A GENERAL MULTISERVER STATE-DEPENDENT QUEUEING SYSTEM

EVSEY MOROZOV

Abstract. The work studies a general multiserver queue in which the service time of an arriving customer and the next interarrival period may depend on both the current waiting time and the server assigned to the arriving customer. Stability of the system is proved under general assumptions on the predetermined distributions describing the model. The proof exploits a combination of the Markov property of the workload process with a regenerative property of the process. The key idea leading to stability is a characterization of the limit behavior of the forward renewal process generated by regenerations. Extensions of the basic model are also studied.

1. Introduction

Including various dependencies in a queueing model to reflect real-life effects makes it more realistic. The aim of this work is to establish stability conditions for a multiserver queue in which the service time of an arriving customer, and the next interarrival period are dependent on the customer's waiting time and its assigned server. This model contains a wide class of state-dependent queues. The main motivation of this work is to present a wider framework in which the regenerative approach combined with the Markov property of the workload process are instrumental in the stability analysis of state-dependent queues. The general model we consider illustrates this framework, but it also may have an independent interest because some known and new systems can be analyzed as special cases. We will not give an extensive description of state-dependent queues, but only mention a few examples.

The simplest example is a workload-dependent model that uses a rejection rule for arriving customers. In this case, the impatience of a customer reflects a simple dependence between workload and service time in that the service time equals zero if the workload exceeds a random time. A general form of this rejection rule for an $M/G/1$ queue is presented in [1], where a state-dependent Poisson arrival rate is also discussed. (An explicit steady-state solution for a Markovian model with bounded waiting time was obtained earlier in [13].)

Another approach is used in [3], which considers systems with workload - dependent arrival rates and service speeds. The results give explicit characterizations of

Key words and phrases. Multiserver queue; state-dependent queue; dependent interarrival time, service time and waiting time; Lindley-type recursion; regeneration; stability analysis.

the steady-state workload in $M/G/1$ type queues and in a general $G/G/1$ queue with state-dependent release. In particular, the latter model generalizes several well-known relations for the workload at arbitrary epochs and embedded instants in the standard $GI/G/1$ queue. A wide class of state-dependent networks under exponential assumptions is considered in [34].

In [10], Lindley's recursion for the waiting time sequence is used to establish stability conditions for a single-server queue where the distribution of the (discrete) service time of an arriving customer is determined by his waiting time. The paper [41] is devoted to the stability analysis of a wide class of state-dependent single-server queues described by a modified Lindley's recursion. The single-server queues with Poisson input and dependence between waiting time and service time given by conditional distributions is considered in [32]. We also mention earlier related works [9, 10, 14, 33].

An interesting interpretation of a reliability/maintenace model as queueing system with a dependence between service time and the next interarrival time is considered in [7]. There are related papers on ruin models, where dependence between claim sizes and claim intervals is allowed. For instance, in [2] the time between two claim occurrences depends on the previous claim size. (This work also contains a list of relevant references.) The key element of the analysis in the mentioned works is the Markovian property of the waiting time process, which allows one to describe the dependence by (a modified) Tákacs integro-differential equation, when the arrival process is Poisson (or state-dependent Poisson). A rather complete bibliography on state-dependent queues satisfying Lindley-type recursions can be found in [8, 40].

Our study concerns a general m-server queue in which the service time of an arriving customer and next interarrival period may depend on both the current waiting time and the server assigned to the arriving customer. A wide class of (conditional) distributions describing the model is considered in such a way that the workload process retains the Markov property. In general, such a complicated system includes as specific cases both some well-studied and less known systems. (This topic is discussed in section 5.) In such a general setting, we do not obtain explicit formulas for the very complicated steady-state performance parameters. Nevertheless, we are able to develop stability analysis under the assumption that the workload Markov process has classical regenerations appearing when arriving customers meet an empty system.

The stability analysis presented in this work is based on renewal theory and a characterization of the limiting behavior of the forward renewal time in the process generated by regenerations of the workload process. This approach turns out to be effective in the stability analysis of many queues including general retrial queue [28], and also multiserver system with non-identical servers [25]. This characterization gives a straightforward way to establish positive recurrence (finiteness of mean regeneration period) of the embedded renewal process of regenerations in the terms of given distributions. (In a general form, this approach is presented in [25].)

The main advantage of the presented approach is that, instead of the direct proof of the finiteness of the mean regeneration period, that is typically a difficult problem, we show that the forward renewal time in the process generated by regenerations does not go to infinity in probability. The latter condition is typically much easier to verify. This verification consists of two steps: i) using *negative drift assumptions* to show that the workload process does not go to infinity; ii) using *regeneration assumptions* to show that the forward regeneration time does not go to infinity. Then the mentioned characterization immediately implies positive recurrence of the basic process. Although the workload process is Markovian, which is useful for intermediate steps of analysis, the presented method also works successfully outside of Markovian models. A recent example is in the paper [28]. Also recent review paper [30] contains detailed description of the method with applications to various known and new models. (The latter paper is based on the report [29].)

In the final section, we discuss in brief a possible relaxation of assumptions implying more general *one-dependent* (or *weak*) regenerations, in which case the workload process turns out to be a Harris Markov chain. In this case a dependence between two adjacent cycles is allowed while the cycle lengths stay i.i.d. A detailed presentation of this topic can be found, for instance, in [1, 16, 20, 39], while some specific aspects related to one-dependent regeneration of queueing processes are presented in [22]–[25], [31], [35]-[37]. A review paper [17] gives sufficient and modern introduction to various techniques applying in stability analysis of queues. Considerable attention in [17] is devoted to a popular and effective stability analysis based on the fluid approximation, which uses the i.i.d. assumptions [11, 12], but it is inapplicable to the models in the present work. The work [17] also discusses a relation between Harris chains and the theory of *renovating events* (developed in [5]).

This work is organized as follows. Section 2 describes our model in detail. Section 3 contains the proof of the main result. An extension of the basic model is considered in Section 4. Finally, Section 5 discusses the model assumptions and compares them with stability assumptions for the known models, including classical multiserver systems. We also discuss a possible relaxation of the assumptions leading to one-dependent regeneration. The work is a considerable revision of [26].

2. Description of the model

Consider a first-come-first-served m-server queue. Let $\{t_n\}$ be the arrival times of customers, $\tau_n = t_{n+1} - t_n$, and let S_n be the service time of customer n, $n \geq 1$, $t_1 = 0$. Denote the (unfinished) workload at server i at instant t_n^- by

$W_n^{(i)}$, $i = 1, \dots, m$. Then $W_n = \min_{1 \le i \le m}\{W_n^{(i)}\}$ is the waiting time of customer n. Unlike conventional models, we allow the random variables τ_n, S_n, W_n to be dependent. To describe these dependencies, we introduce a sequence of random variables $\sigma_n \in \{1, \dots, m\}$, where σ_n is the server assigned to customer n. We assume that on the event

$$\mathsf{E}_n(dy, i) = \{W_n \in dy,\, \sigma_n = i\}$$

the interarrival time τ_n and service time S_n are (conditionally) independent (and independent of everything else) and are distributed as some random variables $\tau_i(y)$ and $S_i(x)$, with given distributions

$$A_{i,y}(x) = \mathsf{P}(\tau_i(y) \le x),\ G_{i,y}(x) = \mathsf{P}(S_i(y) \le x), \tag{1}$$

respectively, $x \ge 0$, $y \ge 0$, $n \ge 1$, $i = 1, \dots, m$. We also use the notations $F_n(x) = \mathsf{P}(W_n \le x)$, and $(x)^+ = \max(0, x)$.

We adopt the major assumption that the assignment of servers to arriving customers regenerates each time the system becomes empty. The assignment is arbitrary for the case when a minimal value W_n is attained by several servers simultaneously. Denote by ν_n the number of customers in the system at instant t_n^-. The customers that meet an empty system are

$$\beta_n = \min(k > \beta_{n-1} : \nu_k = 0),\ n \ge 1\ (\beta_0 = 0). \tag{2}$$

Then β_n, when they are finite, constitute the classical regeneration epochs for the processes $\{\nu_n\}$, $\{W_n\}$ with the i.i.d. regeneration periods $\beta_n - \beta_{n-1}$, $n \ge 2$, distributed as a random variable β. In the zero-delayed case, $\beta_1 = \beta$ and $\nu_1 = t_1 = 0$.

Define the forward renewal time $\beta(n) = \min\{\,\beta_k - n : \beta_k - n > 0\,\}$ at instant $n \ge 0\,(\beta(0) = \beta_1)$. In the zero-delayed case $\beta(0) = \beta$. The key to our stability result is the following dichotomy describing the asymptotic behavior of $\beta(n)$ [15]. For arbitrary initial value β_1,

$$\beta(n) \Rightarrow \infty \text{ if and only if } \mathsf{E}\beta = \infty. \tag{3}$$

($\Rightarrow$ stands for the convergence in probability.) To establish (stability) condition $\mathsf{E}\beta < \infty$, it is sufficient to show that $\beta(n) \not\Rightarrow \infty$, that is,

$$\inf_k \mathsf{P}\left(\beta(n_k) \le L\right) \ge \varepsilon,$$

for some constants $L < \infty$, $\varepsilon > 0$ and a non-random sequence $n_k \to \infty$. If moreover, the regeneration period is aperiodic, then the stationary distribution $\lim_{n\to\infty} \mathsf{P}(\nu_n \in \cdot)$ exists. (In this case the stationary distribution of the vector workload process also exists.)

We prove the condition $\mathsf{E}\beta < \infty$ in two settings. In the first case, for each i, we allow any number of different service time distributions $G_{i,y}$, but exclude heavy-tailed ones. In contrast, in the second setting we assume that the number of different distributions $G_{i,y}$ is finite. Note that we consider the zero-delayed case only.

3. Stability analysis

First, we establish the stability conditions with no restriction on the number of different (conditional) service time distributions $G_{i,y}$.

Theorem 1. *Assume the following conditions hold for each $i = 1, \dots, m$:*

$$\sup_{x\ge 0} \mathsf{E}\tau_i(x) < \infty, \ \ \sup_{x\ge 0} \mathsf{E}S_i(x) < \infty; \tag{4}$$

$$\limsup_{x\to\infty} \mathsf{E}\Big(S_i(x) - m\tau_i(x)\Big) < 0; \tag{5}$$

there exist $\varepsilon > 0$ and a finite constant T such that

$$\inf_{x\ge 0}\inf_{y\ge 0} \frac{G_{i,y}(x+T) - G_{i,y}(x)}{1 - G_{i,y}(x)} \ge \varepsilon; \tag{6}$$

for each $x \ge 0$, there exists a constant $\delta_i(x) > 0$ such that

$$\inf_{y\le x} \mathsf{P}\Big(\tau_i(y) > \delta_i(x) + S_i(y)\Big) := \rho_i(x) > 0. \tag{7}$$

Then

$$\mathsf{E}\beta < \infty. \tag{8}$$

Proof. Define

$$\begin{aligned} V_n &= \sum_{i=1}^m W_n^{(i)}, \quad \rho(x) = \min_i \rho_i(x), \ \delta(x) = \min_i \delta_i(x), \\ S_n^{(i)} &= S_n I_{\{\sigma_n = i\}}, \ i = 1, \dots, m; \ n \ge 1, \end{aligned}$$

where I is the indicator function. Observe that $V_1 = 0$, $\rho(x) > 0$ for each $x \ge 0$ and $\sum_{i=1}^m S_n^{(i)} = S_n$, $n \ge 1$. Instead of Kiefer-Wolfowitz recursion, we use the following relations

$$W_{n+1}^{(i)} = (W_n^{(i)} + S_n^{(i)} - \tau_n)^+, \ i = 1, \dots, m, \ n \ge 1.$$

Denoting the increments $\Delta_n = V_{n+1} - V_n$, we have

$$\begin{aligned} \Delta_n &= \sum_{i=1}^m (W_n^{(i)} + S_n^{(i)} - \tau_n)^+ - \sum_{i=1}^m W_n^{(i)} \\ &= S_n - m\tau_n + \sum_{i=1}^m (\tau_n - W_n^{(i)} - S_n^{(i)})^+, \ n \ge 1. \end{aligned} \tag{9}$$

On the event $\mathtt{E}_n(dx, i)$ denote the difference Δ_n by $\Delta(x, i)$. Also note that

$$W_n^{(i)} = x \le W_n^{(j)}, \ j \ne i \quad \text{on the event } \ \mathtt{E}_n(dx, i).$$

By (9) the following upper bound holds:

$$\begin{aligned}\Delta(x,i) &= S_i(x) - m\tau_i(x) + (\tau_i(x) - x - S_i(x))^+ + \sum_{j\neq i}(\tau_i(x) - W_n^{(j)})^+ \\ &\le S_i(x) + m((\tau_i(x) - x)^+ - \tau_i(x)) := \alpha_i(x),\ n \ge 1. \qquad (10)\end{aligned}$$

We first show, using assumption (5), that the waiting time $W_n \not\Rightarrow \infty$. It follows from (4), (10) that

$$\begin{aligned}\mathsf{E}\Delta_n &\le \int_{x\ge 0}\sum_{i=1}^{m}\mathsf{E}\alpha_i(x)\mathsf{P}(\mathtt{E}_n(dx,i)) \\ &\le \max_i \sup_{x\ge 0}\mathsf{E}\alpha_i(x) := R < \infty,\end{aligned}$$

and thus for each $x \ge 0$,

$$\mathsf{E}\Delta_n \le RF_n(x) + \max_i \sup_{y>x}\mathsf{E}\alpha_i(y)(1 - F_n(x)),\ n \ge 1.$$

Assume that $W_n \Rightarrow \infty$, then we obtain by (5) that $\limsup_{n\to\infty}\mathsf{E}\Delta_n < 0$. This easily implies $\max_n \mathsf{E}V_n < \infty$. Together with inequality $W_n \le V_n$ this contradicts the assumption, and thus $W_n \not\Rightarrow \infty$. Because of this property, there exist constants $\varepsilon_0 > 0$, $T_0 < \infty$ and a non-random sequence $n_k \to \infty$ such that

$$\inf_k \mathsf{P}(W_{n_k} \le T_0) \ge \varepsilon_0 > 0. \tag{11}$$

Consider the events

$$\mathtt{B}_n = \{W_n \le T_0\},\ \mathtt{A}_n = \{S_n + \delta(T_0) \le \tau_n\},\ n \ge 1,$$

and denote $r_0 = \lceil T_0/\delta(T_0)\rceil$. Note that for any $i \ge 0$,

$$\mathtt{B}_{n_k} \cap \cap_{p=0}^{i}\mathtt{A}_{n_k+p} \subseteq \Big\{W_{n_k+i} \le (T_0 - i\delta(T_0))^+\Big\},$$

that is, each occurrence of an event $\mathtt{A}_k$ decreases residual work not less than by $\delta(T_0)$ (as long as all servers are busy). Thus we conclude that

$$\mathtt{B}_{n_k} \cap \cap_{i=0}^{r_0}\mathtt{A}_{n_k+i} \subseteq \Big\{W_{n_k+p} = 0 \text{ for some } p \in [0, r_0]\Big\}.$$

Fix some n_k satisfying (11) and observe that

$$\begin{aligned}\mathsf{P}\Big(\mathtt{B}_{n_k} \cap \mathtt{A}_{n_k}\Big) &= \int_0^{T_0}\sum_{i=1}^{m}\mathsf{P}\Big(S_i(x) + \delta(T_0) \le \tau_i(x)\Big)\mathsf{P}(\mathtt{E}_{n_k}(dx,i)) \\ &\ge \min_i \inf_{x\le T_0}\mathsf{P}\Big(S_i(x) + \delta_i(T_0) \le \tau_i(x)\Big)\mathsf{P}(\mathtt{B}_{n_k}) \\ &= \min_i \rho_i(T_0)\mathsf{P}(\mathtt{B}_{n_k}) \ge \rho(T_0)\varepsilon_0.\end{aligned}$$

It is easy to conclude that

$$\begin{aligned} \mathsf{P}\Big(W_{n_k+p} = 0 \ \text{ for some } \ p \in [0, r_0]\Big) &\geq \mathsf{P}\Big(\mathtt{B}_{n_k} \cap \cap_{i=0}^{r_0} \mathtt{A}_{n_k+i}\Big) \\ &\geq \varepsilon_0[\rho(T_0)]^{r_0} := \delta_0 > 0. \end{aligned} \tag{12}$$

It now follows that $\limsup_{n\to\infty} \mathsf{P}(W_n = 0) \geq \delta_0$, and there exists a non-random sequence $z_k \to \infty$ such that

$$\inf_k \mathsf{P}(W_{z_k} = 0) \geq \delta_0. \tag{13}$$

Next, we show that $\beta(n) \not\Rightarrow \infty$ under assumptions (6), (7). Consider the (right-continuous) unfinished $\tilde{S}_i(t)$ and attained $\bar{S}_i(t)$ service time, respectively, at each server $i = 1, \dots, m$, at instant t, and let

$$\tilde{S}_i(t_{z_k}) = \tilde{S}^{(i)}_{z_k},\ \bar{S}_i(t_{z_k}) = \bar{S}^{(i)}_{z_k}\ \ (\tilde{S}^{(i)}_{z_k} = \bar{S}^{(i)}_{z_k} = 0 \ \text{ for an empty server}),\ \ n \geq 1.$$

Let, for non-empty sever i, $n_i(z_k)$ be the number of the customer being served at instant t_{z_k}. Fix some z_k satisfying (13) and let T satisfy (6). Denote

$$\begin{aligned} &y = (y_1, \dots, y_m),\ x = (x_1, \dots, x_m),\ \tilde{S}_{z_k} = (\tilde{S}^{(1)}_{z_k}, \dots, \tilde{S}^{(m)}_{z_k}), \\ &B_T = [0,\, T] \times \cdots \times [0,\, T] \in \mathsf{R}^m_+, \end{aligned}$$

and introduce the events

$$\mathtt{C}_k(dx, dy) = \Big\{W_{z_k} = 0,\ \bar{S}^{(i)}_{z_k} \in dx_i,\ W_{n_i(z_k)} \in dy_i,\ i = 1,\, \dots, m\Big\},$$

where we put $W_{n_i(z_k)} = 0$ if server i is free at instant t_{z_k}. Then it follows from (6), (13) that

$$\begin{aligned} \mathsf{P}(\mathtt{F}(z_k)) &:= \mathsf{P}\Big(W_{z_k} = 0,\ \tilde{S}_{z_k} \in B_T\Big) \\ &= \int_{x\geq 0}\int_{y\geq 0} \mathsf{P}\Big(\tilde{S}_{z_k} \in B_T \,|\, \mathtt{C}_k(dx, dy)\Big)\mathsf{P}(\mathtt{C}_k(dx, dy)) \\ &\geq \Big[\min_i \inf_{x_i\geq 0,\, y_i\geq 0} \mathsf{P}\Big(S_i(y_i) \leq T + x_i |\, S_i(y_i) \geq x_i\Big)\Big]^m \mathsf{P}(W_{z_k} = 0) \geq \varepsilon^m \delta_0, \end{aligned}$$

where, in the last line, we use the equality

$$\mathsf{P}\Big(S_i(y_i) \leq T + x_i |\, S_i(y_i) \geq x_i\Big) = \frac{G_{i,\,y_i}(x_i + T) - G_{i,\,y_i}(x_i)}{1 - G_{i,\,y_i}(x_i)}$$

and assumption (6). Note that each event $\mathtt{D}_j = \{\tau_j > \delta(0) + S_j\}$ decreases residual work at each server not less than by $\delta(0)$. Denoting $\lceil T/\delta(0)\rceil = r$, we obtain from (7) (and conditioning in the same way as in the derivation of (14)) that

$$\begin{aligned} \mathsf{P}(\beta(z_k) \leq r) &= \mathsf{P}(\nu_{z_k+p} = 0,\ \text{ for some } \ p \in [z_k, z_k + r]) \\ &\geq \mathsf{P}\Big(\mathtt{F}(z_k) \cap \cap_{j=z_k}^{z_k+r} \mathtt{D}_j\Big) \geq [\rho(0)]^r \varepsilon^m \delta_0, \end{aligned} \tag{14}$$

and hence, $\beta(n) \not\Rightarrow \infty$. ■

Note that the proof of (13) does not use assumption (6) and that $\{W_n = 0\} = \{\nu_n = 0\}$. Thus,

$$\inf_k \mathsf{P}(\beta(z_k - 1) = 1) \geq \delta_0$$

and we obtain the following statement.

Corollary 1. *If $m = 1$, then the statement of Theorem 1 holds without assumption (6).*

Remark 1. It follows from (7) that

$$\mathsf{P}(\tau_i(0) > S_i(0)) > 0, \ i = 1, \dots, m,$$

and thus regeneration period β is aperiodic. In particular, the stationary distribution of the m-dimensional workload process $\lim_{n\to\infty} \mathsf{P}\Big(\Big(W_n^{(1)}, \dots, W_n^{(m)}\Big) \in \cdot\Big)$ exists.

Assumption (6) is the most restrictive and introduced to guarantee the tightness of the residual service time process. (The importance of the tightness for the stability analysis is discussed in [24].) Condition (6), in particular, does not hold for long/heavy-tailed distributions, which are discussed say, in [38]. For instance, if (for some i, y) the service time $S_i(y)$ is Pareto with exponent $\alpha_i(y) \in (0, \infty)$, then

$$G_{i,y}(x + T) - G_{i,y}(x) = o(1 - G_{i,y}(x)), \ \ x \to \infty$$

for any constant $T < \infty$, and this contradicts (6).

We now consider the system with (6) replaced by other conditions that ensure the required tightness. Assume that, for each server i, there exist constants

$$0 = b_0^{(i)} < b_1^{(i)} < \dots < b_{M_i}^{(i)}$$

and distribution functions $\hat{G}_{i,k}$, $k = 0, \dots, M_i$, such that $\min_{i,k} \hat{G}_{i,k}(0) < 1$ and

$$G_{i,y} \ = \ \hat{G}_{i,k}, \ \ y \in [b_k^{(i)}, b_{k+1}^{(i)}), \ \ k = 0, \dots, M_i \ (b_{M_i+1}^{(i)} = \infty). \tag{15}$$

Let, for each i, k, $\{S_k^{(i)}(n), \ n \geq 1\}$ be the i.i.d. sequence with distribution $\hat{G}_{i,k}$ and generic element $S_k^{(i)}$.

Assumption (15) seems to be less restrictive and more practical than (6) since allows only a finite number of switching of service time distribution. This simplifies considerably the state-dependence control mechanism governing the behavior of the system. In particular, on the event $\mathtt{E}_n(dy, i)$, the service time S_n is insensitive to change of the waiting time provided $W_n = y \geq b_{M_i}^{(i)}$. The proof of the following result in part is similar to the proof of Theorem 1, but it uses the Kiefer-Wolfowitz representation of the workload vector instead of the unordered workloads. Moreover, another idea is used to establish the tightness of the residual service time process.

Theorem 2. *Assume that the following conditions hold for each $i = 1, \dots, m:$*

$$\sup_{x\ge 0} \mathsf{E}\tau_i(x) < \infty, \quad \max_{1\le k\le M_i} \mathsf{E}S_k^{(i)} < \infty; \tag{16}$$

$$\mathsf{E}S_{M_i}^{(i)} < m \liminf_{x\to\infty} \mathsf{E}\tau_i(x); \tag{17}$$

there exists a constant $\delta >$ such that

$$\inf_{y\in[b_k^{(i)},\, b_{k+1}^{(i)})} \mathsf{P}(\tau_i(y) > \delta + S_k^{(i)}) := \rho_k^{(i)} > 0, \;\; k = 0, \dots, M_i. \tag{18}$$

Then $\mathsf{E}\beta < \infty$.

Proof. Assume that $m > 1$, denote by $\hat W_n^{(i)}$ the i-th smallest residual work at instant t_n (among m components), and consider Kiefer-Wolfowitz sequence

$$\hat W_n = (\hat W_n^{(1)}, \dots, \hat W_n^{(m)}),\; n \ge 1.$$

Assume arbitrary fixed initial state $\hat W_1 = (x_1^{(1)}, \dots, x_1^{(m)})$ and denote

$$D_n = \hat W_n^{(m)} - \hat W_n^{(1)},\;\; A = (m-1)x_1^{(m)} - \sum_{i=1}^{m-1} x_1^{(i)}.$$

It is assumed that $S_n = S_k^{(i)}(n)$ on the event $\mathsf{E}_n(dy, i)$ provided

$$W_n \equiv \hat W_n^{(1)} = y \in [b_k^{(i)},\, b_{k+1}^{(i)}).$$

Also denote

$$\begin{aligned} Y_n &= \max\Big((m-1)S_n,\, (m-1)S_{n-1} - S_n, \\ &\qquad \dots, (m-1)S_1 - S_2 - \dots - S_n,\, A - S_1 - S_2 - \dots - S_n\Big); \\ \alpha(n) &= \max_{1\le i\le m}\, \max_{0\le k\le M_i} S_k^{(i)}(n);\;\; \beta(n) = \min_{1\le i\le m}\, \min_{0\le k\le M_i} S_k^{(i)}(n),\;\; n \ge 1. \end{aligned}$$

Note that $\{\alpha(n)\}$, $\{\beta(n)\}$ are i.i.d. (independent) sequences and that with probability 1 (w.p.1)

$$\beta(n) \le S_n \le \alpha(n),\;\; n \ge 1.$$

Since $\min_i \min_k \hat G_{i,k}(0) < 1$, then $\mathsf{E}\beta(1) > 0$. Following [24], we obtain

$$\mathsf{P}(Y_n \le x_1) \ge \mathsf{P}\Big(\alpha(i) \le \frac{x_1 + \sum_{k=1}^{i-1} \beta(k)}{m-1},\;\; i \ge 1\Big) \;\; \Big(\sum_{\emptyset} = 0\Big), \tag{19}$$

where x_1 is chosen in such a way that $x_1 \ge (m-1)x_m^{(1)}$. Since $\alpha(i)/i \to 0$ and, by the Strong Law of Large Numbers, $\sum_{k=1}^{i-1} \beta(k)/i \to \mathsf{E}\beta(1) > 0$, $i \to \infty$ w.p.1, it follows easily from (19) that the sequence $\{Y_n\}$ (and hence $\{D_n\}$) is tight. (The detailed proof of the tightness is based of the Kiefer-Wolfowitz "key" lemma [19] and its extension presented in [24].) As in Theorem 1, we use (16)–(18) to prove (13). (Recall that $W_n = \hat{W}_n^{(1)}$.) Thus, a constant $C < \infty$ exists such that

$$\inf_k \mathsf{P}(\hat{W}_{z_k}^{(1)} = 0,\, D_{z_k} \le C) \ge \delta_0/2.$$

Then denoting $\lceil C/\delta \rceil = r$, $\rho = \min_i \rho_0^{(i)}$ we obtain as in (14) (using the events $\{\tau_i(0) > S_0^{(i)} + \delta\}$) that

$$\mathsf{P}(\beta(z_k) \le r) \ge \delta_0 \rho^r/2 > 0. \quad \blacksquare$$

Remark 2. We stress that the statement of Theorem 2 holds for the zero-delayed case only, while the tightness of $\{D_n\}$ takes place for arbitrary initial state $\hat{W}_1$. Also instead of a common δ in (18) we could use different $\delta_i > 0$ and then put $\delta = \min_i \delta_i$.

4. An extension

Using the same notation, consider the following extension, which only differs from the basic model in that, on the event

$$\mathtt{D}_n(dx, i, dz) := \mathtt{E}_n(dx, i) \cap \{S_i(x) \in dz\},$$

the interval τ_n is distributed as a random variable $\tau_i(x, z)$ with a given distribution, $n \ge 1$. Also we denote $\tau_i(x, S_i(x))$ the interval τ_n on the event $\mathtt{E}_n(dx, i)$. Note that the β_n in (2) are classical regeneration points of the workload sequence because at each such point

$$S_{\beta_k} = S_{\sigma_{\beta_k}}(0), \;\; \tau_{\beta_k} = \tau_{\sigma_{\beta_k}}(0,\, S_{\sigma_{\beta_k}}(0))$$

and the sequence $\{\sigma_n\}$ also regenerates. Here is an extension of Theorem 1.

Theorem 3. *Assume that (6) and the following conditions hold for each $i = 1, \dots, m$:*

$$\sup_{x \ge 0} \mathsf{E} S_i(x) < \infty, \;\; \sup_{x \ge 0} \sup_{y \ge 0} \mathsf{E}\tau_i(x, y) < \infty; \tag{20}$$

$$\limsup_{x \to \infty} \mathsf{E}(S_i(x) - m\tau_i(x, S_i(x))) < 0; \tag{21}$$

for each $x \ge 0$ there exists a constant $\delta_i(x) > 0$ such that

$$\inf_{u \le x} \mathsf{P}(\tau_i(u,\, S_i(u)) > S_i(u) + \delta_i(x)) := \rho_i(x) > 0. \tag{22}$$

Then $\mathsf{E}\beta < \infty$. If $m = 1$ then assumption (6) is not necessary.

Proof. The proof of Theorem 1 allows us to simplify the following proof. Note that

$$\Delta(x, i) \le S_i(x) - m\tau_i(x, S_i(x)) + m\Big(\tau_i(x, S_i(x)) - x\Big)^+, \tag{23}$$

and that by (20),

$$\lim_{x\to\infty}\sup_{y\geq 0}\mathsf{E}(\tau_i(x,y)-x)^+=0.$$

Then (21), (23) imply the lower bound (11). Denote

$$\rho(x)=\min_i\rho_i(x),\ \delta(x)=\min_i\delta_i(x).$$

Then for any n_k, by (22),

$$\begin{aligned}&\mathsf{P}(W_{n_k}\leq T_0,\,\tau_{n_k}>\delta(T_0)+S_{n_k})\\ \geq\ &\sum_{i=1}^m\int_{u\leq T_0}\mathsf{P}(\tau_i(u,\,S_i(u))>S_i(u)+\delta_i(T_0))\mathsf{P}(\mathsf{E}_n(du,i))\\ \geq\ &\rho(T_0)\varepsilon_0.\end{aligned}$$

Denote $r_0=\lceil T_0/\delta(T_0)\rceil$ and note that

$$\{W_n\leq T_0\}\cap\{\tau_n>S_n+\delta(T_0)\}\subseteq\{W_{n+1}\leq(T_0-\delta(T_0))^+\},\ n\geq 1.$$

As in Theorem 1 one can show that a non-waiting customer arrives in the interval $[n_k,\,n_k+r_0]$ with a probability $\geq\varepsilon_0[\rho(T_0)]^{r_0}$, and thus (13) holds. Then exactly as in Theorem 1 we obtain (14). Then it is easy to show that for any z_k,

$$\begin{aligned}\mathsf{P}\Big(\mathtt{F}(z_k)\cap\{\tau_{z_k}>\delta(0)+S_{z_k}\}\Big)\ &\geq\ \min_i\mathsf{P}\Big(\tau_i(0,S_i(0))>S_i(0)+\delta_i(0)\Big)\mathsf{P}(\mathtt{F}(z_k))\\ &\geq\ \rho(0)\varepsilon^m\delta_0.\end{aligned}$$

The rest of proof is now obvious. Denoting $r=\lceil T/\delta(0)\rceil$ we note that a regeneration occurs in (any) interval $[z_k,\,z_k+r]$ with a probability $\geq[\rho(0)]^r\varepsilon^m\delta_0$. As above, the statement of theorem for the single-server case holds without assumption (6). ■

We note a connection between (22) and given distributions of $\tau_i(u,z)$ and $S_i(u)$ for each $i=1,\dots,m$:

$$\mathsf{P}(\tau_i(u,\,S_i(u))>S_i(u)+\delta_i(x))=\int_0^\infty\mathsf{P}(\tau(u,\,z)>z+\delta_i(x))\mathsf{P}(S_i(u)\in dz).$$

The proof of the following result is omitted since it is analogous to those of Theorems 2, 3.

Theorem 4. *Assume that (15) and the following conditions hold for each* $i=1,\dots,m:$

$$\sup_{x\geq 0}\sup_{y\geq 0}\mathsf{E}\tau_i(x,\,y)<\infty,\quad\max_{1\leq k\leq M_i}\mathsf{E}S_k^{(i)}<\infty; \tag{24}$$

$$\mathsf{E}S_{M_i}^{(i)}<m\liminf_{x\to\infty}\mathsf{E}\tau_i(x,\,S_{M_i}^{(i)}); \tag{25}$$

there exists a constant $\delta > 0$ *such that*

$$\inf_{y\in[b_k^{(i)},\, b_{k+1}^{(i)})} \mathsf{P}(\tau_i(y,\, S_k^{(i)}) > \delta + S_k^{(i)}) > 0,\ k = 0, \dots, M_i. \tag{26}$$

Then $\mathsf{E}\beta < \infty$.

5. Discussion of assumptions

In this section, we discuss our assumptions and compare them with the stability assumptions of classical models. Also, we describe some state-dependent models with various dependencies between random variables, and discuss possible weakening of assumptions leading to one-dependent regeneration.

First of all, note that for a standard $GI/G/m$ queue (that is for the i.i.d. case) with the interarrival time τ (with distribution A) and the service time S (with distribution G), the negative drift assumptions considered above take the well-known form $\lambda := 1/\mathsf{E}\tau < m/\mathsf{E}S := m\mu$. If we assume a dependence between service time and the assigned server only, then we obtain the system with non-identical servers (with service rates μ_i). Perhaps it is surprisingly, that in this case, our negative drift assumptions reduce to assumption $\lambda < m\mu_i$, $i = 1, \dots, m$, which is stronger than *minimal* requirement $\lambda < \sum_{i=1}^m \mu_i$ implying stability [21, 25].

Keeping assumption (15), one can obtain another realistic model assuming that a dependence between interarrival times and waiting time is expressed (for each server i) by a finite number $N_i + 1$ of distributions $\hat{A}_{i,k}$ such that (see (1))

$$A_{i,y} = \hat{A}_{i,k},\ y \in [a_k^{(i)},\, a_{k+1}^{(i)}),\ k = 0, \dots, N_i,$$

where $0 = a_0^{(i)} < a_1^{(i)} < \cdots < a_{N_i}^{(i)}$ are given constants ($a_{N_i+1}^{(i)} = \infty$). In this case the negative drift assumption becomes especially simple:

$$\mathsf{E}S_{M_i}^{(i)} < m\mathsf{E}\tau_{N_i}^{(i)},$$

where $\tau_{N_i}^{(i)}$ has distribution $\hat{A}_{i,N_i}$, $i = 1, \dots, m$. (An interesting case is $N_i \equiv M_i$, $a_k^{(i)} \equiv b_k^{(i)}$.)

Next, we discuss regeneration assumptions (which are reduced to $\mathsf{P}(\tau > S) > 0$ for the i.i.d. case) in more detail. First, to illustrate how requirement (7) can be satisfied, we assume that $\tau_i(y)$ and $S_i(y)$ are exponential with parameters $\lambda_i(y)$ and $\mu_i(y)$, respectively. Then for any $\delta > 0$ and $x \ge 0$

$$\inf_{y\le x} \mathsf{P}(\tau_i(y) > S_i(y) + \delta) > 0, \tag{27}$$

provided the following natural assumptions hold:

$$\inf_{y\le x} \mu_i(y) > 0,\ \sup_{y\le x}(\lambda_i(y) + \mu_i(y)) < \infty.$$

To show a role of local uniformity in (7), consider the following counterexample for a single server system. Assume that, for a constant $C > 2$ w. p. 1

$$\tau(U_k) - S(U_k) := \delta(U_k) = \frac{1}{C^{k+1}}, \quad k = 0, 1, \ldots, \tag{28}$$

where we denote

$$U_0 = C, \;\; U_k = C - \sum_{i=1}^{k} 1/C^i, \;\; k \geq 1.$$

Then

$$U_k \to C - 1/(C-1) > 1, \;\; k \to \infty,$$

and hence,

$$\{W_n = C\} \subseteq \Big\{\beta(n) := \min(k : U_k \leq 0) = \infty\Big\}.$$

Thus, on the event $\{W_n = C\}$ regeneration point is not attainable after instant n. It is because $\inf_{k\geq 0} \delta(U_k) = 0$ and no $\delta(C) > 0$ exists satisfying (7).

We discuss in brief a possible relaxation of regeneration assumptions which may lead to one-dependent regeneration. First, consider the $GI/G/m$ system with $m \geq 3$ and define

$$b = ess \inf G := \inf(x : G(x) = 0), \;\; a = ess \sup A := \sup(x : 1 - A(x) > 0).$$

Then assumption $\lambda < m\mu$ implies $am > b$. If $a > b$, then a positive recurrent process of classical regenerations exists. Otherwise,

$$(k_0 - 1)a < b < k_0 a$$

for some integer $1 < k_0 \leq m$. (We assume that $ak \neq b$, any integer k.) Then the Markov chain $\{\hat{W}_n\}$ is *Harris ergodic* and has one-dependent regenerations with a *regeneration set* $\mathtt{R}$. Denote the events

$$\mathtt{A}_k = \Big\{S_k \in [b, b+\varepsilon), \;\; \tau_k \in (a - \varepsilon, a]\Big\},$$

where we choose $\varepsilon \in (0,\, k_0 a - b)$. Note that a *minimal regeneration set* is

$$\mathtt{R}_0 = \Big\{x \in \mathsf{R}_+^m : x_i \leq (i - 1 + k_0 - m)^+ a, \;\; i = 1, \ldots, m\Big\},$$

in which case the chain regenerates whenever the *renovating event*

$$\Omega_n = \{\hat{W}_n \in \mathtt{R}_0\} \cap \cap_{k=n}^{n+k_0-1} \mathtt{A}_k\}$$

occurs. (Definitions and more details can be found in [1], pp. 345–346; [5]; [17]; [25]; [36]; [39], pp. 369–372.)

Now consider our basic model described in Theorem 1. By (5), for each i, there exists $\delta(x) > 0$ such that

$$\mathsf{P}(S_i(x) + \delta(x) < m\tau_i(x)) > 0$$

for each (large) x. Even if we sharpen the condition requiring a local uniformity, that is

$$\inf_{y \le x} \mathsf{P}(S_i(y) + \delta(x) < m\tau_i(y)) > 0 \quad \text{for each} \quad x \ge 0, \tag{29}$$

(cf. (7)), then it is still not enough to reach a regeneration within a finite interval with a probability which is *uniformly lower bounded* (over z_k) by a positive constant, which is required to apply characterization (3). Indeed, assumption (29) allows to decrease residual work whenever events $\{S_i(y) + \delta(T) < m\tau_i(y)\}$ occurs (provided $y \le T$) as long as all servers are busy. In the $GI/G/m$ queue, construction of a regeneration point on the event $\{\hat{W}_n^{(1)} = 0\}$ is based on realization of a (required) number of events $\mathtt{A}_k$ $(k \ge n)$ and on a monotonicity property and domination of a cyclic queue [1]. By an analogy, we could consider the following (rather unnatural and restrictive) assumptions

$$ess \inf G_{i,y} = b, \; ess \sup A_{i,y} = a \quad \text{for all} \; i \; \text{and} \; y \tag{30}$$

to construct one-dependent regenerations in our model. Furthermore, unlike the $GI/G/m$ system, the order of occupation of free servers during *renovation period* preceding a regeneration point in our model with the non-identical servers may significantly influence the *regeneration measure* (distribution of the workload process at the regeneration instant). To overcome this problem, we note that in the $GI/G/m$ system on the event Ω_n, provided $k_0 < m$, at least two servers are free at each arrival instant (during *minimal* renovation period $[n, \, n + k_0))$, and thus a fixed order of occupation of servers is achieved during at most $m!$ successive renovation periods. As to the mentioned cyclic system, which is inapplicable in our case, a careful analysis shows that, at least in the $GI/G/m$ system, it is not necessary to obtain a regeneration instant. Thus, assumptions (30) and $k_0 < m$ (together with (5)) seem to be sufficient for the existence of positive recurrent one-dependent regenerations in the basic model. (Similar arguments can be applied to other variations of the model.) We will not go further in this direction since, as we see, an analysis using one-dependent regeneration in our models seems to be complicated, requires restrictive and rather unnatural assumptions, and thus a serious motivation is needed to justify this scenario. However, the corresponding proofs could be obtained by a combination of those above and the known construction of one-dependent regeneration in $GI/G/m$ queue.

Finally, we mention a possibility to extend the model to the case, when the interarrival time depends on whole workload vector $\hat{W}_n$ rather than on its minimal component. This dependence still retains the Markov property of the workload process. We conjecture that stability of this model can be studied by the techniques presented above, but the analysis (especially the tight negative drift condition) may be very complicated. To some extent it is demonstrated in a recent related paper [18], where a stability region of an exponential multiclass system with service rates depending on the current (vector) queue-size is studied under monotonicity assumptions.

Acknowledgements

I am indebted to Onno Boxma who supplied me with references [2, 7] and for the suggestion to study the extension of the basic model included now in Section 4. Special thanks to Maria Vlasiou who kindly supplied me with many useful references and her Ph.D thesis [40]. Furthermore, I thank Richard Serfozo for useful comments. I am thankful CRM for nice conditions during my visit in 2009, when the main part of this research has been done.

This work is supported by Russian Foundation for Basic Research under grant 07-07-00088.

References

[1] S. Asmussen, *Applied Probability and Queues* (2nd ed.). Springer-Verlag, NY (2003).

[2] H. Albrecher, O.J. Boxma, A ruin model with dependence between claim sizes and claims intervals. Insurance: Math. and Econ., **35** (2004), 245–254.

[3] R. Bekker, S.C. Borst, O.J. Boxma, and O. Kella, Queues with workload-dependent arrival and service rates, Queueing Systems, **46** (2004), 537–556.

[4] R. Bekker, Finite-buffer queues with workload-dependent service and arrival rates. Queueing Systems, 50 (2005) 231–253.

[5] A. Borovkov, *Asymptotic Methods in Queueing Theory.* Wiley, NY (1984).

[6] A. Borovkov and Foss, S., Stochastically recursive sequences. Siberian Advances in Mathematica, **2(1)** (1992), 16–81.

[7] O.J. Boxma, D. Perry, A queueing model with dependence between service and interarrival times. European J. Oper.Res., **128(3)** (2001), 611–624.

[8] O. Boxma and M. Vlasiou, On queues with service and interarrival times depending on waiting times, Technical Report 2006-008, Eurandom, Eindhoven, The Netherlands (2006).

[9] P.H. Brill, Single-server queues with delay-dependent arrival streams, Probability in the Engineering and Informational Sceinces, **2** (1988), 231–247.

[10] J.R. Callahan, A queue with waiting time dependent service times. Nav. Res. Log. Quart., **20(2)** (1973), 321–324.

[11] H. Chen, Fluid approximation and stability of multiclass queueing networks: work-conserving disciplines, Ann. Appl. Probab., **5** (1995), 637–665.

[12] J. Dai, On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models, Ann. Appl. Probab., **5** (1995), 49–77.

[13] B. Gavish and P.J. Schweitzer, The Markovian queue with bounded waiting time, Man Sci., **23** (1977), 1349–1357.

[14] B. Gnedenko and I. Kovalenko, *An Introduction to Queueing Theory* (2nd Ed.) Birkhäuser (1989).

[15] W. Feller, *An Introduction to Probability Theory and its Applications. I* (2nd ed.) Wiley, NY (1971).

[16] S. Foss and V. Kalashnikov, Regeneration and renovation in queues, Queueing Systems, **8** (1991), 211–224.

[17] S. Foss and T. Konstantopoulos, An overview on some stochastic stability methods, Journal of the Operations Research Society of Japan, **47(4)** (2004), 275–303.

[18] M. Jonckheere, S.Borst, L.Leskelä, Stability of parallel queueing systems with coupled service rates, Discrete Event Dymanic Systems, **18(4)** (2008), 447–471.

[19] J. Kiefer and J. Wolfowitz, On the theory of queues with many servers, Trans. Amer. Math. Soc. **78** (1955), 1–18.

[20] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, London, 1993.

[21] E. Morozov, A comparison theorem for queueing system with non-identical channels, in, Stability Problems for Stochastic Models, Springer-Verlag, NY (1993), 130–133.

[22] E. Morozov, Regeneration of a closed queueing network, Journal of Mathematical Sciences, **69** (1994), 1186–1192.

[23] E. Morozov. Wide sense regenerative processes with applications to multi-channel queues and networks, Acta Applicandae Math., **34** (1994), 189–212.

[24] E. Morozov, The tightness in the ergodic analysis of regenerative queueing processes, Queueing Systems, **27** (1997), 179–203.

[25] E. Morozov, Weak regeneration in modeling of queueing processes. Queueing Systems, **46** (2004), 295–315.

[26] E. Morozov, Stability of a multiserver regenerative queue with a dependence between workload, input and service time. Prepublications du Laboratoire d'Analyse et de Mathematiques Appliquees, UMR CNRS, 8050 10/2004, Decembre, Universite de Marne-la-Vallee (2004).

[27] E. Morozov and R. Delgado, Stability analysis of regenerative queues, Sci. report 812, CRM, Barcelona (2008) `http://www.crm.cat`.

[28] E. Morozov, A multiserver retrial queue: regenerative stability analysis, *Queueing Systems*, **56** (2007), 157–168.

[29] E. Morozov and R. Delgado, Stability analysis of regenerative queues, Sci. report 812, CRM, Barcelona (`http://www.crm.cat`), 1–33, 2008.

[30] E. Morozov and R. Delgado, Stability analysis of regenerative queues, Automation and Remote control, **70(12)** 2009, 1977–1991.

[31] E. Nummelin, Regeneration in tandem queues, Adv. Appl. Prob., **13** (1981), 221–230.

[32] D. Perry and S. Asmussen, Rejection rules in the $M/G/1$ queue, Queueing Systems, **19** (1995), 105–130.

[33] M. J. M. Posner, Single-server queues with service times depending on waiting time, Operations Research, **21(2)** (1973), 610–616.

[34] R. Serfozo, *Introduction to Stochastic Networks*, Springer-Verlag, NY (1999).

[35] K. Sigman, Regeneration in tandem queues with multiserver stations, J. Appl. Prob., **25** (1988), 391–403.

[36] K. Sigman, Queues as Harris recurrent Markov chains, Queueing Systems, **3** (1988), 179–198.

[37] K. Sigman, The stability of open queueing networks, Stochastic Process and Their Applications, **35** (1990), 11–25.

[38] K. Sigman, Appendix: A primer on heavy-tailed distributions, Queueing Systems, **33** (1999), 261–275.

[39] H. Thorisson, *Coupling, Stationarity, and Regeneration* (Springer, NY, 2000).

[40] M. Vlasiou, Lindley-type recursions. Ph.D thesis, Eindhoven University of Technology, The Netherlands (2006).

[41] W. Whitt, Queues with service times and interarrival times depending linearly and randomly upon waiting times. Queueing Systems, **6** (1990), 335–352.

Institute of Applied Mathematical Research
Karelian Research Centre of Russian Academy of Sciences
and Petrozavodsk University