Queueing Systems 38, 149–184, 2001

Some Results for Large Closed Queueing Networks with and without Bottleneck: Up- and Down-Crossings Approach

VYACHESLAV M. ABRAMOV vyachesl@inter.net.il; slava@orikanetworks.com
24/6 Balfour St., Petach Tiqwa 49350, Israel

Received 20 January 2000; Revised 14 December 2000

Abstract. The paper provides the up- and down-crossing method to study the asymptotic behavior of queue-length and waiting time in closed Jackson-type queueing networks. These queueing networks consist of central node (hub) and k single-server satellite stations. The case of infinite server hub with exponentially distributed service times is considered in the first section to demonstrate the up- and down-crossing approach to such kind of problems and help to understand the readers the main idea of the method. The main results of the paper are related to the case of single-server hub with generally distributed service times depending on queue-length. Assuming that the first $k-1$ satellite nodes operate in light usage regime, we consider three cases concerning the kth satellite node. They are the light usage regime and limiting cases for the moderate usage regime and heavy usage regime. The results related to light usage regime show that, as the number of customers in network increases to infinity, the network is decomposed to independent single-server queueing systems. In the limiting cases of moderate usage regime, the diffusion approximations of queue-length and waiting time processes are obtained. In the case of heavy usage regime it is shown that the joint limiting non-stationary queue-lengths distribution at the first $k-1$ satellite nodes is represented in the product form and coincides with the product of stationary GI/M/1 queue-length distributions with parameters depending on time.

Keywords: closed queueing network, martingales and semimartingales, up- and down-crossings, coupling, diffusion and fluid approximation, bottleneck

1. Introduction

1.1. *The history of subject*

Speaking about the history of subject we mean the most closely related papers. Therefore in this subsection we discuss the following two papers: Kogan and Liptser [31], and Abramov [4]. Both those papers study the large closed queueing networks with bottleneck. Other papers concerning the subject will be mentioned in one of the next subsections.

Let us consider closed Jackson-type Markovian queueing network with a single class (chain) consisting of N customers, a single infinite server queue which will be called *central* node or *hub*, numbered by 0 and k single-server satellite nodes numbered by $1, 2, \dots, k$. Suppose that at the initial time moment all customers are at the hub. After service completion at the hub a customer visits satellite node j with probability p_j ($\sum_{j=1}^{k} p_j = 1$). Customers served at a single-server node return to the hub. Assume that a service time at the hub has the expectation λ^{-1} for each customer, and the expectation of a service time of the jth satellite node is equal to $(N\mu_j)^{-1}$.

Let $\{Q_1(t), Q_2(t), \dots, Q_k(t)\}$ denote the vector of queue-lengths at time t at satellite nodes $1, 2, \dots, k$, and $Q_j(0) = 0$, $1 \leqslant j \leqslant k$, according to convention. Queue-lengths $Q_j(t)$, $j = 0, 1, \dots, k$, depend on N. If it is necessary we will point out this dependence: $Q_j(t) = Q_{j,N}(t)$. Sometimes, where it is clear from context $Q_j(t)$ without index N denotes the limiting as $N \to \infty$ queue-length process at the jth satellite node. In those cases it is assumed that all queue-length processes indexed by N are given on the same probability space.

We say that satellite node j is a *bottleneck node*, if $\mu_j \leqslant \lambda p_j$. Otherwise, that node j is called non-bottleneck node. As N increases infinitely, Kogan and Liptser [31] showed that the non-stationary queue-length distribution for every non-bottleneck node is the same as some stationary M/M/1 queue-length distribution whose parameters explicitly depend on t, i.e., the queue-length process is described by the Yule process (see Bharucha-Reid [7]). In a case where there is no bottleneck node in network, the non-stationary queue-length processes for every node is described by geometrical distribution for all $t > 0$.

More accurately, denote $\lambda_j = \lambda p_j$, $1 \leqslant j \leqslant k$, and suppose that the first $k-1$ nodes are non-bottleneck nodes while the kth node is the bottleneck node. It was shown that the distribution of the queue-length $Q_j(t)$, $j = 1, 2, \dots, k-1$, as $N \to \infty$, for any fixed $t > 0$ tends to the stationary distribution of the M/M/1 queue-length process with constant arrival rate $\lambda_j[\pi_k + (1-\pi_k)\exp(-\lambda_k t)]$, where $\pi_k = \lambda_k^{-1}\mu_k$ and service rate μ_j. This fact has the following simple intuitive explanation. As $N \to \infty$, the queue-length at the kth satellite node increases to infinity being equivalent to $N[(1-\pi_k)(1-\exp(-\lambda_k t)]$, and therefore the queue-length at the hub increasing to infinity too as $N \to \infty$ is equivalent to $N[\pi_k + (1-\pi_k)\exp(-\lambda_k t)]$. As $t \to \infty$ as well, the limiting stationary distribution of the queue-length at the jth satellite node, $j = 1, 2, \dots, k-1$, coincides with the stationary distribution of the M/M/1 queue-length process with the arrival rate $\lambda_j \pi_k$ and the service rate μ_j.

In the case where $\mu_j > \lambda_j$ for all $j = 1, 2, \dots, k$, as $N \to \infty$, the limiting queue-length distribution at satellite node j tends to the stationary distribution of the M/M/1 queue-length process with the arrival rate λ_j and the service rate μ_j for any fixed $t > 0$ (see Whitt [57]).

The sequal paper by Abramov [4] developed the paper by Kogan and Liptser [31] assuming that the service mechanism of k satellite stations is autonomous. That is every satellite server j serves customers only in random instants $\xi_{j,1}^N, \xi_{j,1}^N + \xi_{j,2}^N, \dots$, where the sequences $\{\xi_{j,1}^N, \xi_{j,2}^N, \dots\}$ is assumed to consist of generally distributed random variables and be strictly stationary and ergodic sequences of random variables for each j. Under these assumption the author studied the qualitative behavior of the queue-length processes at satellite nodes, taking into account the effect of bottleneck node rather than obtained results in explicit form giving us something in terms of habital for us transformations, Laplace–Stieltjes transform say.

Motivating the current investigation in the next subsection we explain why the models from the earlier abovementioned papers would be developed from another position?

1.2. Motivation and description of the model

The case of queueing network considered in the papers of Kogan and Liptser [31] and Abramov [4] slightly idealizes the computer models arising in practice. Basing on client/server star topology, the models from the both abovementioned papers consider the case where the hub is an infinite server queueing system. However this case cannot precisely approximate the variety of real queueing networks. Therefore in the present paper we suggest a more general model by allowing service time distributions at the hub to depend on the number of customers residing there. A more exact description of the problem is given below.

Consider the Jackson type networks with the same mechanism as it was described above for the network considered in the both abovementioned papers. A service time of the hub is assumed to be a generally distributed random variable while the service times of other nodes are exponentially distributed. It is assumed that the hub is a single-server station a service time which depends on a queue-length as follows: if immediately before a service of sequential customer the queue-length at this node is equal to $K \leqslant N$, the probability distribution function is $G_K(Kx)$, $g_K^{-1} = \int_0^\infty x \, \mathrm{d}G_K(x) < \infty$, and, as $N \to \infty$, the sequence of probability distributions $G_N(x)$ converges weakly to $G(x)$ with $g^{-1} = \int_0^\infty x \, \mathrm{d}G(x) < \infty$ and $G(0+) = 0$.

Along with these assumptions in the further account we shall require in addition the stochastic order relations between two neighbor distribution functions $G_K(Kx)$ and $G_{K+1}(Kx + x)$: we shall assume that $G_K(Kx) \leqslant G_{K+1}(Kx + x)$ for all $x \geqslant 0$. The sense of this order relation is intuitively clear: *a rate of service time at the hub increases, as a queue-length increases there*. Note that this assumption is automatically implied in the special case of $G_1(x) = G_2(x) = \cdots = G_N(x) = G(x) = 1 - \mathrm{e}^{-\lambda x}$, leading to network considered by Kogan and Liptser [31]. In this case the expected service time at the hub linearly depends on its queue-length.

We consider a network, k satellite nodes which are assumed to be the single-server stations with the service rates $\mu_j N$, $j = 1, 2, \ldots, k$. At the initial time moment $t = 0$ all customers are at the hub, and one of them begins to be served there. As in the case of Markovian queueing network it is assumed that after a service completion at the hub a customer visits satellite node j with probability p_j ($\sum_{j=1}^{k} p_j = 1$), and customers, served at a single-server node, return to the hub. Denoting $\gamma_j = gp_j$, it is assumed for this network that $g_N p_j < \mu_j$, and, as $N \to \infty$, in the limiting case $\gamma_j < \mu_j$ for $j = 1, 2, \ldots, k - 1$. For the satellite node k of network we shall consider three cases:

(i) $g_N p_k < \mu_k$ and $\gamma_k = g p_k < \mu_k$;

(ii) $|\mu_k - g_N p_k| = \Delta_N > 0$ and $\Delta_N \to 0$ as $N \to \infty$;

(iii) $g_N p_k > \mu_k$ and $\gamma_k = g p_k > \mu_k$.

The cases (i) and (iii) are called *asymptotically light usage regime* and *asymptotically heavy usage regime*, respectively, for satellite node k, and the case $\gamma_k = \mu_k$ is called *asymptotically moderate usage regime* for that node (see, e.g., Kogan and Liptser [31]). For the sake of simplicity the word "asymptotically" will be omitted in such phrases.

1.3. Methodology of the paper

It is clear that the network considered here is a more general variant of that considered by Kogan and Liptser [31]. Kogan and Liptser [31] studied the queue-length processes for satellite nodes with the aid of traditional "real time" analysis, by using the modern methods of the theory of martingales in continuous time and the Skorohod reflection principle.

The approach of the present paper is based on consideration of the sequences of super- and submartingales generated by up-crossing processes of these queueing networks and proving their convergence to a Galton–Watson branching process.

Thus in contrast to the mentioned paper by Kogan and Liptser [31] we study the similar queueing processes with the aid of discrete time analysis, where we equate a time with an up-crossing level. This enables us essentially simplify the analysis, avoid background technical details and obtain some of results in more general form. Namely, the method permits us to obtain the results both for light and heavy usage regimes in a form of the joint queue-length distributions whereas Kogan and Liptser [31] obtained only the marginal (one-dimensional) distributions. For light usage regime it is not so considerable because intuitively clear that the limiting queue-lengths at satellite nodes are mutually independent. However in the case of heavy usage regime it looks considerable because a behavior of the queue-lengths in non-bottleneck nodes should be correlated with the queue-length in the bottleneck node. The joint queue-length distribution obtained in the paper in product form enables us to conclude that the limiting queue-lengths in non-bottleneck nodes are independent nevertheless.

In all likelihood, the first paper using the up- and down-crossing approach to the waiting time problems of GI/GI/1 queueing system was due to Cohen [14]. An idea of the method of the present paper with different variations was earlier used in a number of works by the author devoted to the simplest Markovian queueing systems and Markovian epidemic models (see Abramov [1,2] as well as the second section of this paper). For other papers related to the up- and down-crossings approach see also Abramov [5], Brill and Posner [10,11], Shanthikumar [48,49], Shanthikumar and Chandra [50].

The traditional idea of the up- and down-crossings approach of those papers is that during regeneration period the number of up-crossings of a level l equals the number of down-crossings of that level. In many cases using of very simple idea of up- and down-crossings is much profitable than analytical techniques.

Note, that another idea of the up- and down-crossings approach is used in Abramov [3,4]. The approach of those papers uses a ‘real time’ analysis, that is the numbers of up- and down-crossings are studied during a ‘real time’ t. The reason is that the processes studied in those papers are not regenerative, and therefore the well understood property is that the number of up-crossings of the level l equals the number of down-crossings of the same level is not valid.

Along with up- and down-crossings approach the paper uses diffusion and fluid approximations. The approach of those approximations is based on sample paths analysis and coupling of stochastic sequences permitting us to deal with the sample paths of random variables and processes given on the same probability space rather than probability distributions. This enables us to reduce the problem to the well-known classic cases applying then diffusion approximations. Amongst a large set of known approachs to diffusion approximations (see Borovkov [8], Prohorov [44], Iglehart and Whitt [18,19], Liptser and Shiryayev [38], Reiman [45], Lemoine [36], etc.) this paper uses an approach of Borovkov [8] being a development of the earlier results by Prohorov [44]. Under the heavy usage regime the paper uses fluid approximation taking into consideration the fact that the GI/M/1 queue-length process can be represented as the process with autonomous service mechanism. This property is evident because of the property of the lack of memory of exponentially distributed service times.

The results related to diffusion approximations for some other network with the same central node and one station with generally distributed service times have been obtained by Krichagina et al. [35]. For other relevant applications of a bottleneck and fluid and diffusion approximations see Knessl and Tier [28] Kogan [29], Kogan and Birman [30], Kogan et al. [32,33], Krichagina [34], McKenna et al. [42], Abramov [4], Chen and Mandelbaum [12,13], Reiman and Simon [46], Pittel [43], Mandelbaum and Pats [41], Mandelbaum and Massey [39], Mandelbaum et al. [40] and others.

1.4. The structure of the paper

The paper is structured as follows. It consists of six sections. The first section is introduction. The second section can be considered as auxiliary for a more clear understanding of the main idea of the method. It considers the Markovian queueing network, investigated by Kogan and Liptser [31] and Whitt [57] under assumption $\gamma_j < \mu_j$ for all $j = 1, 2, \ldots, k$. The main reason of the presence of this section is that the up- and down-crossings method for Markovian models of queues does not contain cumbersome technical details and simpler than that for non-Markovian models of queues and networks. Moreover, the section contains complete notation, definitions and proof, while the different sections are possibly shortened.

The third section considers a network, the service times of hub which are generally distributed random variables as was described above. Under the same assumption $\gamma_j < \mu_j$, $j = 1, 2, \ldots, k$, it focuses on the proof of ergodic theorem, showing that, as $N \to \infty$, the joint limiting queue-length state probabilities are calculated as the product of those probabilities for the stationary GI/M/1 queueing systems.

The fourth section considers the diffusion approximations for queue-length and waiting time distributions under the heavy traffic conditions.

In the fifth section the paper studies the queue-length distribution at the first $k-1$ satellite nodes where the kth satellite node operates in heavy usage regime.

In the last section a number of generalizations of results, where k satellite nodes are many server stations, are discussed.

2. A Markovian queueing network

According to convention at the initial time moment all customers are at the hub. Assume that $\lambda_j < \mu_j$, $j = 1, 2, \dots, k$, and consider one of k satellite nodes. Let it be first.

Suppose that the first customer arrives to this node at time t_0. Let T_1 denote the *first busy period* for the first satellite node, i.e., the time interval from the moment t_0 till the moment when after a service completion of a customer the node becomes empty at the first time after t_0.

Let $f_1(n)$ denote the number of up-crossings of the level n during this busy period, i.e., the number of situations during that period where arriving customer meets n other customers in the node ($f_1(0) = 1$ with probability 1). Denote $t_{n,1}, t_{n,2}, \dots, t_{n,f_1(n)}$ those time moments in ascending order, and let $s_{n,1}, s_{n,2}, \dots, s_{n,f_1(n)}$ be the correspondent time moments where after service completion there are n customers in the node. Thus we have the following time intervals:

$$(t_{n,1}, s_{n,1}),\ (t_{n,2}, s_{n,2}),\ \dots,\ (t_{n,f_1(n)}, s_{n,f_1(n)}). \tag{2.1}$$

Let us consider the similar moments $t_{n+1,1}, t_{n+1,2}, \dots, t_{n+1,f_1(n+1)}, s_{n+1,1}, s_{n+1,2}, \dots, s_{n+1,f_1(n+1)}$ of the level $n+1$ and correspondent intervals:

$$[t_{n+1,1}, s_{n+1,1}),\ [t_{n+1,2}, s_{n+1,2}),\ \dots,\ [t_{n+1,f_1(n+1)}, s_{n+1,f_1(n+1)}). \tag{2.2}$$

It is obvious that the intervals (2.2) are contained in the intervals (2.1). Let us delete the intervals (2.2) from intervals (2.1) and connect the ends, that is the point $t_{n+1,1}$ with $s_{n+1,1}$, the point $t_{n+1,2}$ with $s_{n+1,2}$, etc.

Note, that in mathematical terms the given procedure is explaned as follows. Let $\mathcal{J}_1$ be the set of intervals (2.1) and $\mathcal{J}_2$ be in turn the set of intervals (2.2). Then the set obtained after the procedure described above is $\mathcal{J}_1/\mathcal{J}_2$, which is the factor-set of set $\mathcal{J}_1$ on set $\mathcal{J}_2$.

It is also significant to note here that all queueing systems (with different N) are defined on the same probability space. This allows us to use sample paths techniques such as comparison of the processes and taking a limit as the series parameter N increases to infinity.

Next, let us denote the maximum and minimum numbers of customers in the hub within defined thus residual time intervals by N_n^+ and N_n^-, respectively. It is obvious that

$$N_n^+ \leqslant N. \tag{2.3}$$

with probability 1. Let us show that

$$\mathbb{E}\big\{f_1(n+1) \mid f_1(n)\big\} \leqslant f_1(n)\frac{\lambda_1}{\mu_1}. \tag{2.4}$$

Indeed, consider first a standard M/M/1 queueing system with the input rate $\lambda_1 N$ and the mean service time $(\mu_1 N)^{-1}$ and assume that the intervals (2.1) and (2.2) are defined now for that queueing system. Remaining in force all of the suggested earlier notation, because of the lack of memory property of exponential distribution $\mathbb{E}f_1(1) = \mathbb{E}\{f_1(n+1) \mid f_1(n) = 1\}$ and therefore, since all intervals of (2.1) are independent and identically distributed as a busy period (provided that they do exist), $\mathbb{E}\{f_1(n+1) \mid f_1(n)\} = f_1(n)\mathbb{E}f_1(1)$. Thus $\{f_1(n)\}$ is a Galton–Watson branching process. In turn,

$$\mathbb{E}f_1(1) = \sum_{i=1}^{\infty} i \int_0^{\infty} \mathrm{e}^{-\lambda_1 tN} \frac{(\lambda_1 tN)^i}{i!}\, \mu_1 N\, \mathrm{e}^{-\mu_1 Nt}\, \mathrm{d}t = \frac{\lambda_1}{\mu_1}.$$

For more detailed explanation of this case see Abramov [1]. Coming back to the first satellite node of the network let us take into account that the length of the residual intervals after deleting of the intervals (2.2) from the intervals (2.1) and connecting the ends has the same distribution as in a case of standard M/M/1 queueing system, more accurately, it coincides with the distribution of service time, while interarrival times being depending on the number of customers in the hub and having the parameter non-greater than $\lambda_1 N$, are stochastically non-smaller than in a case of a standard M/M/1 queueing system. Therefore,

$$\mathbb{E}\big\{f_1(n+1) \mid f_1(n)\big\} \leqslant f_1(n)\,\frac{\lambda_1}{\mu_1}\,\frac{N_n^+}{N} \leqslant f_1(n)\frac{\lambda_1}{\mu_1} \tag{2.5}$$

and (2.4) follows. Note that (2.4) easily follows also from the theorem on stochastic monotonicity of Markov processes (see, e.g., Kalmykov [22], Keilson [26])

Denote $\mathcal{F}_{1,n} = \sigma\{f_1(0), f_1(1), \ldots, f_1(n)\}$. By (2.4) it is clear that the stochastic sequence

$$\left\{\frac{\mu_1^n}{\lambda_1^n}\, f_1(n), \mathcal{F}_{1,n}\right\} \tag{2.6}$$

is a supermartingale.

From (2.4)

$$\sum_{i=0}^{N} \mathbb{E}f_1(i) = \sum_{i=0}^{N} \mathbb{E}\big\{f_1(i) \mid f_1(0)\big\} \leqslant \sum_{i=0}^{\infty} \frac{\lambda_1^i}{\mu_1^i} = \frac{\mu_1}{\mu_1 - \lambda_1} < \infty. \tag{2.7}$$

Hence, according to Wald's identity, from (2.7) for the expected busy period we obtain:

$$\mathbb{E}T_1 = (N\mu_1)^{-1} \sum_{i=0}^{N} \mathbb{E}f_1(i) \leqslant \big[N(\mu_1 - \lambda_1)\big]^{-1}, \tag{2.8}$$

and as $N \to \infty$

$$\lim_{N\to\infty} N\mathbb{E}T_1 \leqslant (\mu_1 - \lambda_1)^{-1}, \tag{2.9}$$

where $\mathbb{E}T_1$ in (2.9) is considered as a function of N.

Next, analogously to (2.4) we obtain:

$$\mathbb{E}\big\{f_1(n+1) \mid f_1(n)\big\} \geqslant f_1(n)\frac{\lambda_1 N_n^-}{\mu_1 N}. \tag{2.10}$$

Note, that

$$N_n^- \geqslant N - \sum_{i=0}^{N} f_1(i) - v(T_1 + t_0) = c(N),$$

where $v(T_1 + t_0)$ is the mutual number of served customers in all other satellite nodes during a period $(0, T_1 + t_0)$ which, as $N \to \infty$, will be called *asymptotic regeneration period* of the first satellite node. Hence, it follows from (2.10) that

$$\mathbb{E}\big\{f_1(n+1) \mid f_1(n)\big\} \geqslant f_1(n)\frac{\lambda_1 N_n^-}{\mu_1 N} \geqslant f_1(n)\frac{\lambda_1 c(N)}{\mu_1 N}, \tag{2.11}$$

and the stochastic sequence

$$\left\{\left[\frac{\mu_1 N}{\lambda_1 c(N)}\right]^n f_1(n), \mathcal{F}_{1,n}\right\} \tag{2.12}$$

is a submartingale. It follows from (2.8) that $\mathbb{E}v(T_1 + t_0) < \infty$, and therefore

$$\mathbb{P}\left\{\lim_{N\to\infty}\frac{c(N)}{N} = \lim_{N\to\infty}\frac{N_n^-}{N} = \lim_{N\to\infty}\frac{N_n^+}{N} = 1\right\} = 1. \tag{2.13}$$

Next, let $\{f_1^*(n), \mathcal{F}_{1,n}^*\}$ is the limiting as $N \to \infty$ stochastic sequence, $\mathcal{F}_{1,n}^* = \sigma\{f_1^*(0), f_1^*(1), \ldots, f_1^*(n)\}$.

In view of (2.5), (2.12) and (2.13)

$$\mathbb{E}\big\{f_1^*(n+1) \mid f_1^*(n)\big\} \leqslant f_1^*(n)\,\frac{\lambda_1}{\mu_1}$$

and since the limiting stochastic sequence $\{f_1^*(n), \mathcal{F}_{1,n}^*\}$ is Markovian, the limiting stochastic sequence

$$\left\{\frac{\mu_1^n}{\lambda_1^n}\, f_1^*(n), \mathcal{F}_{1,n}^*\right\} \tag{2.14}$$

forms a supermartingale. On the other hand, in view of (2.10) and (2.13),

$$\mathbb{E}\big\{f_1^*(n+1) \mid f_1^*(n)\big\} \geqslant f_1^*(n)\,\frac{\lambda_1}{\mu_1}$$

and the limiting stochastic sequence (2.14) is a submartingale. Thus, we proved that the stochastic sequence (2.14) is both a super- and submartingale simultaneously, and therefore it is a martingale. From this, coming back to intervals (2.1) we can see that in the limiting case the number of inserted points after the procedure described above has the same distribution for all intervals. This allows us to claim that the limiting process is branching.

Now, let T_j denote the first busy period at satellite node j, and $f_j(n)$ be the number of up-crossings during a busy period T_j at this node. It is obvious that the statement proved above for the first satellite node is also true for all other satellite nodes, i.e., denoting the limiting as $N \to \infty$ process $f_j^*(n)$, $\mathcal{F}_{j,n}^* = \sigma\{f_j^*(0), f_j^*(1), \dots, f_j^*(n)\}$ for jth node we can state that the stochastic sequence $\{f_j^*(n), \mathcal{F}_{j,n}^*\}$ is a Galton–Watson branching process. It is also significant to note that for $i \neq j$ the limiting branching processes $\{f_i^*(n), \mathcal{F}_{i,n}^*\}$ and $\{f_j^*(n), \mathcal{F}_{j,n}^*\}$ are independent on one another although under N given the processes $\{f_i(n), \mathcal{F}_{i,n}\}$ and $\{f_j(n), \mathcal{F}_{j,n}\}$ are dependent.

Next, let us consider the first *generalized busy period*, the time interval from the first arrival of a customer at one of k satellite nodes till the moment when all these nodes are at the first time simultaneously empty. Analogously, under *generalized idle time* we mean the time interval where all k satellite nodes are simulteneously empty. It is clear that the first *regeneration period* (generalized busy period + generalized idle time) contains the random number of only complete busy periods for each of k satellite nodes.

Moreover, since the expected numbers of served customers during the first busy periods are finite for all satellite nodes and the network is stable (see, e.g., Borovkov [9], Kaspi and Mandelbaum [24,25]; see also Abramov [4] where the proof of stability is also appropriate here because the given queue-length process can be also considered as the process with autonomous service mechanism), according to reward renewal theorem (see Karlin and Taylor [23]) the expected number of served customers during a generalized busy period is also finite, and therefore the expectation of generalized busy period has the order $\mathrm{O}(N^{-1})$. Therefore, considering the multitype up-crossing process $\{\widetilde{\mathbf{f}}(\mathbf{n}), \widetilde{\mathcal{G}}_{\mathbf{n}}\}$, where $\widetilde{\mathbf{f}}(\mathbf{n}) = \{\widetilde{f}_j(n_j), j = 1, 2, \dots, k\}$, $\widetilde{f}_j(n_j)$ denotes the number of up-crossings of the level n_j at jth node during the first generalized busy period, $\widetilde{\mathcal{G}}_{\mathbf{n}} = \sigma\{\widetilde{\mathbf{f}}(\mathbf{k})\colon \mathbf{0} \leqslant \mathbf{k} \leqslant \mathbf{n}\}$, $\mathbf{0}$ is the k-dimensional vector of zeros, $\widetilde{\mathbf{f}}(\mathbf{0}) = 1$ with probability 1, and $\mathbf{k} \leqslant \mathbf{n}$ denotes that all components of vector $\mathbf{k}$ are non-greater than all those of vector $\mathbf{n}$, one can conclude that the limiting as $N \to \infty$ multitype process $\{\mathbf{f}^*(\mathbf{n}), \mathcal{G}_{\mathbf{n}}^*\}$, where $\mathcal{G}_{\mathbf{n}}^* = \sigma\{\mathbf{f}^*(\mathbf{k})\colon \mathbf{0} \leqslant \mathbf{k} \leqslant \mathbf{n}\}$, is a multitype Galton–Watson branching process, generated by k independent one-dimensional branching processes. In the given context under multitype Galton–Watson branching process we mean such vector-valued process with vector-valued time argument as it is defined above.

Denote the limiting as $N \to \infty$ multitype up-crossing process $\{\widetilde{\mathbf{f}}^*(\mathbf{n}), \widetilde{\mathcal{G}}_{\mathbf{n}}^*\}$, where $\widetilde{\mathbf{f}}^*(\mathbf{n}) = \{\widetilde{f}_j^*(n_j),\ j = 1, 2, \dots, k\}$. Then we have the following representation:

$$\widetilde{f}_j^*(n_j) = \sum_{l=1}^{\kappa_j} f_{j,l}^*(n_j),$$

where κ_j is the (limiting) number of busy periods in the jth node during the (limiting) generalized busy period, and $\{f_{j,i}^*(n_j)\}_{i\geqslant 1}$ is the sequence of independent identically distributed random variables having the same distribution as the random variable $f_j^*(n_j)$, $n_j \geqslant 0$.

In the further account all limiting processes and their σ-algebras will be provided with asterisk. For example, $\mathbf{f}^*(\mathbf{n})$ denotes the limiting multitype branching process; $\widetilde{f}_j^*(n_j)$ denotes the limiting process of $\widetilde{f}_j(n_j)$ as $N \to \infty$, i.e., the limiting one-dimensional up-crossing process.

Theorem 2.1. For every $t > 0$, the limiting as $N \to \infty$ joint queue-length distribution is

$$\mathbb{P}\bigl\{Q_1(t) = n_1, Q_2(t) = n_2, \ldots, Q_k(t) = n_k\bigr\} = \prod_{j=1}^{k} \rho_j^{n_j}(1-\rho_j), \qquad \rho_j = \frac{\lambda_j}{\mu_j}.$$

Proof. Let us apply the reward renewal theorem. Let $\mathbf{T}(\mathbf{n})$ denote the cumulative time in state $\mathbf{n} = (n_1, n_2, \ldots, n_k)$ during generalized busy period. Keeping in mind the Wald's identity we obtain:

$$\begin{aligned} \mathbb{P}\bigl\{Q_1(t) = n_1, Q_2(t) = n_2, \ldots, Q_k(t) = n_k\bigr\} &= \lim_{N\to\infty} \frac{\mathbb{E}\mathbf{T}(\mathbf{n})}{\sum_{\mathbf{k}} \mathbb{E}\mathbf{T}(\mathbf{k})} \\ &= \frac{\mathbb{E}\widetilde{\mathbf{f}}^*(\mathbf{n})}{\sum_{\mathbf{k}} \mathbb{E}\widetilde{\mathbf{f}}^*(\mathbf{k})}. \end{aligned} \tag{2.15}$$

Because of the independency of the limiting up-crossing processes,

$$\mathbb{E}\widetilde{\mathbf{f}}^*(\mathbf{n}) = \prod_{j=1}^{k} \mathbb{E}\widetilde{f}_j^*(n_j), \tag{2.16}$$

and

$$\sum_{\mathbf{k}} \mathbb{E}\widetilde{\mathbf{f}}^*(\mathbf{k}) = \prod_{j=1}^{k} \left[\sum_{n=0}^{\infty} \mathbb{E}\widetilde{f}_j^*(n)\right]. \tag{2.17}$$

Therefore, substituting (2.16) and (2.17) for (2.15) we have

$$\mathbb{P}\bigl\{Q_1(t) = n_1, Q_2(t) = n_2, \ldots, Q_k(t) = n_k\bigr\} = \prod_{j=1}^{k} \frac{\mathbb{E}\widetilde{f}_j^*(n_j)}{\sum_{n=0}^{\infty} \mathbb{E}\widetilde{f}_j^*(n)}. \tag{2.18}$$

According to Wald's identity

$$\mathbb{E}\widetilde{f}_j^*(n_j) = \mathbb{E}f_j^*(n_j)\mathbb{E}\kappa_j.$$

Therefore (2.18) is rewritten as

$$\mathbb{P}\bigl\{Q_1(t) = n_1, Q_2(t) = n_2, \ldots, Q_k(t) = n_k\bigr\} = \prod_{j=1}^{k} \frac{\mathbb{E}f_j^*(n_j)}{\sum_{n=0}^{\infty} \mathbb{E}f_j^*(n)}. \tag{2.19}$$

Taking into consideration that the stochastic sequence (2.14) forms a martingale, we obtain

$$\mathbb{E} f_j^*(n_j) = \mathbb{E}\bigl\{ f_j^*(n_j) \mid f_j^*(0) \bigr\} = \rho^{n_j},$$

and hence the statement of theorem follows from (2.19). The theorem is proved. □

3. A queueing network with state dependent service times in the hub and k single server satellite stations: Light usage regime at the kth satellite node

In this section we consider a more general queueing network than in the previous section in that the service times at the hub are state dependent and not exponentially distributed. The description of this model is given in introduction. As in the previous section we shall assume that all queueing systems (with different N) are considered on the same probability space.

While interarrival times in each satellite node of the Markovian queueing network have exponential distribution, those interarrival time distributions of a queueing network under assumption of this section are very complicated.

Namely, let $\tau_1, \tau_2, \ldots$ be a sequence of service completions at the hub and $Q_0(\tau_1)$, $Q_0(\tau_2), \ldots$ be sequence of queue-lengths there. Then probability distribution function of the first interarrival time in satellite node j can be represented

$$H_{j,1}(x) = \sum_{i=0}^{\infty} p_j (1-p_j)^i \widetilde{G}_{r_0} * \widetilde{G}_{r_1} * \cdots * \widetilde{G}_{r_i}(x), \tag{3.1}$$

where $\widetilde{G}_{r_0}(x) = G_N(Nx)$, $\widetilde{G}_{r_l} = \sum_{u=1}^{\infty} \mathbb{P}\{Q_0(\tau_l) = u\} G_u(ux)$ $(l \geqslant 1)$, asterisk denotes a convolution. Therefore it cannot be expressed in explicit form. Denote the sequence of interarrival time distributions in satellite node j by $H_{j,1}(x)$, $H_{j,2}(x), \ldots$. Let us define first $H_{j,2}(x)$. The time moments $\tau_1, \tau_2, \ldots, \tau_i, \ldots$ can be the moments of first arrival of a customer at satellite node j with probabilities $p_j, p_j(1-p_j), \ldots, p_j(1-p_j)^{i-1}, \ldots$, respectively, therefore

$$H_{j,2}(x) = \sum_{l=1}^{\infty} \sum_{i=0}^{\infty} p_j^2 (1-p_j)^{i+l-1} \widetilde{G}_{r_l} * \widetilde{G}_{r_{l+1}} * \cdots * \widetilde{G}_{r_{l+i}}(x),$$

and recurrently

$$\begin{aligned} H_{j,\nu}(x) = {} & \sum_{k_1=1}^{\infty} \cdots \sum_{k_{\nu-1}=1}^{\infty} \sum_{i=0}^{\infty} p_j^{\nu} (1-p_j)^{i+k_1+\cdots+k_{\nu-1}-\nu+1} \\ & \times \bigl[\widetilde{G}_{r_{k_1+\cdots+k_{\nu-1}}}(x) * \widetilde{G}_{r_{k_1+\cdots+k_{\nu-1}+1}}(x) * \cdots * \widetilde{G}_{r_{k_1+\cdots+k_{\nu-1}+i}}(x) \bigr]. \end{aligned} \tag{3.2}$$

Let us take into account the assumption $G_K(Kx) \leqslant G_{K+1}(Kx + x)$ for all $K \geqslant 1$. Then, according to properties of this stochastic order relation (see, e.g., Stoyan [53]), from (3.2) one can note that for all $l = 0, 1, \dots, i$

$$\widetilde{G}_{r_{k_1+\cdots+k_{\nu-1}+l}}(x) \leqslant G_N(Nx),$$

and therefore

$$\widetilde{G}_{r_{k_1+\cdots+k_{\nu-1}}}(x) * \widetilde{G}_{r_{k_1+\cdots+k_{\nu-1}+1}}(x) * \cdots * \widetilde{G}_{r_{k_1+\cdots+k_{\nu-1}+i}}(x) \leqslant G_N^{*i+1}(Nx).$$

Thus,

$$H_{j,\nu}(x) \leqslant \sum_{i=0}^{\infty} p_j(1-p_j)^i G_N^{*i+1}(Nx) \tag{3.3}$$

for all $\nu \geqslant 1$ and $x \geqslant 0$, where $G_N^{*i+1}(Nx)$ is the $(i+1)$-fold convolution of the probability distribution $G_N(Nx)$ with itself. According to the formula for the total probability, inequality (3.3) is also valid with probability 1 when ν is a nonnegative integer random variable.

Remaining the notation of the previous section let us construct here the intervals (2.1) and (2.2) by the same manner as it has been done there. Let us take into account the following. Every service completion, $s_{n,r}$ say, occurs due to the following situations: immediately before the moment of preceding arrival there were $n+i$ customers at the first satellite node and $i+1$ customers have been served during interarrival time of the first satellite node. Let N_n^+ and N_n^- denote the maximum and minimum numbers of customers at the hub within residual intervals after described above procedure of deleting intervals and connecting the ends. It is clear that $N_n^+ \leqslant N$. Let us estimate $\mathbb{E}\{f_1(n+1) \mid f_1(n)\}$. Consider first the case of the standard GI/M/1 queueing system interarrival times which have the distribution function

$$\sum_{i=0}^{\infty} p_1(1-p_1)^i G_N^{*i+1}(Nx),$$

and service times have the expectation $(\mu_1 N)^{-1}$. As in the case of the M/M/1 queueing system considered earlier in the previous section we have $\mathbb{E}\{f_1(n+1) \mid f_1(n)\} = f_1(n)\mathbb{E}f_1(1)$, and the stochastic process $\{f_1(n)\}$ is a Galton–Watson branching process. Let us find $\mathbb{E}f_1(1)$. According to the formula for the total probability we have:

$$\begin{aligned}\mathbb{E}f_1(1) = z &= \sum_{l=0}^{\infty} z^l \int_0^{\infty} \mathrm{e}^{-\mu_1 Nx} \frac{(\mu_1 Nx)^l}{l!}\, \mathrm{d}\left[\sum_{i=0}^{\infty} p_1(1-p_1)^i G_N^{*i+1}(Nx)\right] \\ &= \sum_{l=0}^{\infty} \int_0^{\infty} \mathrm{e}^{-\mu_1 x} \frac{(\mu_1 zx)^l}{l!}\, \mathrm{d}\left[\sum_{i=0}^{\infty} p_1(1-p_1)^i G_N^{*i+1}(x)\right] \\ &= \frac{p_1 \widehat{G}_N(\mu_1 - \mu_1 z)}{1-(1-p_1)\widehat{G}_N(\mu_1 - \mu_1 z)},\end{aligned}$$

where

$$\widehat{G}_N(z) = \int_0^\infty \mathrm{e}^{-zx}\,\mathrm{d}G_N(x), \quad \Re(z) \geqslant 0.$$

Since the traffic intensity, being equal to $p_1 g_N \mu_1^{-1}$, is less than 1, then according to theorem by Takacs [54] (see also Gnedenko and Kovalenko [17], Klimov [27]) the functional equation

$$z = \frac{p_1 \widehat{G}_N(\mu_1 - \mu_1 z)}{1 - (1 - p_1)\widehat{G}_N(\mu_1 - \mu_1 z)}$$

has a unique solution belonging to interval $(0, 1)$. Therefore $\mathbb{E} f_1(1) < 1$.

Let us return to the first satellite node of network. Note first that

$$\mathbb{E}\{f_1(n+1) \mid f_1(n)\} \leqslant f_1(n)\mathbb{E} f_1(1). \tag{3.4}$$

This is because the service times have the same distribution as in the case of the GI/M/1 queueing system considered above, however interarrival times satisfy (3.3), and therefore the number of connected point within each interval after the procedure described above is stochastically not smaller than in the case of the GI/M/1 queueing system. A more detailed explanation for (3.4) can be provided by the way similar to the case of Markovian queueing network considered in the previous section. Alternative proof of (3.4) can be also provided with the aid of sample paths analysis of Markov processes and application of the theorem on stochastic monotonicity (see, e.g., Kalmykov [22], Keilson [26]).

Next, taking into account that an interarrival time at the first node depends on the queue-length at the hub, let us formally denote the queue-lengths at the hub, where immediately before the moment of preceding arrival there were $n + i$ customers at the first node, by $l_{n,i}(t_{n+i,u})$, where $t_{n+i,1}, t_{n+i,2}, \ldots, t_{n+i,f_1(n+i)}$ denote the respective arrival instants at the first node (see notation of the previous section). Denote $z_n = \mathbb{E}\{f_1(n+1) \mid f_1(n) = 1\}$. Then according to the formula for the total probability one can write:

$$z_0 = \sum_{i=0}^{\infty} \mathbb{E}\left\{\sum_{v=1}^{f_1(i)} \int_0^\infty \mathrm{e}^{-\mu_1 N x} \frac{(\mu_1 N x)^i}{i!}\,\mathrm{d}H_{1,\nu_{v,i}}(x)\right\}, \tag{3.5}$$

where $\nu_{v,i}$ is the some random number of interarrival time of the first satellite node depending on v and i (the empty sum is assumed to be 0). By (3.3) with probability 1

$$\begin{aligned}\int_0^\infty &\mathrm{e}^{-\mu_1 N x} \frac{(\mu_1 N x)^i}{i!}\,\mathrm{d}H_{1,\nu_{v,i}}(x)\\ &\leqslant \sum_{l=0}^{\infty} p_1(1-p_1)^l \int_0^\infty \mathrm{e}^{-\mu_1 N x} \frac{(\mu_1 N x)^i}{i!}\,\mathrm{d}G_N^{*l+1}(Nx),\end{aligned} \tag{3.6}$$

and therefore from (3.5)

$$z_0 \leqslant \sum_{l=0}^{\infty} p_1(1-p_1)^l \sum_{i=0}^{\infty} \mathbb{E}\left\{\sum_{v=1}^{f_1(i)} \int_0^{\infty} \mathrm{e}^{-\mu_1 N x} \frac{(\mu_1 N x)^i}{i!} \, \mathrm{d}G_N^{*l+1}(Nx)\right\}. \tag{3.7}$$

According to Wald's identity,

$$\begin{aligned} &\mathbb{E}\left\{\sum_{v=1}^{f_1(i)} \int_0^{\infty} \mathrm{e}^{-\mu_1 N x} \frac{(\mu_1 N x)^i}{i!} \, \mathrm{d}G_N^{*l+1}(Nx)\right\} \\ &\quad = \mathbb{E} f_1(i) \int_0^{\infty} \mathrm{e}^{-\mu_1 N x} \frac{(\mu_1 N x)^i}{i!} \, \mathrm{d}G_N^{*l+1}(Nx), \end{aligned}$$

and therefore, substituting it for (3.7), we obtain:

$$z_0 \leqslant \sum_{l=0}^{\infty} p_1(1-p_1)^l \sum_{i=0}^{\infty} \mathbb{E} f_1(i) \int_0^{\infty} \mathrm{e}^{-\mu_1 N x} \frac{(\mu_1 N x)^i}{i!} \, \mathrm{d}G_N^{*l+1}(Nx). \tag{3.8}$$

Next, according to (3.4)

$$z_n \leqslant \mathbb{E} f_1(1) = z_0,$$

and therefore, taking into account that

$$\mathbb{E} f_1(i) = \prod_{n=0}^{i-1} z_n$$

we obtain

$$\mathbb{E} f_1(i) \leqslant z_0^i.$$

Thus from (3.8) we obtain

$$\begin{aligned} z_0 &\leqslant \sum_{l=0}^{\infty} p_1(1-p_1)^l \sum_{i=0}^{\infty} z_0^i \int_0^{\infty} \mathrm{e}^{-\mu_1 N x} \frac{(\mu_1 N x)^i}{i!} \, \mathrm{d}G_N^{*l+1}(Nx) \\ &= \sum_{l=0}^{\infty} p_1(1-p_1)^l \int_0^{\infty} \exp(-\mu_1 N x + \mu_1 N z_0 x) \, \mathrm{d}G_N^{*l+1}(Nx). \end{aligned}$$

According to theorem by Takacs [54], there is the least root of equation

$$z = \sum_{l=0}^{\infty} p_1(1-p_1)^l \int_0^{\infty} \exp(-\mu_1 x + \mu_1 x z) \, \mathrm{d}G_N^{*l+1}(x), \tag{3.9}$$

with respect to z in interval (0, 1], and if

$$-\mu_1 \frac{\mathrm{d}}{\mathrm{d}z} \sum_{l=0}^{\infty} p_1(1-p_1)^l \int_0^{\infty} \mathrm{e}^{-zx} \, \mathrm{d}G_N^{*l+1}(x) \bigg|_{z=0} > 1, \tag{3.10}$$

this root is less than 1.

Denoting the least root of equation (3.9) by $\varphi_{1,N}$ we obtain:

$$\mathbb{E}\big\{f_1(n+1) \mid f_1(n)=1\big\} \leqslant \varphi_{1,N},$$

and therefore

$$\mathbb{E}\big\{f_1(n+1) \mid f_1(n)\big\} \leqslant \varphi_{1,N} f_1(n). \tag{3.11}$$

Next, denoting $\mathcal{F}_{1,n} = \sigma\{f_1(0), f_1(1), \ldots, f_1(n)\}$, by (3.11) one can see that the stochastic sequence

$$\big\{f_1(n)\varphi_{1,N}^{-n}, \mathcal{F}_{1,n}\big\} \tag{3.12}$$

forms a supermartingale.

By (3.12)

$$\sum_{i=0}^{\infty} \mathbb{E} f_1(i) = \sum_{i=0}^{\infty} \mathbb{E}\big\{f_1(i) \mid f_1(0)\big\} \leqslant 1 + \sum_{i=1}^{\infty} \varphi_{1,N}^{i} = \frac{1}{1-\varphi_{1,N}}. \tag{3.13}$$

Hence, according to Wald's identity, for the expected busy period from (3.13) we obtain:

$$\mathbb{E}T_1 = \frac{1}{\mu_1 N} \sum_{i=0}^{\infty} \mathbb{E} f_1(i) \leqslant \frac{1}{N\mu_1(1-\varphi_{1,N})}, \tag{3.14}$$

and, as $N \to \infty$,

$$\lim_{N\to\infty} N\mathbb{E}T_1 \leqslant \frac{1}{\mu_1(1-\varphi_1)}, \tag{3.15}$$

where $\mathbb{E}T_1$ in (3.15) is considered as a function of N, and $\varphi_1 = \lim_{N\to\infty} \varphi_{1,N}$. It is shown below that $0 < \varphi_1 < 1$.

Indeed, let φ_1 is the least root of functional equation

$$z = \sum_{l=0}^{\infty} p_1(1-p_1)^l \int_0^{\infty} \exp(-\mu_1 x + \mu_1 x z)\, \mathrm{d}G^{*l+1}(x), \tag{3.16}$$

being the limit of sequence $\{\varphi_{1,N}\}$, as $N \to \infty$. Denoting the Laplace–Stieltjes transform of $G(x)$ by $\widehat{G}(z)$, one can represent (3.16) in the form:

$$z = \frac{p_1 \widehat{G}(\mu_1 - \mu_1 z)}{1-(1-p_1)\widehat{G}(\mu_1 - \mu_1 z)}. \tag{3.17}$$

Note, that

$$-\mu_1 \frac{\mathrm{d}}{\mathrm{d}z} \frac{p_1 \widehat{G}(z)}{1-(1-p_1)\widehat{G}(z)}\bigg|_{z=0} = -\frac{\mu_1}{p_1}\widehat{G}'(0) = \frac{\mu_1}{p_1 g} = \frac{\mu_1}{\gamma_1} > 1$$

and therefore $0 < \varphi_1 < 1$ according to theorem by Takacs [54] (see also relations (3.9), (3.10)).

Next, analogously to (3.6) with probability 1

$$\begin{aligned}&\int_0^\infty \mathrm{e}^{-\mu_1 Nx}\frac{(\mu_1 Nx)^i}{i!}\,\mathrm{d}H_{1,\nu_{v,i}}(x)\\&\quad\geqslant \sum_{l=0}^{\infty} p_1(1-p_1)^l\int_0^\infty \mathrm{e}^{-\mu_1 Nx}\,\frac{(\mu_1 Nx)^i}{i!}\,\mathrm{d}G_{N_n^-}^{*l+1}(N_n^- x),\end{aligned}\tag{3.18}$$

and therefore with the same probability

$$\mathbb{E}\{f_1(n+1)\mid f_1(n)\}\geqslant f_1(n)\varphi_{1,N_n^-},\tag{3.19}$$

where φ_{1,N_n^-} is the least in absolute value root of functional equation

$$z=\sum_{l=0}^{\infty} p_1(1-p_1)^l\int_0^\infty \exp(-\mu_1 Nx+\mu_1 Nxz)\,\mathrm{d}G_{N_n^-}^{*l+1}(N_n^- x).$$

Note that with probability 1

$$N_n^-\geqslant N-\sum_{l=0}^{N} f_1(l)-v(T_1+t_0)=c(N),$$

where $v(T_1+t_0)$ is the mutual number of served customers in all other satellite nodes during a period $(0,T_1+t_0)$. Therefore, the stochastic sequence

$$\left\{f_1(n)\prod_{i=1}^{n}\varphi_{1,c(N)}^{-1},\mathcal{F}_{1,n}\right\}\tag{3.20}$$

is a submartingale, where $\varphi_{1,c(N)}$ is the least (absolute) root of functional equation

$$z=\sum_{l=0}^{\infty} p_1(1-p_1)^l\int_0^\infty \exp(-\mu_1 Nx+\mu_1 Nxz)\,\mathrm{d}G_{c(N)}^{*l+1}\big(c(N)x\big).$$

It follows from (3.14), (3.15) that $\mathbb{E}v(T_1+t_0)<\infty$, and therefore

$$\mathbb{P}\left\{\lim_{N\to\infty}\frac{c(N)}{N}=\lim_{N\to\infty}\frac{N_n^-}{N}=\lim_{N\to\infty}\frac{N_n^+}{N}=1\right\}=1.\tag{3.21}$$

Further, it follows from (3.11) and (3.21), as $N\to\infty$, the sequence of supermartingales (3.12) converges to the limiting stochastic sequence

$$\{f_1^*(n)\varphi_1^{-n},\mathcal{F}_{1,n}^*\},\quad \mathcal{F}_{1,n}^*=\sigma\{f_1^*(0),f_1^*(1),\dots,f_1^*(n)\}.\tag{3.22}$$

For this stochastic sequence we have

$$\mathbb{E}\{f_1^*(n+1)\mid f_1^*(n)\}\leqslant \varphi_1 f_1^*(n),$$

and since the limiting stochastic sequence $\{f_1^*(n),\mathcal{F}_{1,n}^*\}$ is Markovian, the limiting stochastic sequence (3.22) forms a supermartingale.

On the other hand the relation (3.21) means that, as $N \to \infty$, φ_{1,N_n^-} converges almost surely to φ_1, and in view of (3.20) and (3.21) the limiting stochastic sequence (3.22) forms a submartingale.

Thus, the stochastic sequence (3.22) is both a super- and submartingale simultaneously and therefore it is a martingale. From this it is clear that as in a case of Markovian queueing network considered in previous section, the limiting up-crossing process $\{f_1^*(n), \mathcal{F}_{1,n}^*\}$ is a Galton–Watson branching process. Moreover, considering as in previous section the multitype up-crossing process $\{\mathbf{f}(\mathbf{n}), \mathcal{G}_\mathbf{n}\}$ (the notation is the same as in previous section) one can conclude that, as $N \to \infty$, this process converges to a multitype Galton–Watson branching process (in sense of the definition of the previous section). In turn, define the up-crossing process $\{\widetilde{\mathbf{f}}(\mathbf{n}), \widetilde{\mathcal{G}}_\mathbf{n}\}$ during generalized busy period and limiting processes provided by asterisk (see notation of the previous section) with the aid of reward renewal theorem we obtain the following theorem:

Theorem 3.1. For every $t > 0$, the limiting as $N \to \infty$ joint queue-length distribution is

$$\mathbb{P}\{Q_1(t) = n_1, Q_2(t) = n_2, \ldots, Q_k(t) = n_k\} = \prod_{j=1}^{k} \mathbb{P}\{Q_j(t) = n_j\},$$

and

$$\mathbb{P}\{Q_j(t) = n_j\} = \rho_j \varphi_j^{n_j - 1}(1 - \varphi_j) \quad (n_j \geqslant 1,\ j = 1, 2, \ldots, k),$$

where φ_j is the least (absolute) root of functional equation

$$z = \frac{p_j \widehat{G}(\mu_j - \mu_j z)}{1 - (1 - p_j)\widehat{G}(\mu_j - \mu_j z)}. \tag{3.23}$$

The proof of this theorem is analogous to the proof of theorem 2.1. Therefore we provide its shorten version. The notation is assumed to be the same.

Proof. Since the limiting up-crossing and branching multitype processes are generated by independent processes, one may consider only marginal probabilities and prove that, as $N \to \infty$, the limiting distributions are

$$\mathbb{P}\{Q_j(t) = n\} = \rho_j \varphi_j^{n-1}(1 - \varphi_j) \quad (n \geqslant 1,\ j = 1, 2, \ldots, k). \tag{3.24}$$

By reward renewal theorem for those limiting distributions one can write:

$$\mathbb{P}\{Q_j(t) = n\} = \lim_{N \to \infty} \frac{\mathbb{E}T_j(n)}{\sum_{i=0}^{\infty} \mathbb{E}T_j(i)} \quad (j = 1, 2, \ldots, k), \tag{3.25}$$

where $T_j(i)$ is the cumulative time in state i for jth node. According to Wald's identity

$$\lim_{N \to \infty} \gamma_j \mathbb{E}T_j(n) = \mathbb{E}\widetilde{f}_j^*(n) = \mathbb{E}f_j^*(n)\mathbb{E}\kappa_j = \varphi_j^n \mathbb{E}\kappa_j \quad (n \geqslant 1) \tag{3.26}$$

and

$$\lim_{N\to\infty}\gamma_j\mathbb{E}T_j(0)=\sum_{n=0}^{\infty}\mathbb{E}\widetilde{f}_j^*(n)-\frac{\gamma_j}{\mu_j}\sum_{n=0}^{\infty}\mathbb{E}\widetilde{f}_j^*(n)=\frac{1-\gamma_j/\mu_j}{1-\varphi_j}\,\mathbb{E}\kappa_j. \tag{3.27}$$

Substituting (3.26) and (3.27) for (3.25) we obtain (3.24). The theorem is proved. □

4. A queueing network with state dependent service times in the hub and *k* single-server satellite stations: Diffusion approximations

For the queueing network considered in previous section let us assume now that, as $N\to\infty$, $g_N p_k\mu_k^{-1}$ converges to one, i.e., there is the small value $\Delta_N=\mu_k-g_N p_k$ vanishing as $N\to\infty$. All other satellite nodes $j=1,2,\ldots,k-1$ are assumed to be usual, satisfying the condition $\mu_j>\gamma_j$. Assuming that the family of probability distributions $\{G_K(x)\}$, $K=1,2,\ldots$, satisfies the standard Liendeberg condition (e.g., see Shiryayev [51]) we shall study the diffusion approximations for both queue-length immediately before a moment of arrival of a customer at the first time after time t and waiting time of a customer arriving at the first time after time t under appropriate normalizations. For $N\to\infty$, these random processes will be denoted by $Q_k(t+)$ and $W_k(t+)$, respectively.

Considering the stochastic order relation $G_K(Kx)\leqslant G_{K+1}(Kx+x)$ for all $K\leqslant N-1$ enables us to use the coupling method based on the following statement (see Stoyan [53], Thorisson [56], Shaked and Shanthikumar [47]).

Lemma 4.1. Let $F_1(x)$ and $F_2(x)$ be two probability distribution functions and $F_2(x)\geqslant F_1(x)$ for all x. Then there exist a probability space $(\Omega,\mathcal{F},\mathbb{P})$ and the random variables κ_1 and κ_2 given over that space such that $F_i(x)=\mathbb{P}\{\kappa_i(\omega)\leqslant x\}$, $i=1,2$, and $\kappa_2(\omega)\leqslant\kappa_1(\omega)$ for all $\omega\in\Omega$.

Let us consider the sequential waiting times at kth satellite node, denoting them by $W_{k,1}, W_{k,2},\ldots$. It is well known that $W_{k,1}=0$, $W_{k,n+1}=\max\{0, W_{k,n}+u_n\}$ where u_n denotes the difference between a service time of nth customer and an interarrival time between nth and $(n+1)$th customers for the given node. For every fixed N let us take two numbers $\mathcal{N}_t^+$ and $\mathcal{N}_t^-$, denoting the maximum and minimum number of customers at the hub during the time interval $(0,t+\delta)$, where $t+\delta$ is the moment of the first arrival after time t of a customer from the central to kth satellite node (as $N\to\infty$, δ vanishes).

For fixed N let us consider the two sequences of waiting times

$$W_{k,1}^+=0,\qquad W_{k,n+1}^+=\max\{0,\,W_{k,n}^++u_n^+\}, \tag{4.1}$$

$$W_{k,1}^-=0,\qquad W_{k,n+1}^-=\max\{0,\,W_{k,n}^-+u_n^-\}, \tag{4.2}$$

where sequence (4.1) corresponds to waiting time in GI/M/1 queueing system with the distribution of interarrival time

$$\sum_{i=0}^{\infty} p_k(1-p_k)^i G_N^{*i+1}(Nx)$$

and the service time parameter $\mu_k N$, and the sequence (4.2) corresponds to waiting time in GI/M/1 queueing system with the distribution of interarrival time

$$\sum_{i=0}^{\infty} p_k(1-p_k)^i G_{\mathcal{N}_t^-}^{*i+1}\big(\mathcal{N}_t^- x\big)$$

and the service time parameter $\mu_k N$. To be brief let us denote those queueing systems by Σ^+ and Σ^-, respectively. Note that for (4.1) we use the probability distribution

$$\sum_{i=0}^{\infty} p_k(1-p_k)^i G_N^{*i+1}(Nx)$$

rather than

$$\sum_{i=0}^{\infty} p_k(1-p_k)^i G_{\mathcal{N}_t^+}^{*i+1}\big(\mathcal{N}_t^+ x\big),$$

having more complicated form. We may use the probability distribution indexed by N instead of that indexed by $\mathcal{N}_t^+$ because $\mathcal{N}_t^+ \leqslant N$ with probability 1, and estimations connected with the mentioned probability distribution indexed by N are sufficient for our further purposes.

Next, it is clear that with probability 1

$$\sum_{i=0}^{\infty} p_k(1-p_k)^i G_{\mathcal{N}_t^-}^{*i+1}\big(\mathcal{N}_t^- x\big) \leqslant \widetilde{G}_{r_l}(x)$$

for all $\tau_l < t$ (see notation in section 3) and therefore with the same probability

$$\sum_{i=0}^{\infty} p_k(1-p_k)^i G_{\mathcal{N}_t^-}^{*i+1}\big(\mathcal{N}_t^- x\big) \leqslant H_{k,\nu}(x),$$
$$\nu \in \mathcal{M}_t = \{m\colon\ m\text{th arrival time moment to node } k \text{ is less than } t\}. \qquad (4.3)$$

Further we shall distinguish the probability space $\{\Omega, \mathcal{F}, \mathbb{P}\}$ in lemma 4.1 from the initial probability space, where the random objects $\mathcal{N}_t^-$ and ν are given. For this purpose this initial probability space will be denoted by $\{\Omega', \mathcal{F}', \mathbb{P}'\}$. Thus, inequality (4.3) and a number of inequalities above, where the aforementioned random objects $\mathcal{N}_t^-$ and ν are presented, are fulfilled $\mathbb{P}'$-a.s.

According to (3.3), (4.3) and lemma 4.1 there exist a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and three sequences of waiting times $\{W^+_{k,n}\}$, $\{W^-_{k,n}\}$, $\{W_{k,n}\}$ given over that space, such that

$$W^+_{k,n}(\omega) \leqslant W_{k,n}(\omega) \leqslant W^-_{k,n}(\omega) \tag{4.4}$$

for all n and all $\omega \in \Omega$ $\mathbb{P}'$-a.s. Indeed, according to lemma one can construct the probability space and correspondent random variables given over that space such that

$$u^-_n(\omega) \leqslant u_n(\omega) \leqslant u^+_n(\omega)$$

for all $n = 1, 2, \ldots$ and $\omega \in \Omega$ $\mathbb{P}'$-a.s., and the desired inequality (4.4) follows from (4.1), (4.2) and relation $W_{k,n+1} = \max\{0, W_{k,n} + u_n\}$.

For our further purposes we need that inequalities (4.4) as well as other inequalities between the random variables given on the probability space $\{\Omega, \mathcal{F}, \mathbb{P}\}$ should be fulfilled for almost all $\omega \in \Omega$ ($\mathbb{P}$-a.s.) rather than $\mathbb{P}'$-a.s. That is we have to construct the measure $\mathbb{P}$ such that it would be absolutely continuous with respect to measure $\mathbb{P}'$. To make that note, that it follows from (4.3)

$$\mathbb{P}'\left\{\sum_{i=0}^{\infty} p_k(1-p_k)^i G^{*i+1}_{\mathcal{N}^-_t}\big(\mathcal{N}^-_t x\big) \leqslant H_{k,\nu}(x)\right\} = 1,$$

and therefore

$$\mathbb{P}'\big\{H_{k,\nu}(x) \leqslant y\big\} \leqslant \mathbb{P}'\left\{\sum_{i=0}^{\infty} p_k(1-p_k)^i G^{*i+1}_{\mathcal{N}^-_t}\big(\mathcal{N}^-_t x\big) \leqslant y\right\}.$$

Thus, according to lemma 4.1 there exists an intermediate probability space $\{\Omega'', \mathcal{F}'', \mathbb{P}''\}$ and the random variables χ_1 and χ_2, given over this space, such that

$$\mathbb{P}''\{\chi_2 \leqslant y\} = \mathbb{P}'\big\{H_{k,\nu}(x) \leqslant y\big\}$$

and

$$\mathbb{P}''\{\chi_1 \leqslant y\} = \mathbb{P}'\left\{\sum_{i=0}^{\infty} p_k(1-p_k)^i G^{*i+1}_{\mathcal{N}^-_t}\big(\mathcal{N}^-_t x\big) \leqslant y\right\},$$

and $\chi_1(\omega'') \leqslant \chi_2(\omega'')$ for all $\omega'' \in \Omega''$. Denoting also

$$\mathbb{P}''\{\chi_3 \leqslant y\} = \mathbb{P}'\left\{\sum_{i=0}^{\infty} p_k(1-p_k)^i G^{*i+1}_{N}(Nx) \leqslant y\right\},$$

we obtain $\chi_1(\omega'') \leqslant \chi_2(\omega'') \leqslant \chi_3(\omega'')$ for all $\omega'' \in \Omega''$. Applying lemma 4.1 once again, we achieve slightly more. Inequalities (4.4) are fulfilled for *all* $\omega \in \Omega$ rather than *almost all* $\omega \in \Omega$.

Note, that the inequalities similar to (4.4) hold for sequences of queue-length, given on the suitable probability space. Considering three sequences of queue-length $\{Q_{k,n}^{+}\}$, $\{Q_{k,n}^{-}\}$, $\{Q_{k,n}\}$ immediately before arrival of a customer,

$$\begin{aligned} Q_{k,1}^{+} &= Q_{k,1}^{-} = Q_{k,1} = 0, \\ Q_{k,n+1}^{+} &= \max\{0, Q_{k,n}^{+} + 1 - q_n^{+}\}, \\ Q_{k,n+1}^{-} &= \max\{0, Q_{k,n}^{-} + 1 - q_n^{-}\}, \\ Q_{k,n+1} &= \max\{0, Q_{k,n} + 1 - q_n\}, \end{aligned}$$

where the random variables q_n^{+}, q_n^{-}, q_n are defined as the *possible service completions* during interarrival time between nth and $(n+1)$th customers in queueing systems Σ^{+}, Σ^{-} and kth node of considered network, respectively. Under possible service completions q during interarrival time ξ we mean $q = \max\{m\colon \phi_1 + \phi_2 + \cdots + \phi_m \leqslant \xi\}$, where $\phi_1, \phi_2, \ldots$ is a sequence of service times. Here, in the case of the queueing system Σ^{+} the respective sequence $\{\phi_i\}_{i\geqslant 1}$ consists of independent identically distributed random variables, in the case of the queueing system Σ^{-} that sequence consists of independent random variables conditionally dependent of $\mathcal{N}_t^{-}$, and in the case of the kth node of network the sequence consists of dependent on the states of given and other nodes and not identically distributed. However, it is not so considerable for our purposes.

Analogously to (4.4), by lemma 4.1 one can write

$$Q_{k,n}^{+}(\omega) \leqslant Q_{k,n}(\omega) \leqslant Q_{k,n}^{-}(\omega) \tag{4.5}$$

for all $\omega \in \Omega$. Moreover, by lemma 4.1 on the suitable probability space

$$Q_{j,n}(\omega) \leqslant Q_{k,n}^{+}(\omega) \tag{4.6}$$

for all $j = 1, 2, \ldots, k-1$ and $\omega \in \Omega$.

Let $\Delta_N = \mu_k - g_N p_k$. Assuming that $N \to \infty$ and $\Delta_N \to 0$, let us consider the following cases:

$$N\Delta_N^2 \to \infty, \tag{4.7}$$

$$N\Delta_N^2 \to C > 0, \tag{4.8}$$

$$N\Delta_N^2 \to 0. \tag{4.9}$$

According to theorem 18 by Borovkov [8] (see also Prohorov [44], Ivchenko et al. [20]) under assumptions (4.7)–(4.9) we have the following. Let Δ_N vanishes as $N \to \infty$ such that the sign is not changed. Then, if under assumption (4.7) Δ_N remains positive,

$$\lim_{N\to\infty} \mathbb{P}\left\{\frac{Q_{k,N}^{+}\Delta_N}{p_k g_N} \leqslant x\right\} = 1 - \exp\left(-2x\sigma^{-2}\right), \tag{4.10}$$

where $\sigma = \lim_{N\to\infty} \sigma_N$, $\sigma_N^2 = \mathbf{D}q_1^{+}$, and if Δ_N remains negative,

$$\lim_{N\to\infty} \mathbb{P}\left\{\frac{Q_{k,N}^{+} - \Delta_N (p_k g_N)^{-1} N}{\sqrt{N}} \leqslant x\right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x\sigma^{-1}} \mathrm{e}^{-u^2/2}\,\mathrm{d}u. \tag{4.11}$$

Under assumption (4.8) we have:

$$\lim_{N\to\infty}\mathbb{P}\left\{\frac{Q_{k,N}^{+}|\Delta_N|}{p_k g_N}\leqslant x\right\}=\mathbb{P}\left\{w(u)\leqslant\frac{x+u\,\mathrm{sign}(\Delta_N)}{\sigma},0\leqslant u\leqslant C\right\},\tag{4.12}$$

where $w(u)$ is a standard Wiener process. For explicit expression for the right side of (4.12) see Borovkov [8]. Under assumption (4.9),

$$\lim_{N\to\infty}\mathbb{P}\left\{\frac{Q_{k,N}^{+}}{\sigma\sqrt{N}}\leqslant x\right\}=2\Phi(x)-1,\tag{4.13}$$

where $\Phi(x)$ is the function of standard normal distribution.

Next, for $\mathcal{N}_t^-$ one can write

$$\mathcal{N}_t^-\geqslant N-\sup_{s\leqslant t}Q_k(s)-\sum_{i=1}^{k-1}\sup_{s\leqslant t}Q_i(s).\tag{4.14}$$

It follows from (4.10)–(4.13) that $N^{-1}Q_{k,N}^{+}$ vanishes in probability as $N\to\infty$, therefore by (4.5)

$$\mathbb{P}\text{-}\lim_{N\to\infty}\frac{1}{N}\sup_{s\leqslant t}Q_k(s)=0\tag{4.15}$$

as well, and keeping in mind (4.6) one can also write

$$\mathbb{P}\text{-}\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{k-1}\sup_{s\leqslant t}Q_i(s)=0.\tag{4.16}$$

In view of (4.14)–(4.16) we have

$$\mathbb{P}\text{-}\lim_{N\to\infty}\frac{\mathcal{N}_t^-}{N}=1$$

and thus for the queue-length process of queueing system Σ^- one can state the results similar to that Σ^+ under assumption (4.7)–(4.9). Namely, let Δ_N vanishes as $N\to\infty$ such that the sign is not changed. Then under assumption (4.7) if Δ_N remains positive,

$$\lim_{N\to\infty}\mathbb{P}\left\{\frac{Q_{k,N}^{-}\Delta_N}{p_k g_N}\leqslant x\right\}=1-\exp\bigl(-2x\sigma^{-2}\bigr),\tag{4.17}$$

and if Δ_N remains negative,

$$\lim_{N\to\infty}\mathbb{P}\left\{\frac{Q_{k,N}^{-}-\Delta_N(p_k g_N)^{-1}N}{\sqrt{N}}\leqslant x\right\}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x\sigma^{-1}}\mathrm{e}^{-u^2/2}\,\mathrm{d}u.\tag{4.18}$$

Under assumption (4.8) we have:

$$\lim_{N\to\infty}\mathbb{P}\left\{\frac{Q_{k,N}^{-}|\Delta_N|}{p_k g_N}\leqslant x\right\}=\mathbb{P}\left\{w(u)\leqslant\frac{x+u\,\mathrm{sign}(\Delta_N)}{\sigma},0\leqslant u\leqslant C\right\}.\tag{4.19}$$

Under assumption (4.9),

$$\lim_{N\to\infty} \mathbb{P}\left\{\frac{Q^-_{k,N}}{\sigma\sqrt{N}} \leqslant x\right\} = 2\Phi(x) - 1. \tag{4.20}$$

In view of (4.10)–(4.13), (4.17)–(4.20) as well as (4.5) one can conclude the following. Let Δ_N vanishes as $N \to \infty$ such that the sign is not changed. Then, if under assumption (4.7) Δ_N remains positive,

$$\lim_{N\to\infty} \mathbb{P}\left\{\frac{Q_{k,N}\Delta_N}{p_k g_N} \leqslant x\right\} = 1 - \exp\bigl(-2x\sigma^{-2}\bigr), \tag{4.21}$$

and if Δ_N remains negative,

$$\lim_{N\to\infty} \mathbb{P}\left\{\frac{Q_{k,N} - \Delta_N(p_k g_N)^{-1}N}{\sqrt{N}} \leqslant x\right\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x\sigma^{-1}} \mathrm{e}^{-u^2/2}\,\mathrm{d}u. \tag{4.22}$$

Under assumption (4.8) we have:

$$\lim_{N\to\infty} \mathbb{P}\left\{\frac{Q_{k,N}|\Delta_N|}{p_k g_N} \leqslant x\right\} = \mathbb{P}\left\{w(u) \leqslant \frac{x + u\,\mathrm{sign}(\Delta_N)}{\sigma}, 0 \leqslant u \leqslant C\right\}. \tag{4.23}$$

Under assumption (4.9),

$$\lim_{N\to\infty} \mathbb{P}\left\{\frac{Q_{k,N}}{\sigma\sqrt{N}} \leqslant x\right\} = 2\Phi(x) - 1. \tag{4.24}$$

The limiting relations (4.21)–(4.24) permit us to obtain the following

Theorem 4.1. Let $N \to \infty$ and $\Delta_N \to 0$ not changing the sign such that $N\Delta_N^2 \to \infty$. Then if Δ_N remains negative,

$$\lim_{N\to\infty} \mathbb{P}\left\{\frac{Q_k(t+) - \Delta_{[t\gamma_k N]}tN}{\sqrt{t\gamma_k N}} \leqslant x\right\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x\sigma^{-1}} \mathrm{e}^{-u^2/2}\,\mathrm{d}u,$$

where $[t\gamma_k N]$ in this formula denotes the integer part of $t\gamma_k N$. Under assumption that Δ_N remains positive and there exists the minimum real number $\alpha > 2$ such that $N\Delta_N^\alpha \leqslant C_1$, as $N \to \infty$, where C_1 is a some constant,

$$\lim_{N\to\infty} \mathbb{P}\left\{\frac{Q_k(t+)\Delta_N}{\gamma_k} \leqslant x\right\} = 1 - \exp\bigl[-2x(t\gamma_k)^{-1/\alpha}\sigma^{-2}\bigr]. \tag{4.25}$$

If $N\Delta_N^2 \to C > 0$,

$$\lim_{N\to\infty} \mathbb{P}\left\{\frac{Q_k(t)|\Delta_N|}{\gamma_k} \leqslant x\right\} = \mathbb{P}\left\{\frac{w(u)}{\sqrt{t\gamma_k}} \leqslant \frac{x + u\,\mathrm{sign}(\Delta_N)}{\sigma}, 0 \leqslant u \leqslant C\right\}.$$

If $N\Delta_N^2 \to 0$,

$$\lim_{N\to\infty} \mathbb{P}\left\{\frac{Q_k(t)}{\sigma\sqrt{N}} \leqslant x\right\} = 2\Phi\left(\frac{x}{\sqrt{t\gamma_k}}\right) - 1.$$

Proof. Indeed, denote the number of arrived customer at satellite node k at time moment $t+\delta$ by T. Then, if under assumption (4.7) Δ_T positive,

$$\lim_{T\to\infty} \mathbb{P}\left\{\frac{Q_k(t+)\Delta_T}{\gamma_k} \leqslant x\right\} = 1 - \exp\left(-2x\sigma^{-2}\right), \tag{4.26}$$

and if Δ_T negative,

$$\lim_{T\to\infty} \mathbb{P}\left\{\frac{Q_k(t+) - \Delta_T\gamma_k^{-1}T}{\sqrt{T}} \leqslant x\right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x\sigma^{-1}} \mathrm{e}^{-u^2/2}\,\mathrm{d}u. \tag{4.27}$$

Under assumption (4.8),

$$\lim_{T\to\infty} \mathbb{P}\left\{\frac{Q_k(t+)\Delta_T}{\gamma_k} \leqslant x\right\} = \mathbb{P}\left\{w(u) \leqslant \frac{x + u\,\mathrm{sign}(\Delta_T)}{\sigma}, 0 \leqslant u \leqslant C\right\}. \tag{4.28}$$

Under assumption (4.9),

$$\lim_{T\to\infty} \mathbb{P}\left\{\frac{Q_k(t+)}{\sigma\sqrt{T}} \leqslant x\right\} = 2\Phi(x) - 1. \tag{4.29}$$

Next, as $N \to \infty$, let us show that for every small $\varepsilon > 0$, the probability, that the number of arrivals to satellite node k up to time t varies within interval

$$\big((1-\varepsilon)t\gamma_k N, (1+\varepsilon)t\gamma_k N\big),$$

tends to one. For this purpose let us take into account the following lemma by Krichagina et al. [35].

Lemma 4.2. Let $\mathcal{A}^N = (\mathcal{A}_t^N)_{t\geqslant 0}$, $N \geqslant 1$, be a sequence of increasing right continuous random processes with $\mathcal{A}_0^N = 0$. Let

$$\mathcal{B}_t^N = \inf\{s\colon \mathcal{A}_s^N > t\}, \quad t \geqslant 0,$$

where the infinimum of empty set is assumed to be equal to infinity.

If for every t from dense in $\mathbb{R}_+$ set S, $\mathcal{B}_t^N \to \alpha t$ as $N \to \infty$ ($\alpha > 0$), then, as $N \to \infty$,

$$\sup_{t\leqslant\tau}\left|\mathcal{A}_t^N - \frac{t}{\alpha}\right| \to 0$$

in probability for each $\tau > 0$.

Let us apply this lemma. Let $\xi_{1,N}^{+}, \xi_{2,N}^{+}, \dots$ be the sequence of independent identically distributed random variables having the probability distribution

$$\sum_{i=0}^{\infty} p_k(1-p_k)^i G_N^{*i+1}(Nx),$$

and $\xi_{1,N}^{-}, \xi_{2,N}^{-}, \dots$ be the sequence of independent identically distributed random variables having the probability distribution

$$\sum_{i=0}^{\infty} p_k(1-p_k)^i G_{\mathcal{N}_t^-}^{*i+1}\bigl(\mathcal{N}_t^- x\bigr),$$

and

$$S_N^{+}(t) = \max\left\{m\colon \sum_{i=1}^{m} \xi_{i,N}^{+} \leqslant t\right\},$$

$$S_N^{-}(t) = \max\left\{m\colon \sum_{i=1}^{m} \xi_{i,N}^{-} \leqslant t\right\}.$$

If take $\mathcal{A}_t^N = N^{-1}S_N^{+}(t)$, then

$$\mathcal{B}_t^N = \sum_{i=1}^{[Nt]+1} \xi_{i,N}^{+},$$

and if in turn take $\mathcal{A}_t^N = (\mathcal{N}_t^-)^{-1}S_N^{-}(t)$, then

$$\mathcal{B}_t^N = \sum_{i=1}^{[\mathcal{N}_t^- t]+1} \xi_{i,N}^{-},$$

where $[Nt]$ and $[\mathcal{N}_t^- t]$ denotes the integer parts of Nt and $\mathcal{N}_t^- t$, respectively. According to this lemma, as $N \to \infty$, $N^{-1}S_N(t) \to \gamma_k t$ in probability, and because of $\lim_{N\to\infty} N^{-1}\mathcal{N}_t^- = 1$ with probability 1, we also obtain $(\mathcal{N}_t^-)^{-1}S_N^{-}(t) \to \gamma_k t$ in probability (an extension of the lemma 4.2 for N increasing a.s. to infinity is obvious), i.e., in our notation $N^{-1}T \to \gamma_k t$ in probability as $N \to \infty$, and the required statement follows. Therefore, for sufficiently large N one may take $T = [t\gamma_k N]$, where $[t\gamma_k N]$ denotes the integer part of $t\gamma_k N$. Thus the statements of theorem follow from (4.26)–(4.29) after the little algebraic calculations. □

Note 4.1. If $N\Delta_N^2 \to \infty$ and there is no the minimum real number α such that $N\Delta_N^{\alpha} < C_1$, as $N \to \infty$, then instead of limit relation (4.25) one can at most write the limit relation in the form

$$\lim_{N\to\infty} \mathbb{P}\left\{\frac{Q_k(t+)\Delta_{[t\gamma_k N]}}{\gamma_k} \leqslant x\right\} = 1 - \exp\bigl(-2x\sigma^{-2}\bigr),$$

where $[t\gamma_k N]$ denotes the integer part of $t\gamma_k N$.

Now, let us consider the waiting time process $W(t+)$. In this case $\mathbb{E}u_1^+ = \Delta_N N^{-1}$, and $\mathbf{D}u_1^+ = \sigma_N^2(p_k g_N N)^{-2} = \sigma_{1,N}N^{-2}$, $\sigma_{1,N} \to \sigma_1$ as $N \to \infty$, and we therefore obtain

Theorem 4.2. Let $N \to \infty$ and $\Delta_N \to 0$. Then

$$\lim_{N\to\infty} \mathbb{P}\left\{\frac{W(t+)\sqrt{N}}{\sigma_1} \leqslant x\right\} = 2\Phi\left(\frac{x}{\sqrt{t\gamma_k}}\right) - 1.$$

Note 4.2. In the account of this section we did not keep in mind that the service times are exponentially distributed random variables. The results of this section hold therefore under the more general setting when the service times are independent generally distributed random variables obeying the appropriate moment conditions.

5. A queueing network with state dependent service times in the hub and k single-server satellite stations: Heavy usage regime at the kth satellite node

This section considers the case of network when $\gamma_j < \mu_j$, $j = 1, 2, \dots, k-1$, and $\gamma_k > \mu_k$. In this case satellite node k is said to be operated in heavy usage regime. Providing below a rigorous explanation related to the case of a GI/M/1 queueing system under heavy traffic we show how the queue-length process is approximated by fluid model. The grounds for the fluid model are not new (see, e.g., Chen and Mandelbaum [12] and Dai [15] for applications of the fluid models to different networks of queues). The explanation given below is specific to the case of considered queueing system since it uses explicit representation of the queue-length process. At the same time it is very simple and clear.

Let us consider a standard GI/M/1 queueing system. To be close to the network which is considered afterwards, we use the earlier notation. The probability distribution of interarrival time is

$$\sum_{i=0}^{\infty} p_k(1-p_k)^i G_N^{*i+1}(Nx)$$

and service time parameter $\mu_k N$ under assumption that $\pi_N = p_k g_N \mu_k^{-1} > 1$. (Here it is assumed that N is fixed.)

Let $A_N(t)$ denote arrival process

$$A_N(t) = \max\left\{m\colon \sum_{i=1}^{m} \xi_{i,N} \leqslant t\right\}$$

($\xi_{1,N}, \xi_{2,N}, \dots$ are interarrival times), and let $B_N(t)$ in turn denote departure process

$$B_N(t) = \max\left\{m\colon \sum_{i=1}^{m} \phi_{i,N} \leqslant t\right\}$$

($\phi_{1,N}, \phi_{2,N}, \dots$ are service times, exponentially distributed random variables).

A GI/M/1 queueing system can be considered as the queueing system with autonomous service because of the property of the lack of memory of exponential distribution. Therefore, the queue-length process $q_N(t)$ is described by the equation

$$q_N(t) = A_N(t) - \int_0^t \mathbf{I}\bigl\{q_N(s-) > 0\bigr\}\,\mathrm{d}B_N(s), \tag{5.1}$$

where integral in (5.1) is the Lebesgue–Stieltjes integral. It is evident that $\mathbb{P}$-a.s.

$$q_N(t) \geqslant A_N(t) - B_N(t),$$

and therefore also $\mathbb{P}$-a.s.

$$\lim_{t\to\infty} \frac{q_N(t)}{t} \geqslant \lim_{t\to\infty} \frac{A_N(t) - B_N(t)}{t}. \tag{5.2}$$

Fluid approximation having place under heavy thaffic condition means that $\mathbb{P}$-a.s.

$$\lim_{t\to\infty} \frac{q_N(t)}{t} = \lim_{t\to\infty} \frac{A_N(t) - B_N(t)}{t} = N(p_k g_N - \mu_k). \tag{5.3}$$

Indeed, it is known that there is the probability π_N^{-1} that a busy period is finite (see, e.g., Klimov [27]) and hence the probability $1 - \pi_N^{-1}$ that it is infinite. Therefore the path of the queue-length process $q_N(t)$ includes several busy periods at the beginning and one infinite busy period at the end. Thus, with probability 1 there is only a finite number of points where the queue-length decreases to zero, and (5.3) therefore follows. Note, that it also follows from (5.3) that

$$\lim_{t\to\infty} \frac{\mathbb{E}q_N(t)}{t} = N(p_k g_N - \mu_k). \tag{5.4}$$

Note, that relations similar to (5.3) and (5.4) hold also as t is given and $N \to \infty$, where it is assumed that all queueing systems indexed by N are given on the same probability space. We have:

$$\lim_{N\to\infty} \frac{q_N(t)}{N} = \lim_{t\to\infty} \frac{A_N(t) - B_N(t)}{N} = (\gamma_k - \mu_k)t \quad (\mathbb{P}\text{-a.s.}), \tag{5.5}$$

$$\lim_{N\to\infty} \frac{\mathbb{E}q_N(t)}{N} = (\gamma_k - \mu_k)t. \tag{5.6}$$

Note, that in the limiting case as $N \to \infty$ it follows $\pi_\infty = \lim_{N\to\infty} \pi_N = \gamma_k \mu_k^{-1}$.

Let us consider now the kth satellite node of queueing network. According to (3.3) and lemma 4.1 there exist a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and the random variables $q_N(t)$ and $Q_{k,N}(t)$ given over that space such that $Q_{k,N}(t,\omega) \leqslant q_N(t,\omega)$ for all $\omega \in \Omega$. On the other hand, for $\kappa \geqslant \tau$ denoting $\mathcal{N}^-_{\tau,\kappa} = \inf_{\tau\leqslant s\leqslant\kappa} Q_{0,N}(s)$, and considering the standard GI/M/1 queueing system with the distribution of interarrival time

$$\sum_{i=0}^{\infty} p_k(1-p_k)^i G^{*i+1}_{\mathcal{N}^-_{0,t}}\bigl(\mathcal{N}^-_{0,t}x\bigr)$$

and service time parameter $\mu_k N$, one can write:

$$\sum_{i=0}^{\infty} p_k(1-p_k)^i G^{*i+1}_{\mathcal{N}^-_{0,t}}\big(\mathcal{N}^-_{0,t}x\big) \leqslant H_{k,\nu}(x) \leqslant \sum_{i=0}^{\infty} p_k(1-p_k)^i G^{*i+1}_N(Nx), \quad \mathbb{P}'\text{-a.s.},$$

where $\{\Omega', \mathcal{F}', \mathbb{P}'\}$ is the probability space for the random objects $\mathcal{N}^-_{0,t}$ and ν (see also (4.3) and the notation of previous section). Therefore, according to construction of previous section connected with sequential application of lemma 4.1, we have:

$$q_{\mathcal{N}^-_{0,t}}(t,\omega) \leqslant Q_{k,N}(t,\omega) \leqslant q_N(t,\omega) \tag{5.7}$$

for all $\omega \in \Omega$.

Next, let us consider $Q_{k,N}(s,\omega), 0 < s \leqslant \delta$, and assume that $N\delta \to \infty$ as soon as $N \to \infty$ and $\delta \to 0$. It is assumed that the probability space $\{\Omega, \mathcal{F}, \mathbb{P}\}$ contains all queue-length processes and necessary random objects indexed by N. Under these conditions by virtue of (5.5) $(N\delta)^{-1}q_N(\delta)$ converges in probability to $\gamma_k - \mu_k$. Therefore, as $N\delta \to \infty$ as soon as $N \to \infty$ and $\delta \to 0$, the inequality $N^{-1}q_N(\delta) < f\delta$ is fulfilled for some constant f with increasing to 1 probability. Moreover, since the traffic intensities of all other satellite nodes are less than 1,

$$\mathbb{P} - \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{k-1} \sup_{s\leqslant t} Q_{i,N}(s) = 0 \tag{5.8}$$

(see (4.16) and its explanation).

Therefore, we obtain

$$\mathbb{P}\left\{\lim_{N\to\infty} \frac{\mathcal{N}^-_{0,\delta}}{N} = 1\right\} = 1. \tag{5.9}$$

Analogously, for $Q_k(s,\omega)$, $\delta < s \leqslant 2\delta$, under the same assumption of $N\delta \to \infty$ as soon as $N \to \infty$ and $\delta \to 0$ we have

$$\mathbb{P}\left\{\lim_{N\to\infty} \frac{\mathcal{N}^-_{\delta,2\delta}}{\mathcal{N}^-_{0,\delta}} = 1\right\} = 1.$$

Continuing this procedure we obtain

$$\mathbb{P}\left\{\lim_{N\to\infty} \frac{\mathcal{N}^-_{t,t+\delta}}{\mathcal{N}^-_{t,t}} = 1\right\} = 1 \tag{5.10}$$

for all $t \geqslant 0$. Therefore, there exists $\mathbb{P}$-a.s. continuous decreasing function $C(t)$ obeying the conditions:

$$C(0) = 1, \qquad \mathbb{P}\left\{\lim_{N\to\infty} \frac{\mathcal{N}^-_{t,t}}{N} = C(t)\right\} = 1.$$

The function $C(t)$ describes the normalized queue-length at the hub. Because of the quite general assumptions it cannot be expressed in explicit form. In view of (5.8) the normalized queue-length at the bottleneck node is

$$\mathbb{P}\text{-}\lim_{N\to\infty}\frac{1}{N}Q_{k,N}(t)=1-C(t). \tag{5.11}$$

It is known that if the hub is an infinite-server queueing system with exponentially distributed service times then for normalized queue-length at the bottleneck node we have $\mathbb{P}\text{-}\lim_{N\to\infty}N^{-1}Q_{k,N}(t)=(1-\pi_\infty^{-1})(1-\mathrm{e}^{-\gamma_k t})$ (see Kogan and Liptser [31], Abramov [4]).

Let us consider now one of the first satellite $k-1$ nodes, the first satellite node for example, and construct the intervals (2.1) and (2.2) by the same manner as it has been done in the second section, considering the first busy period after time t. Repeating the all of procedure of the third section, take into account that the queue-length at satellite node k, during the first busy period after time t at the first satellite node, contains the number of customers close to the value $N[1-C(t)]$.

Let us consider this case more carefully.

As at time t satellite node k contains the number of customers equivalent to $N[1-C(t)]$, the hub therefore contains the number of customers equivalent to $NC(t)$. Therefore,

$$\mathbb{P}\left\{\lim_{N\to\infty}\frac{N_n^+}{N}=\lim_{N\to\infty}\frac{N_n^-}{N}=C(t)\right\}=1.$$

Thus, in this case the stochastic sequence $\{f_1^*(n)[\psi_1(t)]^{-n},\mathcal{F}_{1,n}^*\}$ forms a martingale, where $\psi_1(t)$ is the least in absolute value root of functional equation

$$z=\frac{p_1\widehat{G}([C(t)]^{-1}(\mu_1-\mu_1 z))}{1-(1-p_1)\widehat{G}([C(t)]^{-1}(\mu_1-\mu_1 z))}. \tag{5.12}$$

According to the formula for the total probability the limiting as $N\to\infty$ queue-length distribution at the first satellite node is

$$\mathbb{P}\{Q_1(t)=n\}=\rho_1(t)[1-\psi_1(t)][\psi_1(t)]^{n-1},\quad n\geqslant 1, \tag{5.13}$$

where $\psi_1(t)$ is the least root of (5.12),

$$\rho_1(t)=\gamma_1\mu_1^{-1}C(t).$$

The theorem below gives the limiting joint queue-lengths distribution in the first $k-1$ satellite nodes. We do not provide the proof because it is analogous to those in the second and third sections.

Theorem 5.1. For every $t>0$, the limiting as $N\to\infty$ joint queue-length distribution is

$$\mathbb{P}\{Q_1(t)=n_1,Q_2(t)=n_2,\dots,Q_{k-1}(t)=n_{k-1}\}=\prod_{j=1}^{k-1}P_j(n_j),$$

where

$$\begin{aligned}
&P_j(n_j) = \rho_j(t)\big[\psi_j(t)\big]^{n_j-1}\big[1-\psi_j(t)\big], \quad n_j \geqslant 1,\\
&\rho_j(t) = \gamma_j\mu_j^{-1}C(t),\\
&P_j(0) = 1-\rho_j(t), \quad j = 1,2,\dots,k-1,
\end{aligned}$$

and $\psi_j(t)$ is the least in absolute value root of functional equation

$$z = \frac{p_j\widehat{G}([C(t)]^{-1}(\mu_j - \mu_j z))}{1-(1-p_j)\widehat{G}([C(t)]^{-1}(\mu_j-\mu_j z))}. \tag{5.14}$$

The function $C(t)$ is a decreasing positive a.s. continuous function, $C(0) = 1$.

Theorem 5.1 does not give us the limiting joint queue-lengths distribution in explicit form as long as there is not explicit representation for the function $C(t)$. Below we discuss the case where the function $C(t)$ can be evaluated. Approach of the case is analogous to that of the earlier mentioned papers by Kogan and Liptser [31] and Abramov [4], therefore we shall try to be possibly brief omitting the technical details.

Let $G_i(t) \geqslant 1 - \mathrm{e}^{-\lambda t}$ ($\lambda \leqslant g$) for all $i = 1, 2, \dots$, and $A_{j,N}(t)$ denote an arrival process to satellite node j. Then

$$A_{j,N}(t) \geqslant \int_0^t \sum_{i=1}^{N} \mathbf{I}\left\{N - \sum_{l=1}^{k} Q_{l,N}(s-) \geqslant i\right\} \mathrm{d}\Pi_{j,i}(s), \tag{5.15}$$

where $\{\Pi_{j,i}(t)\}$, $i = 1, 2, \dots, N$, is a collection of independent Poisson processes with rate $\lambda p_j = \lambda_j$, and $\lambda_j < \mu_j$ for $j = 1, 2, \dots, k-1$ while $\lambda_k > \mu_k$.

The departure process $B_{j,N}(t)$ is similar to the earlier notation $B_N(t)$ of this section for the GI/M/1 queueing system, and the queue-length process is described as

$$Q_{j,N}(t) = A_{j,N}(t) - B_{j,N}(t) + \int_0^t \mathbf{I}\big\{Q_{j,N}(s-) = 0\big\}\,\mathrm{d}B_{j,N}(s). \tag{5.16}$$

Equation (5.16) implies that $Q_{j,N}(t)$ is the normal reflection of the process

$$X_{j,N}(t) = A_{j,N}(t) - B_{j,N}(t), \qquad X_{j,N}(0) = 0 \tag{5.17}$$

at zero, being a non-negative solution of the Skorohod problem (see Skorohod [52], Tanaka [55], Anulova and Liptser [6]). Therefore the function

$$\varphi_j(t) = \int_0^t \mathbf{I}\big\{Q_{j,N}(s-) = 0\big\}\,\mathrm{d}B_{j,N}(s) \tag{5.18}$$

can be represented as

$$\varphi_j(t) = -\inf_{s\leqslant t} X_{j,N}(s), \tag{5.19}$$

and from (5.16)–(5.19) we have

$$Q_{j,N}(t) = X_{j,N}(t) - \inf_{s \leqslant t} X_{j,N}(s). \tag{5.20}$$

Next, let us take into account that the processes $A_{j,N}(t)$ and $B_{j,N}(t)$ are the semimartingales adapted with respect to the filtration $\mathcal{F}_t$ given on stochastic basis $(\Omega, \mathcal{F}, \mathbf{F} = (\mathcal{F}_t)_{t \geqslant 0}, \mathbb{P})$. Then the process $X_{j,N}(t)$ in (5.20) can be rewritten in the form of the Doob–Meyer semimartingale decomposition (see, e.g., Liptser and Shiryayev [37], Jacod and Shiryayev [21]). Denoting $A^p_{j,N}(t)$ and $B^p_{j,N}(t)$ the compensators of the processes $A_{j,N}(t)$ and $B_{j,N}(t)$, respectively, then the processes $A_{j,N}(t) - A^p_{j,N}(t)$ and $B_{j,N}(t) - B^p_{j,N}(t)$ are the local square integrable martingales, and we therefore obtain:

$$X_{j,N}(t) = A^p_{j,N}(t) - B^p_{j,N}(t) + M_{j,N}(t), \tag{5.21}$$

where

$$M_{j,N}(t) = \left[A_{j,N}(t) - A^p_{j,N}(t)\right] - \left[B_{j,N}(t) - B^p_{j,N}(t)\right]$$

is a local square integrable martingale.

Note that by virtue of (5.15)

$$A^p_{j,N}(t) \geqslant \int_0^t \lambda_j \left\{ N - \sum_{l=1}^{k} Q_{l,N}(s) \right\} \mathrm{d}s \tag{5.22}$$

(for details, see Dellacherie [16], Liptser and Shiryayev [37,38] theorem 1.6.1). Then after normalization we have:

$$\frac{1}{N} Q_{j,N}(t) = \frac{1}{N} A^p_{j,N}(t) - \frac{1}{N} B^p_{j,N}(t) + \frac{1}{N} M_{j,N}(t) - \frac{1}{N} \inf_{s \leqslant t} X_{j,N}(t). \tag{5.23}$$

Next, since $B_{j,N}(t)$ is a Poisson process then

$$\mathbb{P}\text{-}\lim_{N \to \infty} \frac{1}{N} B^p_{j,N}(t) = \mu_j t. \tag{5.24}$$

Show also that

$$\mathbb{P}\text{-}\lim_{N \to \infty} \frac{1}{N} A^p_{j,N}(t) = \mathbb{P}\text{-}\lim_{N \to \infty} \frac{1}{N} A_{j,N}(t). \tag{5.25}$$

Indeed, according to Lenglart–Reboleddo inequality (see, e.g., Liptser and Shiryayev [38, p. 66]) for each $\delta > 0$ and $\varepsilon > 0$ and sufficiently large N we have

$$\mathbb{P}\left\{ \sup_{0 \leqslant s \leqslant t} \left|A^p_{j,N}(s) - A_{j,N}(s)\right| > \delta N \right\} \leqslant \frac{\varepsilon}{\delta^2} + \mathbb{P}\left\{A_{j,N}(t) > \varepsilon N^2\right\}$$

and (5.25) follows. Both (5.24) and (5.25) imply that

$$\mathbb{P}\text{-}\lim_{N \to \infty} \frac{1}{N} M_{j,N}(t) = 0. \tag{5.26}$$

For our further purpose and for the sake of simplicity we shall use the notation

$$\Psi_t(X) = -\inf_{s \leqslant t} X(s)$$

and

$$\Phi_t(X) = X(t) + \Psi_t(X)$$

for any function $X(t)$ from the Skorohod space $\mathbf{D}$ with $X(0) = 0$. Then denoting also $x_{j,N}(t) = N^{-1}X_{j,N}(t)$, $x_j(t) = \mathbb{P}\text{-}\lim_{N\to\infty} x_{j,N}(t)$ $(j = 1, 2, \ldots, k)$, let us consider the set of equations

$$x_j(t) = \int_0^t \left(\lambda_j \left[1 - \sum_{l=1}^{k} \Phi_s(x_l)\right] - \mu_j\right) \mathrm{d}s. \tag{5.27}$$

There is a unique solution of (5.27) because of the Lipshitz condition

$$\sup_{s\leqslant t}\bigl|\Phi_s(X) - \Phi_s(Y)\bigr| \leqslant 2\sup_{s\leqslant t}|X_s - Y_s|, \tag{5.28}$$

which is for $x_k(t)$ is

$$x_k(t) = \left(1 - \frac{\mu_k}{\lambda_k}\right)\bigl(1 - \mathrm{e}^{-\lambda_k t}\bigr). \tag{5.29}$$

Liptser and Kogan [31] proved also that for any fixed $t > 0$ and $\varepsilon > 0$

$$\lim_{N\to\infty} \mathbb{P}\Bigl\{\sup_{s\leqslant t}\bigl|x_{k,N}(s) - x_k(s)\bigr| \geqslant \varepsilon\Bigr\} = 0.$$

It also follows from (5.15), (5.16), (5.29) and Lipshitz condition (5.28) that

$$\mathbb{P}\text{-}\lim_{N\to\infty} \frac{1}{N} Q_{k,N}(t) \geqslant x_k = \left(1 - \frac{\mu_k}{\lambda_k}\right)\bigl(1 - \mathrm{e}^{-\lambda_k t}\bigr). \tag{5.30}$$

On the other words we proved that under assumptions $G_i(t) \geqslant 1 - \mathrm{e}^{-\lambda t}$ $(\lambda \leqslant g)$, $i = 1, 2, \ldots$, and $\lambda_j < \mu_j$ for $j = 1, 2, \ldots, k-1$ while $\lambda_k > \mu_k$,

$$C(t) \leqslant 1 - \left(1 - \frac{\mu_k}{\lambda_k}\right)\bigl(1 - \mathrm{e}^{-\lambda_k t}\bigr). \tag{5.31}$$

Analogously, if $G_i(t) \leqslant 1 - \mathrm{e}^{-\lambda t}$ $(\lambda \geqslant g)$, $i = 1, 2, \ldots$, and $\lambda_j < \mu_j$ for $j = 1, 2, \ldots, k-1$ while $\lambda_k > \mu_k$, we obtain

$$C(t) \geqslant 1 - \left(1 - \frac{\mu_k}{\lambda_k}\right)\bigl(1 - \mathrm{e}^{-\lambda_k t}\bigr). \tag{5.32}$$

Now let us consider the case $G_1(t) = G_2(t) = \cdots = G_N(t) = G(x) = 1 - \mathrm{e}^{-\lambda t}$ $(\lambda = g)$. We have

$$C(t) = 1 - \left(1 - \frac{\mu_k}{\lambda_k}\right)\bigl(1 - \mathrm{e}^{-\lambda_k t}\bigr). \tag{5.33}$$

Then, after algebraic transformations from (5.14) we obtain the multi-dimensional variant of the known result by Kogan and Liptser [31]:

$$\mathbb{P}\{Q_1(t)=n_1,\, Q_2(t)=n_2, \ldots,\, Q_{k-1}(t)=n_{k-1}\} = \prod_{j=1}^{k-1}\left[\left(\frac{\lambda_j C(t)}{\mu_j}\right)^{n_j}\left(1-\frac{\lambda_j C(t)}{\mu_j}\right)\right],$$

$$n_j \geqslant 0,\ \ j = 1, 2, \ldots, k-1,$$

where $C(t)$ is defined by (5.33).

6. Concluding remarks

In previous sections 2, 3 and 5 we showed as the up- and down-crossing method enables us to prove that, as $N \to \infty$, the joint limiting queue-length distributions are calculated in closed product form. In the light usage regime case, considered in second and third sections, as $N \to \infty$, the network is decomposed to independent single-server queueing systems. In the case where the kth satellite node operates in heavy usage regime, there is the correlation between nodes because of the bottleneck. Nevertheless, in this case the network is decomposed too.

The technique of present paper permits us to develop the results for the cases where k satellite nodes are many server stations with finite or infinite admissible queue-lengths. In practice this case is related to computer networks with the same star topology, where we take into account that client computers work in multitask regime.

Examples below consider the case of k many server satellite stations with refusals and with infinite admissible queue-lengths. The first example is useful in the cases of queueing network with blocking to estimate the limiting probability that a number of tasks in a client station exceeds the admissible level of that number of tasks. The second example is useful in the case of queueing network without any assumption on blocking, however here it is interesting to study the case where there is a bottleneck amongst the client stations.

(1) Suppose that k satellite nodes are many server stations with refusals, consisting of $r_1, r_2, \ldots, r_k$ servers, respectively. If a customer, arriving to one of stations, meets all servers busy, he returns to the hub and joins it queue. All servers of each node are numbered, and a customer occupies a free server with the smallest number. All other assumptions of the model are the same as in the single servers case considered in the paper.

The limiting as $N \to \infty$ probability that the first customer, arriving after time t to satellite node j, occupies the server n_j is calculated as

$$\mathbb{P}\{Q_j(t) = n_j\} = P_{n_j}(\mu_j), \quad n_j = 1, 2, \ldots, r_j,$$

$$P_{n_j}(\mu_j) = \left[\sum_{i=0}^{n_j}\binom{n_j}{i}\prod_{l=1}^{i}\frac{1-\theta_l(\mu_j)}{\theta_l(\mu_j)}\right]^{-1},$$

$$\theta_l(\mu_j) = \frac{p_j \widehat{G}(\mu_j l)}{1 - (1 - p_j)\widehat{G}(\mu_j l)}.$$

Note, that $P_m(\mu_j)$ is the steady state loss probability for an GI/M/m/0 queueing system (see, e.g., Takacs [54]), where the distribution of interarrival time is

$$\sum_{i=0}^{\infty} p_j(1 - p_j)^i G^{*i+1}(x),$$

and the service time parameter is μ_j.

(2) Suppose that k satellite nodes are many server queueing systems with infinitely admissible queue-lengths, and $r_1, r_2, \dots, r_k$ servers are at satellite nodes $1, 2, \dots, k$, respectively. Let $Q_j(t)$ denote the number of customers at satellite node j at time t. Under the same assumptions as in the case of single-server queueing stations, operating whole in light usage regime, for the limiting joint distribution of the numbers of customers in queue, as $N \to \infty$, we have:

$$\mathbb{P}\{Q_1(t) = n_1, Q_2(t) = n_2, \dots, Q_k(t) = n_k\} = \prod_{j=1}^{k} P_{n_j}(\mu_j),$$

where $P_{n_j}(\mu_j)$ is the steady state probability that the number of customers in the system is equal to n_j for an GI/M/r_j/∞ queueing system with the interarrival time distribution

$$\sum_{i=0}^{\infty} p_j(1 - p_j)^i G^{*i+1}(x)$$

and exponentially distributed service times with parameter μ_j. For explicit formulae for those probabilities see, e.g., Borovkov [8], Takacs [54].

If the kth satellite node operates in heavy usage regime, and all other satellite nodes operate in light usage regime, the behavior of the system is the same as in the case of single-server satellite stations, i.e., the limiting joint distribution has a representation analogous to respective case of single-server satellite stations, although the explicit formula here is very difficult to observe and its obtaining is cumbersome even in the case of Markovian network.

Acknowledgement

The author thanks the anonymous referee for some comments.

References

[1] V.M. Abramov, *Investigation of a Queueing System with Service Depending on Queue-Length* (Donish, Dushanbe, 1991) (in Russian).

[2] V.M. Abramov, On the asymptotic distribution of the maximum number of infectives in epidemic models with immigration, J. Appl. Probab. 31 (1994) 606–613.

[3] V.M. Abramov, Queueing system with autonomous service: Up- and down-crossings approach, submitted for publication (2000).

[4] V.M. Abramov, A large closed queueing network with autonomous service and bottleneck, Queueing Systems Theory Appl. 35 (2000) 23–54.

[5] V.M. Abramov, Inequalities for the $\mathrm{GI}/\mathrm{M}/1/n$ loss system, J. Appl. Probab. 38 (2001) to appear.

[6] S.V. Anulova and R.Sh. Liptser, Diffusion approximation for the processes with normal reflection, Theory Probab. Appl. 35 (1990) 413–423.

[7] A.T. Bharucha-Reid, *Elements of the Theory of Markov Processes and Their Applications* (McGraw-Hill, New York, 1960).

[8] A.A. Borovkov, *Stochastic Processes in Queueing Theory* (Springer, New York, 1976).

[9] A.A. Borovkov, Limit theorems for networks of queues I, II, Theory Probab. Appl. 31 (1986) 413–427; 32 (1987) 257–272.

[10] P.H. Brill and M.J.M. Posner, Level crossing in point processes applied to queues: Single-server case, Oper. Res. 25 (1977) 662–674.

[11] P.H. Brill and M.J.M. Posner, The system point method in exponential queues: A level crossing approach, Math. Oper. Res. 6 (1981) 31–49.

[12] H. Chen and A. Mandelbaum, Discrete flow networks: Bottleneck analysis and fluid approximations, Math. Oper. Res. 16 (1991) 408–446.

[13] H. Chen and A. Mandelbaum, Discrete flow networks: Diffusion approximations and bottlenecks, Ann. Probab. 19 (1991) 1463–1519.

[14] J.W. Cohen, On up- and down-crossings, J. Appl. Probab. 14 (1977) 405–410.

[15] J.G. Dai, On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models, Ann. Appl. Probab. 5 (1994) 49–77.

[16] C. Dellacherie, *Capacites et Processus Stochastiques* (Springer, Berlin, 1972).

[17] B. Gnedenko and I.N. Kovalenko, *An Introduction to the Queueing Theory* (Israel Program for Scientific Translations, Jerusalem, 1968).

[18] D.L. Iglehart and W. Whitt, Multi-chanel queues in heavy traffic, I, Adv. Appl. Probab. 2 (1970) 150–177.

[19] D.L. Iglehart and W. Whitt, Multi-chanel queues in heavy traffic, II, Adv. Appl. Probab. 2 (1970) 355–369.

[20] G.I. Ivchenko, V.A. Kashtanov and I.N. Kovalenko, *Theory of Queues* (Vysshaya Shkola, Moscow, 1982) (in Russian).

[21] J. Jacod and A.N. Shiryayev, *Limit Theorems for Stochastic Processes* (Springer, Berlin, 1987).

[22] G.I. Kalmykov, On the partial ordering of one-dimensional Markov processes, Theory Probab. Appl. 7 (1962) 456–459.

[23] S. Karlin and H.G. Taylor, *A First Course in Stochastic Processes,* 2nd edn. (Academic Press, New York, 1975).

[24] H. Kaspi and A. Mandelbaum, Regenerative closed queueing networks, Stochastic Stochastic Rep. 39 (1992) 239–258.

[25] H. Kaspi and A. Mandelbaum, On Harris recurrence in continuous time, Math. Oper. Res. 19 (1994) 211–222.

[26] J. Keilson, *Markov Chain Models – Rarity and Exponentiality* (Springer, Heidelberg, 1979).

[27] G.P. Klimov, *Stochastical Service Systems* (Nauka, Moscow, 1966) (in Russian).

[28] C. Knessl and C. Tier, Asymptotic expansions for large closed queueing networks, J. Assoc. Comput. Mach. 37 (1990) 144–174.

[29] Y. Kogan, Another approach to asymptotic expansions for large closed queueing networks, Oper. Res. Lett. 11 (1992) 317–321.

[30] Y. Kogan and A. Birman, Asymptotic analysis of closed queueing networks with bottlenecks, in: *Proc. Internat. Conf. on Perfomance of Distributed Systems and Integrated Communication Networks*, eds. T. Hasegawa, H. Takagi and Y. Takahashi (Kyoto, 1991) pp. 237–252.

[31] Y. Kogan and R.Sh. Liptser, Limit non-stationary behavior of large closed queueing networks with bottlenecks, Queueing Systems Theory Appl. 14 (1993) 33–55.

[32] Y. Kogan, R.Sh. Liptser and M. Shenfild, State dependent Benes buffer model with fast loading and output rates, Ann. Appl. Probab. 5 (1995) 97–120.

[33] Y. Kogan, R.Sh. Liptser and A.V. Smorodinskii, Gaussian diffusion approximation of closed Markov model of computer networks, Problems Inform. Transmission 22 (1986) 38–51.

[34] E.V. Krichagina, Asymptotic analysis of queueing networks (martingale approach), Stochastic Stochastic Rep. 40 (1992) 43–76.

[35] E.V. Krichagina, R.Sh. Liptser and A.A. Puhalskii, Diffusion approximation for system with arrivals depending on queue and with general service, Theory Probab. Appl. 33 (1988) 114–124.

[36] A.J. Lemoine, Networks of queues – a survey of weak convergence results, Manag. Sci. 24 (1978) 1175–1193.

[37] R.Sh. Liptser and A.N. Shiryayev, *Statistics of Random Processes*, Vols. I, II (Springer, Berlin, 1977, 1978).

[38] R.Sh. Liptser and A.N. Shiryayev, *Theory of Martingales* (Kluwer, Dordrecht, 1989).

[39] A. Mandelbaum and W. Massey, Strong approximations for time-dependent queues, Math. Oper. Res. 20 (1995) 33–64.

[40] A. Mandelbaum, W.A. Massey and M.I. Reiman, Strong approximations for Markovian service networks, Queueing Systems Theory Appl. 30 (1998) 149–201.

[41] A. Mandelbaum and G. Pats, State dependent stochastic networks. Part I. Approximations and applications with continuous diffusion limits, Ann. Appl. Probab. 8 (1998) 569–646.

[42] J. McKenna, D. Mitra and K.G. Ramakrishnan, A class of closed Markovian queueing networks: Integral representation, asymptotic expansions and generalizations, Bell Syst. Technical J. 60 (1981) 599–641.

[43] B. Pittel, Closed exponential networks of queues with saturation: the Jackson-type stationary distribution and its asymptotical analysis, Math. Oper. Res. 6 (1979) 357–378.

[44] Y.V. Prohorov, Transient phenomena in queueing processes, Litovsk. Math. Sb. 3 (1963) 199–205 (in Russian).

[45] M.I. Reiman, Open queueing networks in heavy traffic, Math. Oper. Res. 9 (1984) 441–458.

[46] M.I. Reiman and B. Simon, A network of priority queues in heavy traffic: One bottleneck station, Queueing Systems Theory Appl. 6 (1990) 33–58.

[47] M. Shaked and J.G. Shanthikumar, *Stochastic Orders and Their Applications* (Academic Press, Boston, 1994).

[48] J.G. Shanthikumar, Some analysis of the control of queues using level crossing of regenerative processes, J. Appl. Probab. 17 (1980) 814–821.

[49] J.G. Shanthikumar, Level crossing analysis of priority queues and a conservation identity for vacation models, Naval Res. Logist. Quat. 36 (1989) 797–806.

[50] J.G. Shanthikumar and M.J. Chandra, Application of level crossing analysis to discrete state processes in queueing systems, Naval Res. Logist. Quat. 29 (1982) 593–608.

[51] A.N. Shiryayev, *Probability* (Springer, Berlin, 1984).

[52] A.V. Skorohod, Stochastic equations for diffusion processes in a bounded region, Theory Probab. Appl. 6 (1961) 264–274.

[53] D. Stoyan, *Comparison Methods for Queues and Other Stochastic Models* (Wiley, Chichester, 1983).

[54] L. Takacs, *Introduction to the Theory of Queues* (Oxford Univ. Press, New York, 1962).

[55] H. Tanaka, Stochastic differential equations with reflecting boundary condition in convex regions, Hiroshima Math. J. 9 (1979) 163–177.

[56] H. Thorisson, From coupling to shift-coupling, Theory Probab. Appl. 37 (1992) 105–112.

[57] W. Whitt, Open and closed models for networks of queues, AT&T Bell Lab. Technical J. 63 (1984) 1911–1979.