Queueing Systems
https://doi.org/10.1007/s11134-018-9587-9

Stability and busy periods in a multiclass queue with state-dependent arrival rates

Philip A. Ernst[1] · **Søren Asmussen**[2] · **John J. Hasenbein**[3]

Received: 7 March 2017 / Revised: 17 August 2018

Abstract
We introduce a multiclass single-server queueing system in which the arrival rates depend on the current job in service. The system is characterized by a matrix of arrival rates in lieu of a vector of arrival rates. Our proposed model departs from existing state-dependent queueing models in which the parameters depend primarily on the number of jobs in the system rather than on the job in service. We formulate the queueing model and its corresponding fluid model and proceed to obtain necessary and sufficient conditions for stability via fluid models. Utilizing the natural connection with the multitype Galton–Watson processes, the Laplace–Stieltjes transform of busy periods in the system is given. We conclude with tail asymptotics for the busy period for heavy-tailed service time distributions for the regularly varying case.

Keywords Busy periods · Fluid models · Multiclass queues · Regular variation · Stability · State-dependent arrival rates

Mathematics Subject Classification 90B22 · 60K25

1 Introduction

We introduce a multiclass single-server queueing system in which the arrival rates depend on the current job in service. The system is characterized by a matrix of arrival

✉ Philip A. Ernst
philip.ernst@rice.edu

Søren Asmussen
asmus@math.au.dk

John J. Hasenbein
jhas@mail.utexas.edu

1 Rice University, 6100 Main Street, Houston, TX 77005, USA

2 Aarhus University, Ny Munkegade, 8000 Aarhus, Denmark

3 University of Texas at Austin, 204 E. Dean Keeton Street, Austin, TX 78712, USA

Published online: 04 September 2018

rates instead of a vector of arrival rates. The proposed model departs from existing state-dependent models in the literature in which the parameters depend primarily on the number of jobs in the system (see Bekker et al. [3], Cruz and Smith [6], Jain and Smith [12], Miller [14], Perry et al. [17] and Yuhaski and Smith [21], among other sources) rather than the job in service.

Our model is motivated by two practical queueing considerations. The first is a multiclass queueing system in which the arriving customer can observe only the class of the customer in service and no other characteristics of the queue. This information informs the customer's decision to either join or leave the queue. The second concerns local area networks with a central server in which K clients generate requests at individual Poisson rates μ_i. Often, a client does not generate requests when a previous request is being handled by the server. Further, it is conceivable that groups of clients working together may influence each other's Poisson rate. To the best of our knowledge, this simple yet potentially very useful queueing model has never appeared in the literature. This serves as our primary motivation for the manuscript.

The remainder of the work is structured as follows. We formulate the queueing model in Sect. 2 and its corresponding fluid model in Sect. 3. In Sect. 4, we obtain necessary and sufficient conditions for stability via fluid models. Through the natural connection with multitype Galton–Watson processes, we characterize the Laplace–Stieltjes transform of busy periods in the system in Sects. 5 and 6.1. Section 6.2 concerns tail asymptotics of the busy period in the case of heavy-tailed service time distributions. Section 7 offers a brief conclusion and presents ideas for future work.

2 The queueing model

Consider a multiclass single-server queue with K classes of jobs, each arriving according to independent counting processes. We assume that only one job may be serviced at a time. Let the arrival rate depend *on the class of the job in service*. If the server is serving a job of class i, the arrival rate of class j jobs is λ_{ij}, $i, j = 1, \ldots, K$. The matrix of arrival rates is defined as $\mathbf{\Lambda} = (\lambda_{ij})$, $i, j = 1, \ldots, K$. If there is no job in service, then the arrival rate of class j jobs is defined as λ_{0j}, $j = 1, \ldots, K$. The arrival mechanism is described more precisely with dynamical equations in Sect. 3.

We proceed to set notation. Let $\bar{\lambda}^i = \sum_{j=1}^K \lambda_{ij}$ for each $i = 1, \ldots, K$. Service times for class i jobs are assumed to be i.i.d. with distribution function F_i, $i = 1, \ldots, K$. Let S_i be a generic service time for class i jobs, with $\mathbb{E}[S_i] = m_i = \mu_i^{-1}$, $i = 1, \ldots, K$ and $\mathbf{G} = \text{diag}(\mu_1, \mu_2, \ldots, \mu_K)$. We define the "mean offspring matrix" to be $\mathbf{M} = \mathbf{G}^{-1}\mathbf{\Lambda}$ (here, the ijth element $\lambda_{ij} m_i$ is the mean number of arriving class j customers during service of a class i customer). By definition, all the elements of $\mathbf{M}$ are nonnegative, and this is enough to ensure that the dominant eigenvalue $\rho(\mathbf{M})$ is real and positive, cf. [10]. For some results, more restrictive conditions on $\mathbf{M}$ will be required. Further, let ψ_i denote the Laplace–Stieltjes transform (LST) of S_i, $i = 1, \ldots, K$, respectively, that is, $\psi_i(s) = \mathbb{E}[\mathrm{e}^{-sS_i}] = \int_0^\infty \mathrm{e}^{-st} \mathrm{d}F_i(s)$ for $s > 0$. We let Q_i denote the steady-state number of class i jobs in the system, $i = 1, \ldots, K$, and let $\mathbf{Q} = (Q_1, \ldots, Q_K)$. Each state of the system takes nonnegative integer values, that is, $\mathbf{x} = (x_1, \ldots, x_K) \in \mathbb{Z}_+^K$.

The service disciplines we consider are nonidling, i.e., jobs must be served using the full capacity of the server whenever there are jobs in the system. Our results on stability and the busy period are independent of the particular (nonidling) scheduling policy employed in the system.

3 Queueing and fluid dynamics

3.1 Queueing dynamical equations

We now precisely define the arrival mechanism. For $i \in \{1, \ldots, K\}$ and $t \geq 0$, $Q_i(t)$ denotes the number of class i jobs in the system at time t, whether in service or in the queue. Similarly, let $T_i(t)$ denote the amount of time that has been devoted to serving class i jobs in $[0, t]$. Further, let $A_i(t)$ and $D_i(t)$ be, respectively, the total number of class i jobs that have arrived and departed from the system in $[0, t]$. We then have the following input–output equation for each class i:

$$Q_i(t) = Q_i(0) + A_i(t) - D_i(t). \tag{3.1}$$

For each class i, the counting process $\mathcal{E}_i^j(t)$ is the number of class i jobs that arrive during the first t time units devoted to processing class j. $\mathcal{E}_i^0(t)$ counts the number of class i arrivals during the first t time units that no job is being processed at the server. The total number of class i arrivals in $[0, t]$ is then given by

$$A_i(t) = \mathcal{E}_i^0\big(T_0(t)\big) + \sum_{j=1}^{N} \mathcal{E}_i^j\big(T_j(t)\big), \tag{3.2}$$

where the counting processes $\mathcal{E}_i^j$ for $i = 1, \ldots, K$ and $j = 0, \ldots, K$ are assumed to be mutually independent.

As for the service processes, for each i, $1 \leq i \leq K$, and positive integer n, we let $V_i(n)$ denote the total service requirement for the first n class i jobs. Assuming an HL service discipline, we have that

$$V_i(D_i(t)) \leq T_i(t) \leq V_i(D_i(t) + 1) \tag{3.3}$$

for each $t \geq 0$ and $1 \leq i \leq N$.

We define the workload in the system at time t to be

$$W(t) = \sum_{i=1}^{K} V_i(A_i(t) + Q_i(0)) - \sum_{i=1}^{K} T_i(t), \tag{3.4}$$

and the cumulative idle time process to be

$$Y(t) = t - \sum_{i=1}^{K} T_i(t). \tag{3.5}$$

It is important to note that Y is a nondecreasing function. We assume that the queueing policy is nonidling, which specifically means that Y can increase only when $W(t) = 0$. More precisely,

$$\int_0^\infty W(t)\, dY(t) = 0.$$

3.2 Fluid model

For purposes of determining the stability conditions of a more general version of our model, we formulate a fluid network version of the model. For references to important definitions and results in the fluid model literature, we refer the reader to Bramson [5] and Gamarnik [9].

For $i \in \{1, \ldots, K\}$ and $t \geq 0$, let $Q_i(t)$ denote the amount of fluid of class i in the system at time t. Similarly, let $T_i(t)$ denote the amount of time that has been devoted to serving class i fluid in $[0, t]$. We also define $A_i(t)$ and $D_i(t)$ which are, respectively, the total amount of class i fluid that has arrived and departed from the system in $[0, t]$. We then have the following standard equation:

$$Q_i(t) = Q_i(0) + A_i(t) - D_i(t), \tag{3.6}$$

for each $i \in \{1, \ldots, K\}$ and $t \geq 0$. The departure processes in this system also obey the standard relation $D_i(t) = \mu_i T_i(t)$ for all $t \geq 0$.

The unusual feature of our model lies in the arrival process, which is dependent on the current class in service. In the queueing model, processor sharing is not allowed. Hence, there is (at most) one class in service at any given time and "the customer in service" is defined unambiguously. Here, we provide a more general formulation that reduces to the queueing model presented in earlier sections, under appropriate restrictions on the allowable queueing disciplines. First, we recall the usual condition

$$\sum_{i=1}^{N} \dot{T}_i(t) \leq 1, \tag{3.7}$$

which simply indicates that the server cannot devote more than 100% of its time to serving fluids of all classes. Since we assume that the queueing discipline is nonidling, $\sum_{i=1}^{N} \dot{T}_i(t) = 1$ whenever there is a positive amount of fluid in the system. We also define the idle time in $[0, t]$ to be

$$Y(t) = t - \sum_{i=1}^{N} T_i(t).$$

Note that the current arrival rate of class j fluid is given by $\dot{A}_j(t)$. In the queueing model, if a job of class i is in service then the arrival rate of class j jobs is λ_{ij}. Let $\boldsymbol{\lambda}_j$ be the column vector $(\lambda_{1j}, \ldots, \lambda_{Nj})^\perp$ and let $\dot{\mathbf{T}}(t)$ be the column vector

$\left(\dot{T}_1(t), \ldots, \dot{T}_N(t)\right)^{\perp}$, where $^{\perp}$ means transposition. We define the fluid arrival rate of class j to be

$$\dot{A}_j(t) = \lambda_{0j}\dot{Y}(t) + \lambda_j^{\perp}\dot{\mathbf{T}}(t). \tag{3.8}$$

In particular, when there is fluid in the system, the class j arrival rate is a convex combination of the elements of $\boldsymbol{\lambda}_j$. If we restrict to policies in which only one class can be served at any time, then Eq. (3.8) assigns an arrival rate of λ_{ij} to class j fluid when class i fluid is in service. Note that this concurs with the queueing model formulation. Combining the above, we have

$$Q_j(t) = Q_j(0) + \int_0^t \left(\lambda_{0j}\dot{Y}(u) + \lambda_j^{\perp}\dot{\mathbf{T}}(u)\right) du - \mu_j T_j(t) \tag{3.9}$$

$$= Q_j(0) + \lambda_{0j}Y(t) + \lambda_j^{\perp}\mathbf{T}(t) - \mu_j T_j(t). \tag{3.10}$$

Writing equations (3.9) and (3.10) in matrix form yields

$$\mathbf{Q}(t) = \mathbf{Q}(0) + (\mathbf{M}^{\perp} - \mathbf{I})\mathbf{D}(t) + Y(t)\boldsymbol{\lambda}_{\mathbf{0}}. \tag{3.11}$$

We define the vector of fluid work in the system at time t to be

$$\mathbf{W}(t) = \mathbf{G}^{-1}\mathbf{Q}(t). \tag{3.12}$$

3.2.1 Fluid limits

Thus far we have described a fluid model, but it remains to show that the fluid limits of the queueing model satisfy the fluid model equations. In this subsection only, we use a bar to denote a fluid limit. As usual, we define the fluid limits of the queue-length processes to be

$$\bar{Q}_i(t) = \lim_{n\to\infty} \frac{Q_i(nt)}{n},$$

with other fluid limits defined in an analogous manner. We make the usual assumptions on the stochastic primitives and initial conditions, i.e., for all i and j

$$\lim_{n\to\infty} \frac{\mathcal{E}_i^j(nt)}{n} = \lambda_{ji}t, \tag{3.13}$$

$$\lim_{n\to\infty} \frac{\mathcal{E}_i^0(nt)}{n} = \lambda_{0i}t, \tag{3.14}$$

$$\lim_{n\to\infty} \frac{V_i(n)}{n} = m_i, \tag{3.15}$$

$$\lim_{n\to\infty} \frac{Q_i(0)}{n} = \bar{Q}_i(0), \tag{3.16}$$

where the convergence is almost surely, uniformly on compact sets. Under these assumptions, the fluid model equations can be derived in a straightforward way from

the queueing dynamical equations, since all but the arrival rate process is identical to the standard multiclass queueing network model. For the arrival process, we have

$$\bar{A}_i(t) = \lim_{n\to\infty} \frac{A_i(nt)}{n} = \lim_{n\to\infty} \frac{\mathcal{E}_i^0(T_0(nt))}{n} + \lim_{n\to\infty} \sum_{j=1}^{N} \frac{\mathcal{E}_i^j(T_j(nt))}{n}$$
$$= \lambda_{0i}\bar{Y}(t) + \lambda_i^{\perp}\bar{\mathbf{T}}(t).$$

The last equality follows from assumptions (3.13) and (3.14) and similar arguments as found in Proposition 4.12 in [5]. Finally, the connection between fluid stability and queueing network stability follows from straightforward modification of existing stability results, under the usual assumptions that the interarrival times for all job classes are unbounded and spread out. We refer the reader to Chapter 4 of Bramson [5] for full details.

4 Stability results for fluid model

In this section, we prove a number of results regarding the stability, or instability, of the fluid model. The proofs rely on the following two observations:

1. $T_i(\cdot)$ is Lipschitz continuous for each i and hence so is any linear function f of $(T_1, \ldots, T_K)$. Thus, f is absolutely continuous and its derivative exists almost everywhere.
2. If $\dot{f}(t)$ exists for $t > 0$, t is called a regular point.

We define $\mathbf{e} = (1, \ldots, 1)^{\perp}$ and assume this column vector is of size K. Finally, we set $\mathbf{H} = \mathbf{GMG^{-1}}$.

Theorem 1 *If $\rho(\boldsymbol{M}) < 1$, then $f(t) = \mathbf{e}^{\perp}(\boldsymbol{I} - \boldsymbol{H}^{\perp})^{-1}\boldsymbol{G}^{-1}Q(t)$ is a Lyapunov function for the fluid model.*

Proof Note that $\rho(\mathbf{H}) = \rho(\mathbf{M}) < 1$. Hence, $\mathbf{I} - \mathbf{H}$ is an $\mathcal{M}$-matrix. Therefore $\mathbf{I} - \mathbf{H}$ is invertible with a nonnegative inverse.

Let us assume that the fluid system starts from a nonempty state, i.e., $\mathbf{Q}(0) \neq \mathbf{0}$. By the continuity of $\mathbf{Q}$, $\mathbf{Q}(t) \neq 0$ for all t in some interval $[0, s)$. Then, we have $Y(t) = 0$ for all $t \in [0, s)$. Using Eqs. (3.11) and (3.12) we have

$$\mathbf{W}(t) = \mathbf{W}(0) - (\mathbf{I} - \mathbf{H}^{\perp})\mathbf{T}(t),$$

for $t \in [0, s)$. Multiplying by $\mathbf{e}^{\perp}(\mathbf{I} - \mathbf{H}^{\perp})^{-1}$ yields

$$\mathbf{e}^{\perp}(\mathbf{I} - \mathbf{H}^{\perp})^{-1}\mathbf{W}(t) = \mathbf{e}^{\perp}(\mathbf{I} - \mathbf{H}^{\perp})^{-1}\mathbf{W}(0) - \mathbf{e}^{\perp}\mathbf{T}(t).$$

As in the statement of the theorem, set

$$f(t) = \mathbf{e}^{\perp}(\mathbf{I} - \mathbf{H}^{\perp})^{-1}\mathbf{G}^{-1}\mathbf{Q}(t),$$

and note that $f(t) = 0$ if and only if $\mathbf{Q}(t) = \mathbf{0}$. Then, we have

$$f(t) = \mathbf{e}^{\perp}(\mathbf{I} - \mathbf{H}^{\perp})^{-1}\mathbf{W}(t) \tag{4.1}$$
$$= \mathbf{e}^{\perp}(\mathbf{I} - \mathbf{H}^{\perp})^{-1}\mathbf{W}(0) - \mathbf{e}^{\perp}\mathbf{T}(t). \tag{4.2}$$

Taking derivatives, we obtain

$$\dot{f}(t) = -\mathbf{e}^{\perp}\dot{\mathbf{T}}(t) = -1,$$

for any $t \in [0, s)$ and regular point t. Therefore, the draining time of the system under any feasible policy is

$$f(0) = \mathbf{e}^{\perp}(\mathbf{I} - \mathbf{H}^{\perp})^{-1}\mathbf{W}(0),$$

which can be interpreted as the initial unfinished "potential" work, defined as the work due to the current workload and work generated in the future by the initial workload's "offspring." The above argument implies that $\dot{f}(t) = -1$ whenever $\mathbf{W}(t) \neq \mathbf{0}$ and thus the system stays drained once $\mathbf{Q}(t) = \mathbf{0}$. This completes the proof. □

The corollary below now immediately follows.

Corollary 1 *The fluid model is globally stable if* $\rho(\boldsymbol{M}) < 1$.

4.1 Weak instability

Next we show that the fluid model is weakly unstable if $\rho(\mathbf{M}) > 1$. We begin by introducing the following lemma.

Lemma 1 *Suppose* $\rho(\boldsymbol{M}) = \rho(\boldsymbol{H}) > 1$ *and that each row of* $\mathbf{M}$ *has at least one strictly positive element. Then, for all nonnegative vectors* $\mathbf{T}(t) > 0$, $\mathbf{V}(t) = (\boldsymbol{I} - \boldsymbol{H})\mathbf{T}(t)$ *must have some component* $V_i(t) < 0$ *for some* $i \in \{1, ..., K\}$.

Proof We argue to the contrary. Note that each row of $\mathbf{H}$ has at least one strictly positive element, by the same assumption on $\mathbf{M}$. Also, for some $\alpha \in (0, 1)$, $\rho(\alpha\mathbf{H}) = 1$. For the sake of contradiction, assume that there exists a nonnegative vector $\mathbf{T}(t) > 0$ s.t. $\mathbf{V}(t) = (\mathbf{I} - \mathbf{H})\mathbf{T}(t) \geq 0$. Further, define $\mathbf{V}'(t) = (\mathbf{I} - \alpha\mathbf{H})\mathbf{T}(t)$. We now consider

$$\mathbf{V}(t) - \mathbf{V}'(t) = (\mathbf{I} - \mathbf{H})\mathbf{T}(t) - (\mathbf{I} - \alpha\mathbf{H})\mathbf{T}(t) = (\alpha\mathbf{H} - \mathbf{H})\mathbf{T}(t) < 0.$$

The above equations imply that $\mathbf{V}'(t) > \mathbf{V}(t)$ and that there exists some $\mathbf{T}(t) > 0$ with $(\mathbf{I} - \alpha\mathbf{H})\mathbf{T}(t) > 0$. Thus, $(\mathbf{I} - \alpha\mathbf{H})$ is semipositive, and by condition I_{27} in Chapter 6 of Berman and Plemmons [4], $(\mathbf{I} - \alpha\mathbf{H})$ is a nonsingular $\mathcal{M}$-matrix. This implies that $\rho(\alpha\mathbf{H}) < 1$, yielding a contradiction. □

We are now ready to prove Theorem 2, the main result of this subsection.

Theorem 2 *The fluid model is weakly unstable if* $\rho(\mathbf{M}) > 1$ *and each row of* $\mathbf{M}$ *has at least one strictly positive element.*

Proof Assume $\mathbf{Q}(0) = \mathbf{W}(0) = \mathbf{0}$. Then, for any $t > 0$ we have by (3.11) and (3.12) that

$$\mathbf{W}(t) \geq \mathbf{W}(0) - (\mathbf{I} - \mathbf{H}^{\perp})\mathbf{T}(t) = (\mathbf{H}^{\perp} - \mathbf{I})\mathbf{T}(t).$$

By Lemma 1, there exists some component of $\mathbf{W}(t)$ s.t. $W_i(t) > 0$. This implies that $\mathbf{Q}(t) \neq \mathbf{0}$ for all $t > 0$. Thus, the fluid model is weakly unstable.

□

4.2 Weak stability

Theorem 3 *Suppose that* $\mathbf{M}$ *is an irreducible nonnegative matrix. Then the fluid model is weakly stable if* $\rho(\mathbf{M}) \leq 1$.

Proof It suffices to show the result for the case $\rho(\mathbf{M}) = 1$, since we have already shown that the fluid model is "strongly" stable when $\rho(\mathbf{M}) < 1$.

Let $\mathbf{Q}(0) = 0$. We argue to the contrary. For the sake of contradiction, let us assume that $\mathbf{Q}(t) \neq \mathbf{0}$ for some $t > 0$. Then, since $\mathbf{Q}$ is continuous, there must be an interval (t_1, t_2), with $t_2 > t_1$, for which $\|\mathbf{Q}(t)\| > 0$ for all $t \in (t_1, t_2)$, $\mathbf{Q}(t_1) = 0$ and $\|\mathbf{Q}(t_2)\| > 0$. In particular, we may set $t_1 = \inf\{t : \mathbf{Q}(t) \neq \mathbf{0}\}$. Now, recall that

$$\mathbf{Q}(t) = (\mathbf{M}^{\perp} - \mathbf{I})\mathbf{D}(t) + Y(t)\boldsymbol{\lambda}_{\mathbf{0}}. \tag{4.3}$$

Since $\mathbf{M}$ is a positive matrix, it follows by the Perron–Frobenius Theorem that there exists a positive left (row) eigenvector $\mathbf{w}$ of $\mathbf{M}$ with $\mathbf{wM} = \mathbf{w}$, $w_i > 0$ for $i \in \{1, ..., K\}$. Multiplying both sides of (4.3) by $\mathbf{w}$ we obtain

$$\mathbf{wQ}(t) = \mathbf{w}\big[(\mathbf{M}^{\perp} - \mathbf{I})\mathbf{D}(t) + Y(t)\boldsymbol{\lambda}_{\mathbf{0}}\big] = \mathbf{w}Y(t)\boldsymbol{\lambda}_{\mathbf{0}},$$

for all $t \geq 0$. Recalling $\mathbf{Q}(t_1) = \mathbf{0}$ and $\|\mathbf{Q}(t_2)\| > 0$ we have

$$\mathbf{w}\big(Y(t_2) - Y(t_1)\big)\boldsymbol{\lambda}_{\mathbf{0}} = \mathbf{w}\big(\mathbf{Q}(t_2) - \mathbf{Q}(t_1)\big) > 0.$$

This implies $Y(t_2) > Y(t_1)$ and thus there is positive idle time in (t_1, t_2). However, since the fluid level is positive in this entire interval, this violates the nonidling condition. Thus such a fluid solution is not feasible. A contradiction has been reached. This concludes the proof.

□

5 Branching process connection

In the remainder of the paper, we investigate a special case of the multiclass model discussed so far. In particular, we now assume that arrivals to each class form a Poisson process, i.e., the model is an $M/G/1$ multiclass queue, rather than a $GI/G/1$ queue. Although more general stability conditions for the $GI/G/1$ case were proven

in Sect. 4, we begin by reproving them in the Poisson setting, by making a connection to branching processes. There are two reasons to do this. First, the stability results arise in a somewhat more intuitive manner using this methodology. Secondly, we find the connection to branching processes illuminating and useful in later sections.

A classical tool for the simple $M/G/1$ queue and related systems is to interpret customers as individuals in a branching process, such that the children of a customer are the customers arriving during his or her service. This is useful because the stability condition for the queueing system is the same as the condition for almost sure extinction. Carrying out the same idea for our multiclass systems leads to a K-type Crump–Mode–Jagers branching process $\left\{\mathbf{Z}_n = \left(Z_n^{(1)}, \ldots, Z_n^{(K)}\right) : n \geq 1\right\}$, such that the lifetime of an individual of type j has the same distribution as S_j. In the results below, we consider a branching process with a single ancestor of type i. Whenever $\mathbb{E}_i$ and $\mathbb{P}_i$ are used, they are with reference to the probability measure induced by such a single ancestor. The offspring mechanism is then described by the probabilities

$$p_{ij}(k) = \mathbb{P}_i\left(Z_1^{(j)} = k\right) = \mathbb{P}\big(\mathrm{Pois}(\lambda_{ij} S_j) = k\big) = \int_0^\infty \frac{(\lambda_{ij} s)^k \mathrm{e}^{-\lambda_{ij} s}}{k!} \mathrm{d}F_j(s)\,. \quad (5.1)$$

The offspring matrix $\mathbf{M} = (M_{ij})_{i,j=1,\ldots,K}$ is given by $M_{ij} = \mathbb{E}_i\left[Z_1^{(j)}\right] = \lambda_{ij}/\mu_i$ and is assumed irreducible. Thus, Perron–Frobenius theory applies to $\mathbf{M}$ and, as before, $\rho = \rho(\mathbf{M})$ the largest eigenvalue. Note that the ijth element of the matrix $\sum_{n=0}^\infty \mathbf{M}^n$ gives the expected number of type j progeny of an individual of type i; of course, when $\rho < 1$, we have $\sum_{n=0}^\infty \mathbf{M}^n = (\mathbf{I} - \mathbf{M})^{-1}$.

5.1 Stability conditions

Let $|\mathbf{Z}_n| = \sum_{j=1}^K Z_n^{(j)}$ denote the total number of individuals in the nth generation and T^* the extinction time. Let $\mathbb{P}_i(T^* < \infty)$ be the extinction probability of type i of the branching process. Then, by classical results, we have the following theorem.

Theorem 4

$$\mathbb{P}_i(T^* < \infty) = 1,\ \ i = 1, \ldots, K,\ \textit{if and only if}\ \rho \leq 1. \quad (5.2)$$

Proof By the classical result for the extinction time of branching processes [11, Chap II. Theorem 7.1], if and only if $\rho \leq 1$, the total number of generations for each type is finite with probability 1 and thus $\sum |\mathbf{Z}_n| < \infty$, which further implies $\mathbb{P}_i(T^* < \infty) = 1$ for every i. □

Consider $K = 2$. Straightforward algebra gives that $\rho \leq 1$ is equivalent to

$$\frac{\frac{\lambda_{11}}{\mu_1} + \frac{\lambda_{22}}{\mu_2} + \sqrt{\left(\frac{\lambda_{11}}{\mu_1} - \frac{\lambda_{22}}{\mu_2}\right)^2 + \frac{4\lambda_{12}\lambda_{21}}{\mu_1\mu_2}}}{2} \leq 1. \quad (5.3)$$

Theorem 5 $\mathbb{E}_i T^* < \infty$ *for all* i *if and only if* $\rho < 1$.

Proof For a simple proof of sufficiency, assume $\rho < 1$ and let $S_j(m;n)$ denote the lifetime of the mth individual of type j in the nth generation, $\underline{\mu} = \min_1^K \mu_j$. Then

$$\mathbb{E}_i T^* = \mathbb{E}_i \sum_{n=0}^{\infty} \sum_{j=1}^{K} \sum_{m=1}^{Z_n^{(j)}} S_j(m;n) = \mathbb{E}_i \sum_{n=0}^{\infty} \sum_{j=1}^{K} \frac{Z_n^{(j)}}{\mu_j}$$
$$\leq \underline{\mu}^{-1} \mathbb{E}_i \sum_{n=0}^{\infty} \sum_{j=1}^{K} Z_n^{(j)} = \underline{\mu}^{-1} \sum_{n=0}^{\infty} \sum_{j=1}^{K} M_{ij}^n < \infty,$$

where the second step above uses that $S_j(m;n)$ is independent of $\mathbf{Z}_0, \ldots, \mathbf{Z}_n$ (but not $\mathbf{Z}_{n+1}, \mathbf{Z}_{n+2}, \ldots$). Further, the strict inequality

$$\sum_{n=0}^{\infty} \sum_{j=1}^{K} M_{ij}^n < \infty \tag{5.4}$$

follows from $\rho < 1$. To prove the necessity, let $\bar{\mu} = \max_1^K \mu_j$. Then, by the same reasoning we get $\mathbb{E}_i T^* \geq \bar{\mu}^{-1} \sum_{n=0}^{\infty} \sum_{j=1}^{K} M_{ij}^n = \infty$ for $\rho \geq 1$ (see Berman and Plemmons [4]). Hence, $\mathbb{E}_i T^* < \infty$ is also necessary for $\rho < 1$. □

We now have the following corollary to Theorem 4.

Corollary 2 *The busy period $T < \infty$ w.p.1 if and only if the matrix $\mathbf{M}$ given by*

$$M_{ij} = \frac{\lambda_{ij}}{\mu_i}, \quad i, j = 1, \ldots, K,$$

has largest eigenvalue $\rho(\mathbf{M}) \leq 1$.

Similarly, a corollary to Theorem 5 is stated below.

Corollary 3 *For the busy period T, $\mathbb{E}T < \infty$ if and only if $\rho < 1$.*

5.2 Further applications

Let $B_{i;z}$ denote the length of the busy period initiated by a class i customer with service requirement z (B_i is that of the standard busy period initiated by a class i customer, that is, taking $z = S_i$). Let further

$$\tau_j = \mathbb{E}_i \sum_{n=0}^{\infty} \sum_{m=1}^{Z_n^{(j)}} S_j(m;n) \tag{5.5}$$

be the expected total time in $[0, B_j)$ where the customer being served is of class i. As before, $\mathbf{G}$ is a diagonal matrix with the μ_i on the diagonal.

Lemma 2 *Assume $\rho < 1$. Then:*

(i) $(\mathbb{E}_i \tau_j)_{i,j=1,\ldots,K} = (\mathbf{I} - \mathbf{M})^{-1}\mathbf{G}^{-1}$; (ii) $\mathbb{E}B_i = \mathbf{e}_i^\top (\mathbf{I} - \mathbf{M})^{-1}\mathbf{G}^{-1}\mathbf{e}$;
(iii) $\mathbb{E}B_{i;z} = z\beta_i$, *where* $\beta_i = \mathbf{e}_i^\top \mathbf{\Lambda}(\mathbf{I} - \mathbf{M})^{-1}\mathbf{G}^{-1}\mathbf{e}$;
(iv) $B_{i;z}/z \to \beta_i$ *in probability as* $z \to \infty$.

Proof (i) follows immediately since the ijth element of $(\mathbf{I} - \mathbf{M})^{-1}\mathbf{G}^{-1}$ is

$$\sum_{n=0}^{\infty} M_{ij}^n/\mu_j = \mathbb{E}_i \sum_{n=0}^{\infty} Z_n^{(j)}/\mu_j = \mathbb{E}\tau_i,$$

and (ii) follows from (i) by summing over j. For (iii) and (iv), we may (by work conservation) assume that the discipline is preemptive resume. The workload process during service of a class i customer evolves as a standard compound Poisson process with arrival rate $\bar{\lambda}^i = \sum_{j=1}^N \lambda_{ij}$ and with cumulative distribution function

$$\sum_{j=1}^{K} \frac{\lambda_{ij}}{\bar{\lambda}^i} \mathbb{P}(B_j \le x)$$

for the jumps. For this system, the rate of arriving work is $\bar{\lambda}^i \sum_{j=1}^K \lambda_{ij}/\bar{\lambda}^i \, \mathbb{E}B_j$, which is the same as β_i. Now we may simply appeal to standard compound Poisson results to obtain (iii) and (iv). This concludes the proof. □

6 Busy period results

In this section, we begin by assuming $\rho(M) \le 1$. Let $B_{\mathbf{x}}$ denote the busy period when the system starts from the state $\mathbf{x} \in \mathbb{Z}_+^K$, that is, the time period until the system becomes empty. In particular, when $\mathbf{x}$ consists of a single customer of class i, we denote the busy period as B_i, and $B_{i,s}$ is the busy period when his remaining service is s. Define $g_{\mathbf{x}}$ to be the LST of $B_{\mathbf{x}}$, i.e., $g_{\mathbf{x}}(\theta) = \mathbb{E}_{\mathbf{x}}[\mathrm{e}^{-\theta B_{\mathbf{x}}}]$ for $\mathbf{x} \in \mathbb{Z}_+^K$, and similarly for g_i, $g_{i,s}$.

6.1 The busy period Laplace transform

For the $M/G/1$ queue, when $K = 1$, it is well known that the LST of the busy period B is given by

$$g(\theta) = \psi(\theta + \lambda - \lambda g(\theta)), \tag{6.1}$$

where ψ is the LST of the service time and λ is the arrival rate. See, for example, Neuts [15] or Wolff [20]. We shall use the branching process connection to derive a similar fixed point equation for our model.

We first observe that the busy period of the system corresponding to an arbitrary initial state $\mathbf{x} = (x_1, \ldots, x_K) \in \mathbb{Z}_+^K$ is the independent sum of busy periods, each of which corresponds to the branching process starting with a single customer. This gives immediately that

$$g_{\mathbf{x}}(\theta) \;=\; g_1^{x_1}(\theta) \cdots g_K^{x_K}(\theta) \quad \text{when } \mathbf{x} = (x_1, \ldots, x_K) \in \mathbb{Z}_+^K. \tag{6.2}$$

Hence, it is sufficient to calculate g_i, $g_{i,s}$. Recall that ψ_i is the LST of the service time distribution F_i of a class i customer.

Theorem 6 *For $\theta \geq 0$,*

$$g_{i,s}(\theta) \;=\; \exp\left\{-s\left(\theta + \bar{\lambda}^i - \sum_{j=1}^{K} \lambda_{ij} g_j(\theta)\right)\right\}. \tag{6.3}$$

Further,

$$g_i(\theta) \;=\; \psi_i\left(\theta + \bar{\lambda}^i - \sum_{j=1}^{K} \lambda_{ij} g_j(\theta)\right), \quad i = 1, \ldots, K, \tag{6.4}$$

and the vector $\big(g_1(\theta), \ldots, g_K(\theta)\big)$ is the minimal nonnegative and nonincreasing solution of this system of equations.

Proof Clearly, $B_{i,s}$ is the service time s plus the busy periods of all customers arriving during service. But the number of such customers of class j is Poisson$(\lambda_{ij}s)$ and so their busy periods add up to a compound Poisson random variable with LST $\exp\big\{\lambda_{ij}s\big(g_j(\theta) - 1\big)\big\}$. The independence for different j then gives

$$g_{i,s}(\theta) \;=\; \mathrm{e}^{-\theta s} \prod_{j=1}^{K} \exp\big\{\lambda_{ij}s\big(g_j(\theta) - 1\big)\big\},$$

which is the same as (6.3). Integrating with respect to $F_i(\mathrm{d}s)$ then gives (6.4).

Now consider another nonnegative solution $\big(\widetilde{g}_1(\theta), \ldots, \widetilde{g}_K(\theta)\big)$ of (6.4). Define the depth D of the multitype Galton–Watson family tree as $D = \max\big\{n \geq 0 : \mathbf{Z}_n \neq \mathbf{0}\big\}$ and let $g_i^{(n)}(\theta) = \mathbb{E}[\mathrm{e}^{-\theta B_i};\ D \leq n]$. Here $D = 0$ means no arrivals during service. This occurs with probability $\mathrm{e}^{-\bar{\lambda}^i S_i}$ given S_i, and so $g_i^{(0)}(\theta) = \psi_i(\theta + \bar{\lambda}^i)$. The assumptions on $\widetilde{g}_j(\theta)$ then give $\widetilde{g}_i(\theta) \geq g_i^{(0)}(\theta)$. Further, the same reasoning as that leading to (6.4) gives

$$g_i^{(n+1)}(\theta) \;=\; \psi_i\left(\theta + \bar{\lambda}^i - \sum_{j=1}^{K} \lambda_{ij} g_j^{(n)}(\theta)\right).$$

By induction starting from $g_i^{(0)}(\theta) \leq \widetilde{g}_i(\theta)$ we then get $g_i^{(n)}(\theta) \leq \widetilde{g}_i(\theta)$ for all n. The proof is completed by observing that $\rho(\mathbf{M}) \leq 1$ implies $D < \infty$ and hence $g_i^{(n)}(\theta) \uparrow g_i(\theta)$. □

Example 1 Consider a network with $K = 2$ users, $\lambda_{11} = \lambda_{22} = 0$ and F_i exponential(μ_i). Then, (6.4) has the form

$$g_1 = \frac{\mu_1}{\mu_1 + \theta + \lambda_{12} - \lambda_{12} g_2}, \quad g_2 = \frac{\mu_2}{\mu_2 + \theta + \lambda_{21} - \lambda_{21} g_1},$$

where, for brevity, g_i means $g_i(\theta)$. This gives

$$g_1 = \frac{\mu_1\lambda_{21} - \mu_2\lambda_{12} + (\mu_1 + \theta + \lambda_{12})(\mu_2 + \theta + \lambda_{21}) - \sqrt{\Delta}}{2\lambda_{21}(\mu_1 + \theta + \lambda_{12})},$$
$$g_2 = \frac{-\mu_1\lambda_{21} + \mu_2\lambda_{12} + (\mu_1 + \theta + \lambda_{12})(\mu_2 + \theta + \lambda_{21}) - \sqrt{\Delta}}{2\lambda_{12}(\mu_2 + \theta + \lambda_{21})},$$

where

$$\Delta = [\mu_1\mu_2 + \lambda_{12}\lambda_{21} + \theta^2 + \theta(\mu_1 + \mu_2 + \lambda_{12} + \lambda_{21})]^2 - 4\mu_1\mu_2\lambda_{12}\lambda_{21}.$$

6.2 Busy period asymptotics

In this section, we offer some observations on the tail asymptotics of the busy period in the case of heavy-tailed service time distributions. For light-tailed service time distributions, we refer the reader to the recent work of Palmowski and Rolski [16]. For the current case of heavy tails, we refer the reader to Zwart [22], Jelenković and Momcilović [13] and Denisov and Shneer [7].

The key idea in both Jelenković and Momcilović [13] and in Zwart [22] (as in many other instances of heavy-tailed behavior) is the principle of *one big jump*. For busy periods, this leads us to expect a large busy period to occur as consequence of one large service time. For concreteness, consider the standard $M/G/1$ queue with $\rho < 1$ and suppose there is a single large service time of size $S = z$. The workload after the large jump is $u + z$ for some small or moderate u. The workload then decreases at the rate $1 - \rho$ until it reaches 0 and the busy period terminates. By the Law of Large Numbers (LLN), the time of termination is approximately $(z + u)/(1 - \rho)$. Since the time before the big jump can be neglected, we have $B > x$ if and only if $z > (1 - \rho)x$. Both Asmussen [2] and Foss and Zachary [8] show the probability of this large jump is asymptotically equal to $\mathbb{P}\big(S > (1 - \rho)x\big)\mathbb{E}\sigma$ for large x, where σ is the number of customers served in a busy period. But $\mathbb{E}\sigma = \sum_{n=0}^{\infty} \rho^n = 1/(1 - \rho)$. Indeed, 1 corresponds to the customer initiating the busy period, ρ is the number of customers arriving while he is in service (the first generation), ρ^2 is the number of customers arriving while they are in service, and so forth. In the framework of branching processes, ρ^n is the number of individuals in the nth generation. These considerations lead to

$$\mathbb{P}(B > x) \sim \frac{1}{1-\rho}\mathbb{P}\big(S > (1-\rho)x\big), \tag{6.5}$$

which Jelenković and Momcilović [13] show to be the correct asymptotics if the service time distribution is subexponential and square root insensitive, i.e., with a heavier tail than $e^{-\sqrt{x}}$.

Generalizing this approach to our multiclass system, we recall that $\beta_i = \sum_{j=1}^{K} \lambda_{ij}\mathbb{E}B_j$ and we introduce a subexponential and square root insensitive reference distribution F for which the individual service time distributions are related as

$$\overline{F}_i\big(x/(1+\beta_i)\big) \sim c_i\overline{F}(x). \tag{6.6}$$

In practice, one chooses $\overline{F}(x)$ as $\sup_i \overline{F}_i\big(x/(1+\beta_i)\big)$. This is common in heavy-tailed studies involving distributions with different degrees of heavy-tailedness. In particular, it allows some F_j to be light-tailed ($c_j = 0$).

Recalling the interpretation of β_i as the rate of arriving work while a class i customer is in service, a big service time S_i of a class i customer will lead to $B_i > x$ precisely when $S_i(1+\beta_i) > x$. Using the same reasoning as for (6.5), we first note that $(\mathbf{M}^n)_{ij}$ is the number of type j progeny of a type i ancestor. Hence, if $\rho(\mathbf{M}) < 1$, the probability that one of these large service times occur in $[0, B_i)$ is approximately

$$\sum_{n=0}^{\infty}\sum_{j=1}^{K}(\mathbf{M}^n)_{ij}\overline{F}_j\big(x/(1+\beta_j)\big) \sim d_i\overline{F}(x),$$

where

$$d_i = \sum_{n=0}^{\infty}\sum_{j=1}^{K}(\mathbf{M}^n)_{ij}c_j = \sum_{j=1}^{K}(\mathbf{I}-\mathbf{M})^{-1}_{ij}c_j.$$

Equivalently, the d_i solve

$$d_i = c_i + \sum_{j=1}^{K} m_{ij}d_j\,. \tag{6.7}$$

As for the standard $M/G/1$ queue, it is straightforward to verify that this is an asymptotic lower bound.

Proposition 1 *Assume that F in (6.6) is subexponential with finite mean, so that $c_k > 0$ for some k and $\rho(\mathbf{M}) < 1$. Then, for each $i = 1, \ldots, K$,*

$$\liminf_{x\to\infty}\frac{\mathbb{P}(B_i > x)}{\overline{F}(x)} \geq d_i. \tag{6.8}$$

Remark 1 Square root insensitivity of F is not needed for Proposition 1. The assumption

$$\overline{F}_i(x) \sim \widetilde{c}_i \overline{F}_0(x) \tag{6.9}$$

may apriori be more appealing than (6.6) since it does not involve evaluation of the β_i. However, it is closely related. The reason is that if F is regularly varying with $\overline{F}(x) = L(x)/x^\alpha$, then (6.6) and (6.9) with $F_0 = F$ are equivalent, with the constants related by $c_i = \widetilde{c}_i(1+\beta_i)^\alpha$. For F_0 lognormal or Weibull with tail e^{-x^δ} (where $\delta < 1/2$ in the square root insensitive case), one has, for $\gamma_1 > \gamma_2$, $\overline{F}_0(\gamma_1 x) = o\big(\overline{F}_0(\gamma_2 x)\big)$. Hence, if (6.9) holds, we may define $\beta^* = \max_1^K \beta_j$ and take $\overline{F}(x) = \overline{F}_0\big(x/(1+\beta^*)\big)$, where $c_j = 1$ if $\beta_j = \beta^*$ and $c_j = 0$ if $\beta_j < \beta^*$.

The $M/G/1$ literature leads to the conjecture that further contributions to $\mathbb{P}(B_i > x)$ can be neglected, i.e., that $\mathbb{P}(B_i > x) \sim d_i \overline{F}(x)$ in the square root insensitive case. However, the upper bound is much more difficult (even in the single-class $M/G/1$ setting) and follows in the regular varying case from more general results recently established in Asmussen & Foss [1]:

Theorem 7 *Assume, in addition to the conditions of Proposition 1, that F is regularly varying. Then, $\mathbb{P}(B_i > x) \sim d_i \overline{F}(x)$ for each $i = 1, \ldots, K$.*

In the proof, we need:

Lemma 3 *Let S be subexponential and let the conditional distribution of N given $S = s$ be Poisson(λs). Then, $\mathbb{P}(S+N > x) \sim \mathbb{P}\big(S(1+\lambda) > x\big)$ as $x \to \infty$. Further, the conditional distribution of $(S, N)/(S+N)$ given $S + N > x$ converges to the one-point distribution at $((1/1+\lambda), \lambda/(1+\lambda))$.*

Proof The argument is standard, with the key intuition being that the variation in S dominates that of the Poisson distribution, so that N can be replaced by its conditional expectation λS given S. Firstly, note that if x is so large that $x - x^{1/2} > 2\lambda x^{1/2}$ and $N(x^{1/2})$ is Poisson$(\lambda x^{1/2})$, then

$$\mathbb{P}(S+N > x, S < x^{1/2}) \le \mathbb{P}\big(x^{1/2} + N(x^{1/2}) > x\big) \le \mathbb{P}\big(N(x^{1/2}) > 2\lambda x^{1/2}\big),$$

which (by large deviations theory) tends to zero faster than $\mathrm{e}^{-\delta x^{1/2}}$ for some $\delta > 0$, and hence faster than $\mathbb{P}\big(S(1+\lambda) > x\big)$. Secondly, $N/S \to \lambda$ as $y \to \infty$ given $S > y$ and so

$$\mathbb{P}(S+N > x, S \ge x^{1/2}) \sim \mathbb{P}\big(S(1+\lambda) > x, S \ge x^{1/2}\big),$$

the latter equaling $\mathbb{P}\big(S(1+\lambda) > x\big)$ for large x. This proves the first statement, and the second follows since (asymptotically) only large values of S contribute to large values of $S+N$, and in this regime $N/S \sim \lambda$. □

The setup of [1] is a set of random variables $(B_1, \ldots, B_K)$ satisfying

$$B_i \stackrel{\mathcal{D}}{=} S_i + \sum_{j=1}^{K} \sum_{m=1}^{N_{j;i}} B_{m;j}. \tag{6.10}$$

The assumptions for (6.10) are that all $B_{m;j}$ are independent of the vector $(S_i, N_{1;i}, \ldots, N_{K;i})$, that they are mutually independent, and that $B_{m;j} \stackrel{\mathcal{D}}{=} B_j$. Further, all random variables are nonnegative. In our multiclass queue, B_i is the length of the busy period initiated by a class i customer, S_i is the service time, and $N_{j;i}$ is the number of class j customers arriving during his service. In the following, we omit the index i and instead express the dependence on i in terms of a governing probability measure $\mathbb{P}_i$.

Proof of Theorem 7 To apply the results of [1], we first need to verify a condition on multivariate regular variation (see [18] for background) of the vector $(S, N_1, \ldots, N_K)$. Its first part is that $\mathbb{P}(S+N_1+\cdots+N_K > x) \sim b_i \overline{F}(x)$ for some b_i. This is immediate from Lemma 3 by taking $N = N_1 + \cdots + N_K$, $\lambda = \overline{\lambda}_i$, $b_i = \widetilde{c}_i(1+\overline{\lambda}_i)^\alpha$. A minor extension of the proof of Lemma 3 further yields that, given $S + N_1 + \cdots + N_K > x$,

$$\frac{1}{S+N_1+\cdots+N_K}\big(S, N_1, \ldots, N_K\big) \to \frac{1}{1+\overline{\lambda}_i}\big(1, \lambda_{i1}, \ldots, \lambda_{iK}\big), \qquad (6.11)$$

where the limit is taken as $x \to \infty$. This establishes the second part, namely the existence of the so-called angular measure (in this case a one-point distribution at the right-hand side of (6.11)).

It now follows from [1] that $\mathbb{P}(B_i > x) \sim d_i^* \overline{F}(x)$, where the d_i^* solve the set of linear equations

$$d_i^* = c_i^* + \sum_{j=1}^{K} m_{ij} d_j, \qquad (6.12)$$

and

$$c_i^* = \lim_{x\to\infty} \frac{1}{\overline{F}(x)} \mathbb{P}_i(S + N_1\overline{r}_1 + \cdots + N_K\overline{r}_K > x), \quad \text{with } \overline{r}_j = \mathbb{E}_j B.$$

Comparing with (6.7), we see that we need only check that $c_i^* = c_i$. By similar arguments to those above,

$$\begin{aligned}\mathbb{P}_i(S + N_1\overline{r}_1 + \cdots + N_K\overline{r}_K > x) &\sim \mathbb{P}(S(1+\lambda_{i1}\overline{r}_1 + \cdots + \lambda_{iK}\overline{r}_K) > x)\\ &= \mathbb{P}(S(1+\beta_i) > x) \sim \widetilde{c}_i(1+\beta_i)^\alpha \overline{F}(x) = c_i \overline{F}(x),\end{aligned}$$

where parts (ii) and (iii) of Lemma 2 are employed in the second step. □

Remark 2 The general subexponential case seems much more difficult. One obstacle is that theory and applications of multivariate subexponentiality are much less developed than for the regular varying case. See, however, Samorodnitsky and Sun [19] for a recent contribution and for further references.

7 Conclusion

We have introduced a multiclass single-server queueing model in which the arrival rates depend on the current job in service. The model departs from existing state-dependent models in the literature in which the parameters depend primarily on the number of jobs in the system rather than the job in service.

The main contributions of this paper can be summarized as follows. Firstly, we formulate the multiclass queueing model and its corresponding fluid model and provide motivation for its practical importance. The necessary and sufficient conditions for stability of the queueing system are obtained via the corresponding fluid model. Secondly, by appealing to the natural connection with multitype Galton–Watson processes, we utilize Laplace–Stieltjes transforms to characterize the busy period of the queueing system. Thirdly, we present a preliminary study of busy period tail asymptotics for heavy-tailed service time distributions and give a complete set of results for the regularly varying case, using recent results of Asmussen and Foss [1]. Tail asymptotics in our multiclass setting for nonregularly varying heavy-tailed service time distributions, as well as for light-tailed service time distributions, are much more difficult and will be attempted in a separate manuscript.

Acknowledgements We are very grateful to a referee for pointing out a problem in our initial proof of the upper bound in Sect. 6.2. We also thank a second referee and an associate editor for many useful suggestions. The first author thanks Dr. Quan Zhou and Dr. Guodong Pang for helpful conversations. The first author is grateful to the Dobelman Family for support in the form of the Dobelman Family Junior Chair. Finally, the first author gratefully acknowledges the support of ARO-YIP-71636-MA, NSF DMS-1811936, and ONR N00014-18-1-2192.

References

1. Asmussen, S., Foss, S.: Regular variation in a fixed-point problem for single- and multiclass branching processes and queues. arXiv:1709.05140. Accepted for Advances in Applied Probability, vol. 50A (Festschrift for Peter Jagers) (2017)
2. Asmussen, S.: Subexponential asymptotics for stochastic processes: extremal behavior, stationary distributions and first passage probabilities. Ann. Appl. Probab. **8**, 354–374 (1998)
3. Bekker, R., Borst, S.C., Boxma, O.J., Kella, O.: Queues with workload-dependent arrival and service rates. Queueing Syst. **46**, 537–556 (2004)
4. Berman, A., Plemmons, R.J.: Nonnegative Matrices in the Mathematical Sciences. Academic Press, Cambridge (1979)
5. Bramson, M.: Stability of Queueing Networks. Springer, Berlin (2008)
6. Cruz, F.R.B., Smith, J.M.: Approximate analysis of M/G/c/c state-dependent queueing networks. Comput. Oper. Res. **34**, 2332–2344 (2007)
7. Denisov, D., Shneer, S.: Global and local asymptotics for the busy period of an $M/G/1$ queue. Queueing Syst. **64**, 383–393 (2010)
8. Foss, S., Zachary, S.: The maximum on a random time interval of a random walk with long-tailed increments and negative drift. Ann. Appl. Probab. **13**, 37–53 (2003)
9. Gamarnik, D.: Fluid models of queueing networks. In: Wiley Encyclopedia of Operations Research and Management Science (2010)
10. Gantmacher, F.R.: Matrix Theory. Chelsea Publishing Company, New York (1960)
11. Harris, T.E.: The Theory of Branching Processes. Springer, Berlin (1963)
12. Jain, R., Smith, M.J.: Modeling vehicular traffic flow using M/G/C/C state-dependent queueing models. Transp. Sci. **31**, 324–336 (1997)

13. Jelenković, P., Momcilović, P.: Large deviations of square root insensitive random sums. Math. Oper. Res. **29**, 398–406 (2004)
14. Miller, D.R.: Computation of steady-state probabilities for M/M/1 priority queues. Oper. Res. **29**, 945–958 (1981)
15. Neuts, M.F.: The Markov renewal branching process. In Proc. Conf, Mathematical Methods in the Theory of Queues, Kalamazoo (1974)
16. Palmowski, Z., Rolski, T.: On the exact asymptotics of the busy period in GI/G/1 queues. Adv. Appl. Probab. **38**, 792–803 (2006)
17. Perry, D., Stadje, W., Zacks, S.: A duality approach to queues with service restrictions and storage systems with state-dependent rates. J. Appl. Probab. **50**, 612–631 (2013)
18. Resnick, S.: Heavy-Tail Phenomena: Probabilistic and Statistical Modeling. Springer, Berlin (2007)
19. Samorodnitsky, G., Sun, J.: Multivariate subexponential distributions and their applications. Extremes **19**, 171–196 (2016)
20. Wolff, R.W.: Stochastic Modeling and the Theory of Queues. Prentice-Hall, Upper Saddle River (1989)
21. Yuhaski, S.J., Smith, J.M.: Modeling circulation systems in buildings using state-dependent queueing models. Queueing Syst. **4**, 319–338 (1989)
22. Zwart, B.: Tail asymptotics for the busy period in the GI/G/1 queue. Math. Oper. Res. **26**, 485–493 (2001)