Inequalities for Workload Process in Queues with NBU/NWU Input

Evsey Morozov[1,2], Alexander Rumyantsev[1,2](✉), and Kseniya Kalinina[1]

[1] Institute of Applied Mathematical Research, Karelian Research Centre, RAS,
Petrozavodsk, Karelia, Russia
emorozov@karelia.ru, ar0@krc.karelia.ru, kalininaksenia90@gmail.com
[2] Petrozavodsk State University, Petrozavodsk, Karelia, Russia

Abstract. We give a simple proof of the well-known property PASTA for the workload and queue size process in the queueing systems with Poisson input. The proof is based on a relation connecting the workload process at an arbitrary instants and the arrival instants of the customers and, in particular, yields famous Pollaczeck-Khintchine equality. It is then shown that this equality transforms to the corresponding inequality when the interarrival time has New-Better-than-Used (NBU) or New-Worse-than-Used (NWU) distribution. Then the inequalities for the stationary workload are extended to classical multiserver system and multiserver model with simultaneous service, describing the modern high performance cluster. The analysis is illustrated by simulation of the single-server and cluster models with Weibull interarrival time, covering NBU, NWU and Poisson inputs. The obtained results are of practical interest for Quality-of-Service estimation of modern high-performance computing systems.

Keywords: Property PASTA · NBU/NWU input · Regeneration · Pollaczeck–Khintchine formula · Inequalities for workload · High performance cluster

1 Introduction

It is well-known that the so-called PASTA (Poisson Arrivals See Time Average) property holds if and only if the input to a queueing system is Poisson. Under PASTA, the limiting fraction of time that the system spends in an arbitrary fixed state equals the limiting fraction of the arrivals which meet the system in this state. Moreover, the property PASTA allows to establish the celebrated Pollaczeck-Khintchine formula expressing the mean stationary waiting time (stationary workload) via predefined moments of the governing sequences (interarrival times and service times). In general, the relation between the stationary distributions of the queueing systems at the *event instants* and arbitrary instants have been studied in a number of the works, with the pioneering one [10], see also [3–5,7,9,11].

A. Gruca et al. (eds.), *Man-Machine Interactions 5*,
Advances in Intelligent Systems and Computing 659, DOI 10.1007/978-3-319-67792-7_52

In this note, we first give a simple regenerative proof of PASTA for the workload and queue-size process in FIFO single-server model with Poisson input. Then we show that replacement of the exponential interarrival time distribution (corresponding to the Poisson input) by the NBU/NWU distribution yields stochastic *inequalities* between the basic processes and, as a result, transforms Pollaczeck-Khintchine equality to the corresponding inequality. The earlier known proofs of this result are rather complicated and, in particular, based on the theory of stationary point processes [3]. Our proof is based on a modified Lindley recursion, connecting the (*imbedded*) sequence of waiting times of customers in the queue, and the waiting time process at arbitrary instant. This relation, properly modified, is then used to establish the property PASTA in the multiserver system $GI/G/m$ and $GI/G/m$-type model with simultaneous service, describing the workload of a high-performance cluster. Then we show how the property PASTA is modified if the exponential interarrival time is replaced by the NBU/NWU interarrival time. As a result, the upper (respectively, lower) bound for the stationary workload, both at the imbedded and the arbitrary instants is obtained. An important ingredient of the research is the verification of the obtained inequalities by simulation of the systems with Weibull interarrival time, where we consider NBU, NWU and Poisson inputs. In particular, we demonstrate how the *tightness* of the obtained bounds depends on the proximity between the interarrival time NBU/NWU distribution and the exponential distribution. The obtaining these bounds is also an important motivation of this research, because the mean stationary workload, being the key QoS parameter of the system, is typically analytically unavailable in case of the non-Poisson inputs.

Thus, the contribution of this research is as follows. A novel simple proof of property PASTA both for classical queueing systems and for the cluster model with simultaneous service. The inequalities for the stationary workload and queue size process in the models with NBU/NWU input. Verification of the tightness of the obtained inequalities (as the corresponding bounds of the stationary performance) depending on the given parameters, by simulation of the systems with the Weibull interarrival time.

The paper is organized as follows. In Sect. 1, we derive the property PASTA both for the workload process and for the queue size process. Although this result is well-known, we give very simple and transparent proof based on coupling method, regenerative arguments and relation (2) connecting continuous-time and imbedded (discrete-time) processes. The latter relation is then allows to deduce corresponding inequalities in the Pollaczeck-Khintchine formula in the case when the input belongs to the class of NBU/NWU distributions (Sect. 3). In Sect. 3 we also discuss Little's formula and derive corresponding inequalities for the workload vector process in classical multiserver queue $GI/G/m$, while, in Sect. 4, we consider the cluster model with simultaneous service with the NBU/NWU input process. In this analysis we apply a modification of the classical Kiefer-Wolfowitz recursion defining the workload sequence. Finally, in Sect. 5 we present results of simulation the workload process in single-server system and cluster model with Weibull interarrival time, which covers NBU, NWU and Poisson inputs.

2 PASTA Property

Consider a single-server infinite buffer FIFO queueing system $GI/G/1$ with the renewal arrival instants t_n, independent identically distributed (iid) interarrival times $\tau_n := t_{n+1} - t_n$ and the iid service times $S_n, \geq 1$. Here and in what follows, we omit serial index to denote a generic element of an iid sequence, and introduce input rate $\lambda := 1/\mathsf{E}\tau \in (0, \infty)$ and service rate $\mu := 1/\mathsf{E}S \in (0, \infty)$. Assume stability, that is $\rho := \lambda/\mu < 1$ and denote by $W(t)$ the workload (current work) in the system at instant t^-. Also denote $W(t_n^-) = W_n$, and let $Z(t)$ be the number of arrivals in $[0, t)$. First of all we deduce the property PASTA for the workload of the system [1]. To this end, we write well-known Lindley recursion defining the sequence W_n:

$$W_{n+1} = [W_n + S_n - \tau_n]^+, \; n \geq 0, \tag{1}$$

where W_0 is the initial workload. It is easy to see that the following relation holds

$$W(t) = [W_{Z(t)} + S_{Z(t)} - \bar{\tau}(t)]^+, \; n \geq 0, \tag{2}$$

where $\bar{\tau}(t) := \inf_n (t - t_n : \; t - t_n > 0)$ is the attained interarrival time at instant t and $S_{Z(t)}$ is the service time of customer $Z(t)$, the last customer arriving before instant t. In particular, $Z(t_{n+1}) = n$. Since, by construction, $\bar{\tau}(t_{n+1}) = \tau_n$, then relation (2) transforms to (1) for $t = t_n$. Since stability condition $\rho < 1$, or equivalently, $\mathsf{E}\tau > \mathsf{E}S$, implies $\mathsf{P}(\tau > S) > 0$, then the sequence $\{W_n\}$ is aperiodic. Now we define the regeneration instants of this sequence as

$$\beta_{n+1} = \inf(k > \beta_n : W_k = 0), \; n \geq 0 \;\; (\beta_0 = 0). \tag{3}$$

In other words, regenerations generated by arrivals of the non-waiting customers. It is well-known that under $\rho < 1$, this sequence is positive recurrent, that is the mean *generic* regeneration period is finite, $\mathsf{E}\beta < \infty$ [1]. By the aperiodicity of the regeneration period, there exists the weak limit $W_n \Rightarrow W_\infty, \; n \to \infty$, where W_∞ is the stationary workload and $\Rightarrow$ stands for the convergence in distribution.

Now we consider continuous-time workload process $W(t)$, $t \geq 0$ with regeneration instants defined as $T_n = t_{\beta_n}$, $n \geq 0$. Let T be generic regeneration period in continuous time, that is, $T =_{st} T_{n+1} - T_n$. Let β be the generic regeneration period in discrete time. Then, by the Wald's identity, $\mathsf{E}T = \mathsf{E}\tau\mathsf{E}\beta$. Assume that interarrival time τ is *non-lattice* (it holds automatically for the Poisson input). Then the regeneration period T is so, and (provided $\rho < 1$) the weak limit

$$W(t) \Rightarrow W(\infty), \; t \to \infty, \tag{4}$$

exists as well [1]. Note that $Z(t) \to \infty$ with probability (w.p.1), as $t \to \infty$, and *stochastic* equivalence $S_{Z(t)} =_{st} S$ holds.

Now we *assume that* τ *is exponential.* Then $\bar{\tau}(t) =_{st} \tau$ for any t. Also note that the mapping $[\cdot]^+$ is continuous. Then, taking limits in (1) and (2) as $n \to \infty$, $t \to \infty$, respectively, we apply continuous mapping theorem [2] to obtain the stochastic equivalence,

$$W(\infty) =_{st} [W_\infty + S - \tau]^+ =_{st} W_\infty. \tag{5}$$

This relation expresses property PASTA [10] of the workload process. The similar analysis can be easily developed for $\nu(t)$, the number of customers in the system, and for $Q(t)$, the queue size, at arbitrary instant t, and at the arrival instants. Really, denote $\nu(t_n^-) = \nu_n$, $Q_n = Q(t_n^-)$, and let $d(\tau_n)$ be the number of departures during interarrival time τ_n, provided the server is *permanently busy* in interval $[t_n,\, t_{n+1}]$. Then we can write evident balance relations:

$$\nu_{n+1} = [\nu_n + 1 - d(\tau_n)]^+, \tag{6}$$

$$\nu(t) = [\nu_{Z(t)} + 1 - d(\bar{\tau}(t))]^+, \tag{7}$$

which as above give (in evident notation) stochastic equivalence between the limiting number of customers in the system, $\nu(\infty)$, ν_∞, at arbitrary and arrival instants, respectively. Then we obtain stochastic equality,

$$\nu_\infty =_{st} [\nu_\infty + 1 - d(\exp)]^+ =_{st} \nu(\infty), \tag{8}$$

expressing property PASTA of the the queue size process. (We denote by exp the exponential r.v.). This property is also known as ESTA and ASTA, see [9]. Now we apply regenerative approach [6], and using geometrical considerations we obtain

$$\begin{aligned}\mathsf{E}W(\infty) =: \lim_{t\to\infty} \frac{1}{t}\int_0^t W(u)du &= \frac{\mathsf{E}\int_0^T W(u)du}{\mathsf{E}T} \\ &= \lambda \frac{\mathsf{E}\Big[\sum_{i=0}^{\beta-1}(W_n\,S_n + S_n^2/2)\Big]}{\mathsf{E}\beta} = \lambda(\mathsf{E}W_\infty\,\mathsf{E}S + \frac{\mathsf{E}S^2}{2}).\end{aligned} \tag{9}$$

By (5), for the Poisson input, we denote $W(\infty) =_{st} W_\infty =_{st} W$, and, by (9), immediately arrive to Pollaczeck-Khintchine formula:

$$\mathsf{E}W = \frac{\lambda \mathsf{E}S^2}{2(1-\rho)}. \tag{10}$$

3 NBU/NWU Input

Now we replace exponential τ by random variable with the NBU distribution $F_\tau := F$ meaning that (denoting tail $\bar{F} = 1 - F$) for any $x,\, y \geq 0$,

$$\bar{F}(x+y) \leq \bar{F}(y)\bar{F}(x). \tag{11}$$

Now we need to define also the remaining interarrival time at instant t as

$$\hat{\tau}(t) := \inf_n (t_n - t:\ \ t_n - t \geq 0). \tag{12}$$

Note that $\hat{\tau}(t_n) = 0$. Ordering (11) is equivalent to (by coupling) the following inequality between original variable τ and the remaining interarrival time $\hat{\tau}(t)$

$$\tau \geq_{st} \hat{\tau}(t). \tag{13}$$

Recall that, by the renewal theory, both variables, $\hat{\tau}(t)$ and $\bar{\tau}(t)$, are equivalent in the limit as $t \to \infty$, $\hat{\tau}(\infty) =_{st} \tau(\infty) := \hat{\tau}$, and have the so-called integrated-tail distribution (with $\lambda := 1/\mathsf{E}\tau$) [1]:

$$G(x) := \mathsf{P}(\hat{\tau} \le x) = \lambda \int_0^x (1 - F(u))du. \tag{14}$$

Now (2) and (13) yield

$$W(\infty) =_{st} [W_\infty + S - \hat{\tau}]^+ \ge_{st} [W_\infty + S - \tau]^+ =_{st} W_\infty. \tag{15}$$

Because relation (9) holds for arbitrary distributions, in particular, for NBU and NWU distributions, then Pollaczeck-Khintchine equality (10) transforms, in the NBU case, in the following inequality

$$\mathsf{E}W_\infty \le \mathsf{E}W(\infty) \le \frac{\lambda \mathsf{E}S^2}{2(1-\rho)}. \tag{16}$$

Analogously, if τ has NWU distribution, that is, for any x, $y \ge 0$,

$$\bar{F}(x+y) \ge \bar{F}(y)\bar{F}(x), \tag{17}$$

then

$$\mathsf{E}W_\infty \ge \mathsf{E}W(\infty) \ge \frac{\lambda \mathsf{E}S^2}{2(1-\rho)}. \tag{18}$$

Also note the evident inequalities for the k-moments,

$$\mathsf{E}W^k(\infty) \ge \mathsf{E}W_\infty^k, \quad \mathsf{E}W^k(\infty) \le \mathsf{E}W_\infty^k,\ k > 0, \tag{19}$$

for NBU and NWU distributions, respectively.

Denote $Q(t)$ the queue size at instant t, and let $Q(t) \Rightarrow Q(\infty)$. Then we obtain, by regenerative arguments as $t \to \infty$, the famous Little's formula

$$\begin{aligned}\mathsf{E}\nu(\infty) := \lim_{t\to\infty} \frac{1}{t}\int_0^t \nu(u)du = \lambda \frac{\mathsf{E}\Big[\sum_{j=0}^{\beta-1}(W_j + S_j)\Big]}{\mathsf{E}\beta} \\ =\lambda \mathsf{E}W_\infty + \lambda \mathsf{E}S = \mathsf{E}Q(\infty) + \rho,\end{aligned} \tag{20}$$

connecting stationary mean queue size and workload, which holds for general interarrival time τ. Using (19), we can obtain the corresponding inequality instead of the equality (20). For instance, for the NBU input, we obtain

$$\lambda \mathsf{E}W(\infty) \ge \mathsf{E}\nu(\infty), \tag{21}$$

while for the NWU input,

$$\lambda \mathsf{E}W(\infty) \le \mathsf{E}\nu(\infty). \tag{22}$$

We now consider a multiserver FCFS system $GI/G/m$ with m stochastically equivalent servers and the same iid sequence of interarrival times $\{\tau_n,\ n \ge 1\}$ [8].

We assume stability condition $\rho < m$ to be held. Denote by $W_k^{(i)}$ the ith smallest remaining workload at the arrival instant t_k^- of customer $k \geq 1$, and let $W_k := (W_k^{(1)}, \dots, W_k^{(m)})$, where $W_k^{(i)} \leq W_k^{(i+1)}$. The sequence of vectors $\{W_n\}$ satisfies the following celebrated Kiefer-Wolfowitz recursion

$$W_{n+1} = R\Big((W_n^{(1)} + S_n - \tau_n)^+, \dots, (W_n^{(m)} - \tau_n)^+\Big), \tag{23}$$

where operator R puts the components of vector in an increasing order. Denote by $W_i(t)$ the ith smallest workload at instant t, and let $W(t) = (W_1(t), \dots, W_m(t))$ be continuous-time analogue of the Kiefer-Wolfowitz vector. As above and using the same notation, we can write continuous-time analogue of (23)

$$W(t) = R\Big((W_{Z(t)}^{(1)} + S_{Z(t)} - \bar{\tau}(t))^+, \dots, (W_{Z(t)}^{(m)} - \bar{\tau}(t))^+\Big). \tag{24}$$

For the Poisson input, (24) implies property PASTA as the stochastic vector equality $W(\infty) =_{st} W_\infty$. Moreover, we obtain (component-wise) inequality $W(\infty) \geq_{st} W_\infty$ for the NBU interarrival time τ, while for the NWU interarrival time, it follows that $W(\infty) \leq_{st} W_\infty$. In particular, for the mean stationary summary workload, it gives the inequality

$$\sum_{i=1}^{m} \mathsf{E}W_i(\infty) \geq \sum_{i=1}^{m} \mathsf{E}W_\infty^{(i)}, \tag{25}$$

when τ is NBU, and the opposite inequality when τ is NWU.

4 Cluster Model

We now consider a recently proposed m-server FCFS $GI/G/m$-type system with *simultaneous service* (describing modern high-performance clusters) [8], the same input $\{\tau_n\}$, in which customer k needs N_k servers simultaneously for a random service time S_i being *identical* at all N_k occupied servers. Variables $\{N_i\}$ are iid. For the cluster model, the following modified Kiefer-Wolfowitz recursion holds true,

$$\begin{aligned} W_{i+1} =& R\Big(\overbrace{(W_{i,N_i} + S_i - \tau_i)^+, \dots, (W_{i,N_i} + S_i - \tau_i)^+}^{N_i\ components}, \\ & (W_{i,N_i+1} - \tau_i)^+, \dots, (W_{i,m} - \tau_i)^+\Big), \quad i \geq 0, \end{aligned} \tag{26}$$

where we denote by $W_{i,k}$ the kth minimal component of the workload vector W_i at the arrival instant of customer i. A key difference between recursion (26) and classical recursion (2) is that, in the cluster model, service time S_i of customer i is added, at the arrival instant t_i of customer i, to the remaining work in the *first* N_i servers having *minimal remaining works.* Recall that $Z(t)$ is the number of arrivals in the interval $[0, t)$. Then continuous-time analogue of expression (26) is (in an evident notation)

$$\begin{aligned} W(t) =& R\big(\overbrace{(W_{Z(t),\,N_{Z(t)}} + S_{Z(t)} - \bar{\tau}(t))^+, \dots, (W_{Z(t),\,N_{Z(t)}} + S_{Z(t)} - \bar{\tau}(t))^+}^{N_{Z(t)}\ components}, \\ & (W_{Z(t),\,N_{Z(t)}+1} - \bar{\tau}(t))^+, \dots, (W_{Z(t),\,m} - \bar{\tau}(t))^+\Big), \quad t \geq 0. \end{aligned} \tag{27}$$

Note that (27) becomes (26) with $t = t_{i+1}$. Now we assume stability, see [8], implying the existence of the corresponding weak limits. Comparing (26) and (27), it is easy to check that, for exponential τ, PASTA holds again,

$$W(\infty) =_{st} W_\infty. \tag{28}$$

Moreover, as above we obtain vector analogue of (component-wise) inequality (3) for NBU interarrival time τ, and the inequality $W(\infty) \leq_{st} W_\infty$ for the NWU interval τ.

It is worth mentioning that recursions (1), (26) define discrete-time Markov chains, while recursions (2), (27) define continuous-time Markov processes.

5 Verification of the Inequalities by Simulation

In this section, we apply simulation to verify the *accuracy* of the obtained bounds for the stationary workload in single-server system and in the cluster model. We use Weibull interarrival time τ with the density

$$f(x) = \frac{a}{b}\left(\frac{x}{b}\right)^{a-1} e^{-(x/b)^a}, x \geq 0, \tag{29}$$

which is NBU distribution, if the exponent $a > 1$, NWU distribution, if $a < 1$, and exponential with parameter $1/b$, if $a = 1$.

We consider first a single-server system $Weibull/M/1$ with service time parameter $\mu = 1$, and study two scenarios: light traffic, $\rho = 0.1$, and heavy traffic, $\rho = 0.9$, where $\rho := 1/\mathsf{E}\tau$. We vary parameter $a \in [0.5, 2]$ (with step 0.02), and, for ρ fixed, find parameter b from the expression

$$\frac{1}{\rho} = \mathsf{E}\tau = b\Gamma(1 + \frac{1}{a}), \tag{30}$$

where Γ is the gamma-function. To evaluate the proximity between $\mathsf{E}W(\infty)$, $\mathsf{E}W_\infty$ and *Pollaczeck-Khintchine value* (the r.h.s of (10)), we perform 100 experiments with $N = 10^5$ arrivals in each experiment, and then take the sample mean. Two plots, corresponding to light and heavy traffics, are depicted on Fig. 1, where $\hat{W}(\infty)$, $\hat{W}_\infty$ denote the (sample mean) estimates of $\mathsf{E}W(\infty)$ and $\mathsf{E}W_\infty$, respectively. Note that, because for given parameters, $\mathsf{E}S^2/2 = 1$, then the r.h.s. of Pollaczeck-Khintchine (10) equals 1/9 for $\rho = 0.9$, and 9 for $\rho = 0.1$, see Fig. 1.

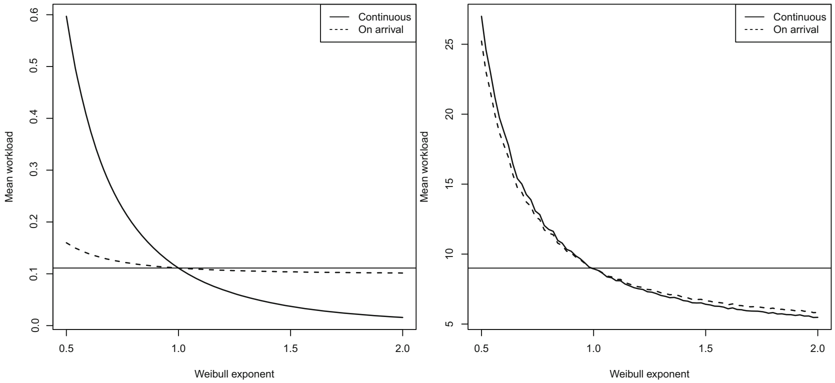

Fig. 1. The Pollaczeck–Khintchine value (straight line) vs. sample mean estimates of the stationary workloads, $\hat{W}(\infty)$, $\hat{W}_\infty$, in system $Weibull/M/1$: light traffic $\rho = 0.1$ (left) and heavy traffic $\rho = 0.9$ (right).

It is seen that, for the light traffic, the mean workload at the arrival epochs corresponding to the actual delays of the customers, is very close to Pollaczeck–Khintchine value, for all values of parameter a (Weibull exponent). It indicates that, in the light traffic system with NBU/NWU input, Pollaczeck–Khintchine formulae can be used for an accurate upper/lower bound, respectively, for the average stationary actual delay of the customers. In heavy traffic, the difference between the estimates and the Pollaczeck–Khintchine value is significant, however both estimates are remarkably close, and it shows that in this case, instead of the two estimates, we can calculate only the more simple estimate $\hat{W}_\infty$.

Next simulation experiments describe a $Weibull/M/10$ cluster system (where an analog of the Pollaczeck–Khintchine formulae does not exist) with exponential service time with parameter $\mu = 1$. We have assumed that the (sufficient) stability condition of this system is

$$\rho := \frac{1}{\mu \mathsf{E}\tau} \sum_{k=1}^{m} \frac{1}{k} \sum_{j=k}^{m} p_j^{k*} \sum_{t=m+1-j}^{m} p_t < 1, \tag{31}$$

where $p_j^{*i} := \mathsf{P}(N_1 + \cdots + N_i = j)$ with N_i being the number of servers required by customer i, and p_i is the probability that a customer requires i servers. (This stability criterion has been proved for $MAP/M/$-type cluster model in [8].) As the experiments show, condition (31) indeed implies stability of the cluster model under consideration.

We see that, for light traffic, the difference between two estimates is significant, while for heavy traffic, both estimates almost coincide, see Fig. 2. In practice it allows to perform the QoS check of a high-performance cluster only in the (discrete) moments of the arrival of customers, to estimate the continuous-time performance. This effect seems to be quite important, since extremely

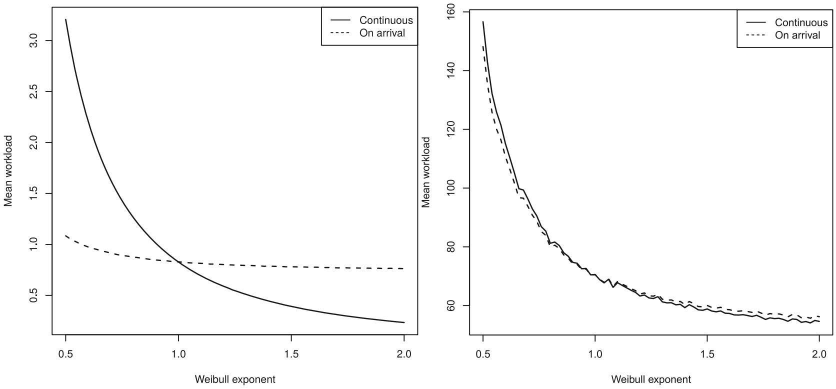

Fig. 2. The estimates of the mean stationary summary workload in continuous time, $\hat{\mathcal{W}}(\infty)$, and at arrival epochs, $\hat{\mathcal{W}}_\infty$, in $Weibull/M/10$ cluster model: light traffic $\rho = 0.1$ (left) and heavy traffic $\rho = 0.9$ (right).

frequent observations (to calculate continuous-time performance), being time- and energy-consuming, have negative impact on the overall QoS of the system.

It is instructive to discuss in brief how to simulate and estimate the workload vector process in continuous time for the cluster model (which include the single-server model as well). Note that (26) describes the evolution of the workload vector process at the arrival epochs only, but for simulation of the continuous-time process we need to include also the departure epochs. To this end, we recall that W_{n,N_n} is the delay of the nth customer, so its departure epoch is

$$D_n := t_n + W_{n,N_n} + S_n,\ n \ge 1. \tag{32}$$

Now we consider together arrival and departure epochs, and denote by $\hat{t}_n$ the nth order statistics among $\{t_n, D_n, n \geqslant 1\}$. If $\hat{t}_n$ is a departure epoch, then we declare this epoch to be *an arrival of observer*, with the following properties: $N_n = 1,\ S_n = 0$. In other words, an observer does not bring any additional work to the system and does not induce additional idle time of servers included in the workload vector and caused by insufficient resources for the head-of-line customer in the queue, see (26). (Note that we also have to renumber the driving sequences $\{S_n, N_n, n \ge 1\}$ properly.) Now recursion (26) formally covers also departure epochs, and interarrival times become $\tau_n := \hat{t}_{n+1} - \hat{t}_n$. It now follows that the vector $W(t)$ can be obtained via the imbedded vector W_n by recursion (27), where now the counting process $Z(t) := \max(k : \hat{t}_k < t)$. An important QoS parameter is the summary workload at time t, $\mathcal{W}(t) := \sum_{j=1}^m W_j(t)$, which can now be calculated by means of the imbedded sequence $\mathsf{W}_i := \mathcal{W}(t_i)$ as

$$\mathcal{W}(t) := \mathsf{W}_{Z(t)} + S_{Z(t)} - \overline{\tau}(t)\sum_{j=1}^{m} I\left(W_j\left(t^+_{Z(t)}\right) > 0\right), \tag{33}$$

where I is the indicator function. It remains to define the consistent estimator of the mean stationary summary workload in continuous time, $\mathsf{E}\mathcal{W}(\infty)$, as

$$\hat{\mathcal{W}}(\infty) = \frac{1}{T} \sum_{i<Z(T)} \left[(\mathsf{W}_i + S_i)\tau_i - \frac{\tau^2}{2} \sum_{j=1}^{m} I\left(W_j\left(t_i^+\right) > 0 \right) \right], \tag{34}$$

which is based on a given simulation period $[0, T]$, where we apply simple geometrical properties of the summary workload process trajectory between arrivals. The sample mean estimator of the stationary summary workload at the arrival epochs, $\mathsf{E}\mathcal{W}_\infty$ (based on N observations) is $\hat{\mathcal{W}}_\infty = 1/N \sum_{i=1}^{N} \mathsf{W}_i$, see Fig. 2.

6 Conclusion

We give a simple proof of the known property PASTA. For the NBU/BWU input, we obtain the corresponding inequalities for the stationary workload at arbitrary instants and at arrival instants both for classical (multiserver) system and for the cluster model with simultaneous service. The analysis is verified by simulation of the systems with Weibull input.

Acknowledgements. The authors thank Masakiyo Miyazawa who has attracted our attention to the papers [3–5]. The research is supported by Russian Foundation for Basic Research, projects 15–07–02341, 15–07–02354, 15–07–02360, 15–29–07974, 16–07–00622, RF President's Grant No. MK-1641.2017.1.

References

1. Asmussen, S.: Applied Probability and Queues. Springer, New York (2003)
2. Billingsley, P.: Convergence of Probability Measures. Wiley, Hoboken (1999)
3. König, D., Schmidt, V.: Stochastic inequalities between customer-stationary and time-stationary characteristics of queueing systems with point processes. J. Appl. Probab. **17**(3), 768–777 (1980)
4. Miyazawa, M.: Stochastic order relations among GI/G/1 queues with a common traffic intensity. J. Oper. Res. Soc. Jpn. **19**(3), 193–208 (1976)
5. Miyazawa, M.: A formal approach to queueing processes in the steady state and their applications. J. Appl. Probab. **16**(2), 332–346 (1979)
6. Morozov, E., Delgado, R.: Stability analysis of regenerative queueing systems. Autom. Remote Control **70**(12), 1977–1991 (2009)
7. Müller, A., Stoyan, D.: Comparison Methods for Stochastic Models and Risks. Wiley, Hoboken (2002)
8. Rumyantsev, A., Morozov, E.: Stability criterion of a multiserver model with simultaneous service. Ann. Oper. Res. **252**(1), 29–39 (2017)
9. Serfozo, R.: Introduction to Stochastic Networks. Springer, New York (1999)
10. Wolff, R.W.: Work-conserving priorities. J. Appl. Probab. **7**(2), 327–337 (1970)
11. Wolff, R.W.: Poisson arrivals see time averages. Oper. Res. **30**(2), 223–231 (1982)