue size $\mathsf{E}Q$ for different configurations of considered 2-server system with $L = 2$ speed modes, assuming that service times a exponentially distributed with a rate μ and speeds are switched at arrival/departure instants, according to transition matrices A and D, defined in (1).

The mean queue $\mathsf{E}Q_{maj}$ in majorant system M/M/1 with a load coefficient

$$\overline{\rho} := \frac{\lambda}{[\mu_1 \mu]}$$

is defined by [17]

$$\mathsf{E}Q_{maj} = \frac{\overline{\rho}^2}{1 - \overline{\rho}}. \tag{14}$$

We additionally construct *minorant* queueing system M/M/2 with the same sequence of arrival instants as in the original system, but with exponential service times with a rate $\mu_2 \mu$. Such a stochastically minorant system (cf. monotonicity results in [16]) has the load coefficient

$$\underline{\rho} := \frac{\lambda}{[\mu_2 \mu]},$$

and its mean queue $\mathsf{E}\underline{Q}$ if defined by

$$\mathsf{E}Q_{min} = \frac{\underline{\rho}^3}{4 - \underline{\rho}^2}.$$

Thus, we obtain an obvious relation

$$\mathsf{E}Q_{min} \leq \mathsf{E}Q \leq \mathsf{E}Q_{maj}. \tag{15}$$

Our goal is to compare an interval (12) based on regenerative estimation with (rather rough) theoretical bounds (15) for different configurations of considered system. Note, that experiments were done under the condition

$$\overline{\rho} < 1,$$

which guarantees the stability of the majorant system and allows to apply statement (14). The simulation model is based on $100\,000$ arrivals. We illustrate the dynamics of r_n and corresponding 95% confidence intervals, where n is *the number of regeneration cycles*. Note, that in this case, the reliability $\gamma = 0.05$, which corresponds to $z_\gamma \approx 1.6449$.

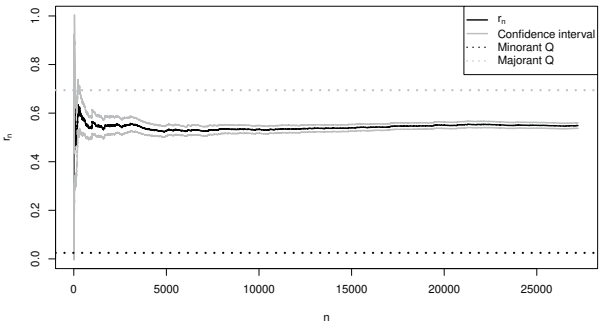

Fig. 2. Regenerative estimation of $\mathsf{E}Q$, for $\rho = 0.5$, $\mu_1 = 0.9$, $\mu_2 = 1.1$

Fig. 2 presents the case for $\lambda = 1, \mu = 2$. The corresponding probabilities are defined by $a = 0.3$, $d = 0.4$, $p_1 = 0.5$ and speed scaling parameters are $\mu_1 = 0.9$, $\mu_2 = 1.1$. Thus,

$$\overline{\rho} = 0.56,\ \mathsf{E}Q_{maj} = 0.6944,\ \mathsf{E}Q_{min} = 0.0248.$$

For considered case we obtained $n = 27286$ regeneration cycles, thus more than 25% customers arrive into totally empty system. The interval (grey solid lines) becomes thinner with a growth of n, this fact is easily explained by the construction of Δ_n, see (13). The confidence interval built by regeneration method is much more accurate than theoretical interval (dash lines), based on bounds (15). Consider the relative measure

$$\delta := \frac{\mathsf{E}Q_{maj} - \mathsf{E}Q_{min}}{2},$$

which is equal to a half of theoretical interval. As,

$$\frac{\delta}{\Delta_n} = \frac{0.3348}{0.0097} = 34.52,$$

the regenerative confidence estimation provides in 34 times more accurate interval for $\mathsf{E}Q$, than bounding based on monotonicity properties.

Table I illustrate results of interval estimating of $\mathsf{E}Q$ for different configurations of considered system, having load coefficient λ/μ fixed and equal to 0.5. Note, that columns Q_{maj}, Q_{min} present the results for *mean* queue length in majorant and minorant systems, respectively.

The first block of the Table I corresponds to the case of "medium" probabilities a, d, p_1. Obviously, greater values of deviation $[\mu_2 - \mu_1]$ provide less accurate theoretical intervals (greater δ). Moreover, with the growth of $[\mu_2 - \mu_1]$, we

TABLE I. SIMULATION RESULTS FOR $\lambda = 1$, $\mu = 2$.

μ_1	μ_2	$\overline{\rho}$	n	r_n	Q_{maj}	Q_{min}	Δ_n	δ
$a = 0.3$, $d = 0.4$, $p_1 = 0.5$.								
1.0	1.0	0.50	28528	0.4835	0.5000	0.0333	0.0088	0.2333
0.9	1.1	0.56	27286	0.5358	0.6944	0.0248	0.0097	0.3348
0.7	1.4	0.71	21582	0.9305	1.7857	0.0118	0.0241	1.8870
0.6	2.0	0.91	15509	1.7671	9.0909	0.0040	0.0721	4.5435
$a = 0.9$, $d = 0.1$, $p_1 = 0.5$.								
0.9	1.1	0.56	27742	0.5107	0.6944	0.0248	0.0098	0.3348
0.7	1.4	0.71	23229	0.8652	1.7857	0.0118	0.0241	1.8870
0.6	2.0	0.91	18023	1.5712	9.0909	0.0040	0.0611	4.5435
$a = 0.1$, $d = 0.9$, $p_1 = 0.5$.								
0.9	1.1	0.56	27132	0.5465	0.6944	0.0248	0.0108	0.3348
0.7	1.4	0.71	20051	1.0198	1.7857	0.0118	0.0241	1.8870
0.6	2.0	0.83	17119	1.3701	4.1667	0.0040	0.0399	2.0814

observed a slight increase in the load of a speed scaling system, which implies the larger values of r_n, less regeneration (the system becomes empty less number of times) and larger values of Δ_n. Anyway, the benefit of regenerative confidence estimation in comparison with theoretical bounding is illustrated for the case with high speed scaling. Namely, for $\mu_1 = 0.55$, $\mu_2 = 2$: $\delta/\Delta_n = 63.02$, while for the "worst" case (without speed scaling) $\mu_1 = \mu_2 = 1$: $\delta/\Delta_n = 26.51$.

The second block of the Table I illustrates the case for $a = 0.9$, thus with a high probability the system switches to the second speed regime (with a faster service) at arrival instants. Small $d = 0.1$ also contributes to keep the second mode at departure instants, which implies a slightly less loaded system (in comparison with the first block of the Table I), more regeneration epochs, smaller r_n and more accurate confidence interval.

The last block of the Table I illustrates the case for $a = 0.1$, $d = 0.9$, and such a configuration tries to keep the first speed regime, which corresponds to the slower service and provides slightly larger load, less regeneration points and increase in average queue.

Simulation results for other configuration and under the condition $\overline{\rho} < 1$ had shown rather similar results: the most illustrative advantage of regenerative estimation in comparison with theoretical interval is obtained to the cases with high speed scaling. Variation of parameters a, b, p_1, ρ does not strongly affect the fraction δ/Δ_n.

VI. CONCLUSION

In this paper, we considered a specific case of non-work-conserving queueing model with simultaneous service and speed scaling policy. For $M/M/2$-type of such a system the stability condition was derived and sensitivity of the system throughput was studied by means of matrix-analytic method for various system configurations. Next, considering the $M/G/2$-type model in steady state, we applied the regenerative method, which is a strong instrument in stochastic analysis. We presented some numerical results related to regenerative confidence estimation of the mean queue size, and illustrated the advantages of such a method in comparison with interval estimation, based on monotonicity properties, for the different load coefficients and speed scaling parameters.

ACKNOWLEDGMENT

The study was carried out under state order to the Karelian Research Centre of the Russian Academy of Sciences (Institute of Applied Mathematical Research KRC RAS).

The research is partially supported by Russian Foundation for Basic Research, projects 18-07-00147, 18-07-00156, 19-07-00303, 19-57-45022.

REFERENCES

[1] V. Kartsch, M. Guermandi, S. Benatti, F. Montagna and L. Benini, "An Energy-Efficient IoT node for HMI applications based on an ultra-low power Multicore Processor," 2019 IEEE Sensors Applications Symposium (SAS), Sophia Antipolis, France, 2019, pp. 1-6.

[2] M. A. Crane and A. J. Lemoine, *An Introduction to the Regenerative Method for Simulation Analysis*. Berlin: Springer-Verlag, 1977.

[3] A. Law and D. Kelton, *Simulation Modeling and Analysis*, 5th edn. McGraw-Hill, 2014.

[4] E. Morozov, "Weak regeneration in modeling of queueing processes", *Queueing Systems*, vol. 46, 2004, pp. 295-315.

[5] E. Morozov, "A multiserver retrial queue: Regenerative stability analysisWeak regeneration in modeling of queueing processes", *Queueing Systems*, vol. 56, 2007, pp. 157-168.

[6] E. Morozov, R. Nekrasova R, "Stability Conditions of a Multiclass System with NBU Retrials", *Queueing Theory and Network Applications. QTNA 2019. Lecture Notes in Computer Science*, vol. 11688, 2019, pp. 51-63, doi: 10.1007/978-3-030-27181-7-4.

[7] D.G. Feitelson, *Workload modeling for computer systems performance evaluation*. Cambridge University Press, 2015, doi: 10.1017/CBO9781139939690

[8] A. Rumyantsev and E. Morozov, "Stability criterion of a multiserver model with simultaneous service", *Annals of Operations Research*, vol. 252, no. 1, 2017, pp. 29–39, doi: 10.1007/s10479-015-1917-2.

[9] G. Rama Murthy and A. Rumyantsev, "On an exact solution of the rate matrix of g/m/1-type markov process with small number of phases", *Journal of Parallel and Distributed Computing*, vol. 119, 2018, pp. 172–178.

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

[11] P. H. Brill and L. Green, "Queues in which customers receive simultaneous service from a random number of servers: a system point approach", *Management Science*, vol. 30, no. 1, 1984, pp. 51–68, doi: 10.1287/mnsc.30.1.51.

[12] K. Sigman and R.W. Wolff, "A review of regenerative processes", *SIAM Review*, vol. 35, 1993, pp. 269-288.

[13] S. Asmussen, *Applied Probability and Queues*. Wiley: New York, 1987.

[14] E. Morozov, "The tightness in the ergodic analysis of regenerative queueing processes", *Queueing Systems*, vol. 27, 1997, pp. 179-203.

[15] E. Morozov and R. Delgado, "Stability analysis of regenerative queues", *Automation and Remote control*, vol. 70, 2009, pp. 1977-1991.

[16] E. Morozov, A. Rumyantsev and I. Peshkova, "Monotonicity and stochastic bounds for simultaneous service multiserver systems", *2016 8th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, 2016, pp. 294-297.

[17] T. Saaty, *Elements of queueing theory, with applications*. New York: McGraw-Hill, 1961.