Operations Research Letters 46 (2018) 492–499

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

Time-varying many-server finite-queues in tandem: Comparing blocking mechanisms via fluid models

Noa Zychlinski [a,*], Petar Momčilović [b], Avishai Mandelbaum [a]

[a] Faculty of Industrial Engineering and Management, Technion–Israel Institute of Technology, Haifa 3200003, Israel
[b] Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA

ARTICLE INFO

Article history:
Received 28 March 2018
Received in revised form 6 June 2018
Accepted 8 July 2018
Available online 18 July 2018

Keywords:
Many-server flow lines with blocking
Communication blocking
Blocking Before Service (BBS)
Time-varying queueing networks
Fluid models
Functional strong law of large numbers

ABSTRACT

This paper focuses on the mechanism of Blocking Before Service (BBS), in time-varying many-server queues in tandem. BBS arises in telecommunication networks, production lines and healthcare systems. We model a stochastic tandem network under BBS and develop its corresponding fluid limit, which includes reflection due to jobs lost. Comparing our fluid model against simulation shows that the model is accurate and effective. This gives rise to design/operational insights regarding network throughput, under both BBS and BAS (Blocking After Service).

1. Introduction

Tandem queueing networks with blocking arise in many communication, production and service systems [6,29,30]. This paper focuses on communication blocking, which is also known as Blocking Before Service (BBS) or two-stage blocking [6]. Under this mechanism, a service cannot begin at Station i if there is no available capacity (buffer space or idle server) at Station $i+1$.

1.1. Motivation and examples

Clearly, the BBS mechanism is prevalent in telecommunication networks [16,30,31]. However, BBS is not uncommon in production lines; for example, in the steel, plastic molding and food processing industries [19], as well as in the chemical and pharmaceutical industries [14]. In the latter, for example, a work-in-process can be unstable or unsafe and, thus, cannot be detained/blocked after certain processes but rather should be immediately transferred to crystallization. Therefore, a process/reaction in certain stations cannot begin before the crystallizer in the subsequent stations is available. BBS can also be found in healthcare systems, for example in short procedures such as cataract surgery, cardiac catheterization and hernia repair; the procedure begins only when there is available room for the patient in the recovery room. Other examples are the hospital boarding ward between the emergency department and the inpatient wards, and the emergency care chain of cardiac in-patient flow [12]. In this latter chain, patients are refused or diverted at the beginning (First cardiac Aid (FCA) and Coronary Care Unit (CCU)) due to unavailability of beds downstream the care chain.

Besides communication, manufacturing and healthcare systems, our fluid models with blocking also have the potential to support transportation implementations. Fluid models originated, in fact, from transportation networks, in which entities that flow through the system are animated as continuous fluid [9]. Such implementations could support/evaluate the practice of releasing cars to highways during rush hours [7], or estimate travel times by navigation software (autonomous vehicles).

1.2. Results

In this paper we develop (Section 2) a stochastic model for a many-server tandem network under the BBS mechanism, time-varying arrivals and finite buffers before the first station and between stations. This model includes reflection, since an arriving job is forced to leave the system if Station 1 is full. Then, using the Functional Strong Law of Large Numbers (FSLLN), we develop and prove a fluid limit of the stochastic model in the many-server regime: system capacity (number of servers) increases indefinitely jointly with demand (arrival rates). Fluid models have proven to be accurate approximations for time-varying stochastic models, which are otherwise intractable [18,21–24,33,34].

* Corresponding author.
E-mail address: noazy@tx.technion.ac.il (N. Zychlinski).

https://doi.org/10.1016/j.orl.2018.07.002
0167-6377/

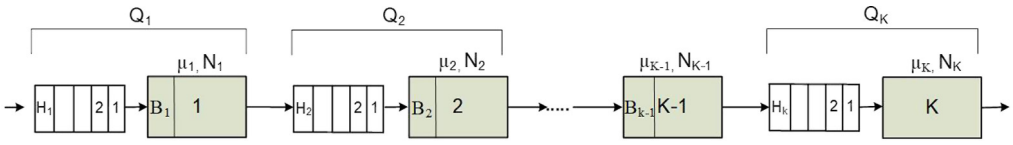

Fig. 1. A network with k stations in tandem under the BBS mechanism.

We establish existence and uniqueness of the fluid approximation, which is characterized by differential equations with reflection. In order to easily implement the differential equations numerically, we transform them into differential equations with a discontinuous right-hand side (RHS) [15,35], but no reflection. We validate the accuracy of our fluid models against stochastic simulation, which amplifies the simplicity and flexibility of fluid models in capturing the performance of time-varying networks altering between overloaded and underloaded periods.

Finally (Section 3), we develop steady-state closed-form expressions for the number of jobs in service at each station under the BAS (Blocking After Service) and BBS mechanisms. These expressions facilitate comparisons of network performances; in particular, comparing the number of jobs at each station, network throughput and job loss rate. In Section 3.2, we conclude the paper with an example of designing transfer protocols from surgery to recovery rooms in hospitals.

1.3. Brief literature review

There exists vast research on tandem flow lines with blocking [17,20,26]. However, research on time-varying multi-server flow lines is scarce. The most common types of blocking mechanisms for tandem flow lines are BAS and BBS [1,6,29]. The BBS mechanism can be sub-categorized into several types; we focus on Server Occupied, where a server can store a blocked job before its service begins [13]. Thus, under this mechanism, a job can enter Station i, but cannot begin service until there is available capacity (buffer space or idle server) at Station $i+1$. Another BBS mechanism is Server Not Occupied, where a blocked job cannot occupy a server. Thus, a job can enter a station (occupy a server), and begin its service, only when there is available capacity at the next station. We focus on BBS — Server Occupied, in order to compare it with the BAS mechanism, in which blocked jobs can also occupy servers [6].

In [5], a steady-state analysis under the BAS mechanism was conducted, for a single-server network with two tandem stations, Poisson arrival process and no intermediate buffers. This system was generalized to k stations with deterministic service times in [2] and to the BBS mechanism in [4]. Under the analyzed BBS, a job begins service at a station only when the next k stations are available. In [3], a k-station single-server network, with no intermediate buffers and an unlimited buffer before the first station, was analyzed under BAS and BBS. Note that the methodology we develop can, with slight modification (see Remark 2), accommodate any k-stage blocking, $k \geq 2$.

Approximation techniques, usually via the decomposition approach, were applied to tandem networks in steady-state under BAS [8,11,17,28,32]. Several papers have developed algorithms for approximating the steady-state throughput of closed single-server cyclic queueing networks with finite buffers (under both BBS and BAS in [27] and under BBS in [16,31]).

1.4. Contribution

Our contributions enrich existing models by adding predictable time variability, multi-server stations and a finite buffer before the first station, which leads to job loss when it is full. Moreover, we provide an analytic comparison between BBS and BAS, that yields operational insights. In particular, we quantify the differences between throughputs and job loss rate under BBS and BAS, including the conditions under which they coincide.

2. The model

2.1. Notations and assumptions

We model a network with k stations in tandem, as illustrated in Fig. 1.

This FCFS system is characterized, to a first order, by the following (deterministic) parameters:

1. Arrival rate to Station 1: $\lambda(t)$, $t \geq 0$;
2. Service rate $\mu_i > 0$, $i = 1, 2, \ldots, k$;
3. Number of servers N_i, $i = 1, 2, \ldots, k$;
4. Buffer size H_i, $i = 1, 2, \ldots, k$; H_i can vary from 0 to ∞, inclusive.

The stochastic model is created from the following stochastic building blocks: A, D_i, $Q_i(0)$, $i = 1, 2, \ldots, k$, all of which are assumed to be independent. Specifically:

1. External arrival process $A = \{A(t), t \geq 0\}$; A is a counting process, in which $A(t)$ represents the external cumulative number of arrivals up to time t; we assume the existence of

$$\mathbb{E}A(t) = \int_0^t \lambda(u)\mathrm{d}u, \qquad t \geq 0. \tag{1}$$

2. "Basic" nominal service processes $D_i = \{D_i(t), t \geq 0\}$, $i = 1, 2, \ldots, k$, where $D_i(t)$ is a standard (rate 1) Poisson process.
3. The stochastic process $Q = \{Q_1(t), \ldots, Q_k(t), t \geq 0\}$ denotes a stochastic queueing process in which $Q_i(t)$ represents the total number of jobs at Station i at time t (queued and in service).
4. Initial number of jobs in each station, denoted by $Q_i(0)$, $i = 1, 2, \ldots, k$.

2.2. The stochastic model

Service at Station i begins only when there is an available server at Station i and available capacity (idle server or buffer space) at Station $i+1$. If there is an available server at Station i, but no available capacity at Station $i+1$, the job is blocked at Station i (occupies a server, but not receiving service). If there is no available server at Station i, the job waits at Buffer i. If Buffer 1 is full, an arriving job is forced to leave the system and is lost. Note that in Fig. 1, B_i denotes the blocked jobs at Station i; their service is delayed until capacity becomes available at Station $i+1$.

The process Q, which represents the total number of jobs at each station, is characterized by the following equations:

$$\begin{aligned} Q_1(t) &= Q_1(0) + A(t) - \int_0^t 1_{\{Q_1(u-)=H_1+N_1\}}\mathrm{d}A(u) \\ &\quad - D_1\left(\mu_1 \int_0^t [Q_1(u) \wedge N_1 \wedge (H_2 + N_2 - Q_2(u))]\,\mathrm{d}u\right), \end{aligned} \tag{2}$$

$$\begin{aligned} Q_i(t) &= Q_i(0) + D_{i-1}\left(\mu_{i-1} \int_0^t [Q_{i-1}(u) \wedge N_{i-1} \wedge (H_i + N_i \right. \\ &\quad \left. - Q_i(u))\right]\mathrm{d}u\Big) \end{aligned}$$

$$- D_i\left(\mu_i \int_0^t [Q_i(u) \wedge N_i \wedge (H_{i+1} + N_{i+1} - Q_{i+1}(u))]\,\mathrm{d}u\right),$$
$$i = 2, \ldots, k-1;$$
$$Q_k(t) = Q_k(0) + D_{k-1}\left(\mu_{k-1} \int_0^t [Q_{k-1}(u) \wedge N_{k-1} \wedge (H_k + N_k - Q_k(u))]\,\mathrm{d}u\right)$$
$$- D_k\left(\mu_k \int_0^t [Q_k(u) \wedge N_k]\,\mathrm{d}u\right); \quad t \geq 0.$$

The integral in the first line of (2) represents the number of jobs that were forced to leave the system up until time t, as when they arrived, Station 1 was full. Note that when $H_1 = \infty$, the integral equals zero since no customers are forced to leave the system. This simplifies the model, since there is no reflection. The second line in (2) represents the number of jobs that completed service at Station 1, up until time t. Since the available storage capacity at Station 2 at time t is $H_2 + N_2 - Q_2(t)$, the term in the rectangle parenthesis represents the number of jobs at service in Station 1.

Now, we rewrite (2), as follows:

$$\begin{cases} \begin{bmatrix} Q_1(t) \\ Q_2(t) \\ \vdots \\ Q_k(t) \end{bmatrix} = \begin{bmatrix} Y_1(t) - L(t) \\ Y_2(t) \\ \vdots \\ Y_k(t) \end{bmatrix} \leq \begin{bmatrix} H_1 + N_1 \\ H_2 + N_2 \\ \vdots \\ H_k + N_k \end{bmatrix}, \quad t \geq 0, \\ \mathrm{d}L(t) \geq 0, \quad L(0) = 0, \\ \int_0^\infty 1_{\{Q_1(u-) < H_1 + N_1\}} \mathrm{d}L(u) = 0, \end{cases} \tag{3}$$

where

$$Y_1(t) = Q_1(0) + A(t) - D_1\left(\mu_1 \int_0^t [Q_1(u) \wedge N_1 \wedge (H_2 + N_2 - Q_2(u))]\,\mathrm{d}u\right),$$
$$Y_i(t) = Q_i(t), \quad i = 2, \ldots, k,$$
$$L(t) = \int_0^t 1_{\{Q_1(u-) = H_1 + N_1\}} \mathrm{d}A(u). \tag{4}$$

The last equation of (4) is a complementary relation between L and Q: $L(\cdot)$ increases at time t only if $Q_1(t) \geq H_1 + N_1$ (see [35], Section 2.1 for details).

We simplify (3), so that the reflection will occur at zero, by letting

$$R_i(t) = N_i + H_i - Q_i(t), \quad i = 1, \ldots, k, \quad t \geq 0, \tag{5}$$

which gives rise to the following equivalent to (3):

$$\begin{cases} \begin{bmatrix} R_1(t) \\ R_2(t) \\ \vdots \\ R_k(t) \end{bmatrix} = \begin{bmatrix} \tilde{Y}_1(t) + L(t) \\ \tilde{Y}_2(t) \\ \vdots \\ \tilde{Y}_k(t) \end{bmatrix} \geq \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad t \geq 0, \\ \mathrm{d}L(t) \geq 0, \quad L(0) = 0, \\ \int_0^\infty 1_{\{R_1(t) > 0\}} \mathrm{d}L(t) = 0, \end{cases} \tag{6}$$

where $\tilde{Y}_i = H_i + N_i - Y_i$. From (6), we see that $L(t) \geq -\tilde{Y}_1(t)$ and therefore, $L(t) = \sup_{0 \leq s \leq t}\left(-\tilde{Y}_1(s)\right)^+$. Note that this solution (or rather representation) applies even though $\tilde{Y}_1$ depends on R (see [25,35] for details).

2.3. Fluid approximation

We now develop a fluid limit for our queueing model through the Functional Strong Law of Large Numbers (FSLLN). We begin with (6) and scale up the arrival rate and the size of the system (servers and waiting rooms) by a factor of $\eta > 0$, $\eta \to \infty$. This parameter η will serve as an index of a corresponding queueing process R^η, which is the unique solution to the following Skorokhod's representation:

$$\begin{cases} R_1^\eta(t) = \tilde{Y}_1^\eta(t) + L^\eta(t), \\ R_i^\eta(t) = \tilde{Y}_i^\eta(t), \quad i = 2, \ldots k, \quad t \geq 0, \end{cases} \tag{7}$$

where

$$\tilde{Y}_1^{\,\eta}(\cdot) = R_1^\eta(0) - A^\eta(\cdot) + D_1\left(\mu_1 \int_0^\cdot \left[\left(\eta H_1 + \eta N_1 - R_1^\eta(u)\right) \wedge \eta N_1 \wedge R_2^\eta\right]\mathrm{d}u\right);$$
$$\tilde{Y}_i^{\,\eta}(\cdot) = R_i^\eta(0) - D_{i-1} \times \left(\mu_{i-1} \int_0^\cdot \left[\left(\eta H_{i-1} + \eta N_{i-1} - R_{i-1}^\eta(u)\right) \wedge \eta N_{i-1} \wedge R_i^\eta\right]\mathrm{d}u\right) + D_i\left(\mu_i \int_0^t \left[\left(\eta H_i + \eta N_i - R_i^\eta\right) \wedge \eta N_i \wedge R_{i+1}^\eta(u)\right]\mathrm{d}u\right),$$
$$i = 2, \ldots k-1;$$
$$\tilde{Y}_k^{\,\eta}(\cdot) = R_k^\eta(0) - D_{k-1}\left(\mu_{k-1} \int_0^\cdot \left[\left(\eta H_{k-1} + \eta N_{k-1} - R_{k-1}^\eta(u)\right) \wedge \eta N_{k-1} \wedge R_k^\eta\right]\mathrm{d}u\right) + D_k\left(\mu_k \int_0^\cdot \left[\left(\eta H_k + \eta N_k - R_k^\eta\right) \wedge \eta N_k\right]\mathrm{d}u\right);$$
$$L^\eta(\cdot) = \int_0^\cdot 1_{\{R_1^\eta(u-) = 0\}} \mathrm{d}A^\eta(u).$$

Here, $A^\eta = \{\eta A(t), \quad t \geq 0\}$ is the arrival process under our scaling; thus,

$$\mathbb{E}A^\eta(t) = \eta \int_0^t \lambda(u)\mathrm{d}u, \qquad t \geq 0.$$

We now introduce the scaled processes $r^\eta = \{r^\eta(t), \quad t \geq 0\}$, $l^\eta = \{l^\eta(t), t \geq 0\}$ and $y^\eta = \{y^\eta(t), t \geq 0\}$, by $r^\eta(t) = \eta^{-1}R^\eta(t)$, $l^\eta(t) = \eta^{-1}L^\eta(t)$, $y^\eta(t) = \eta^{-1}Y^\eta(t)$, respectively. Applying the methodology developed in [35], Theorem 1, yields the following asymptotic behavior of r^η. Suppose that, as $\eta \to \infty$

$$\left\{\eta^{-1}A^\eta(t),\ t \geq 0\right\} \to \left\{\int_0^t \lambda(u)du,\ t \geq 0\right\}, \quad u.o.c.\ \ a.s., \tag{8}$$

as well as

$$\lim_{\eta \to \infty} r^\eta(0) = r(0), \quad a.s., \tag{9}$$

where $r(0)$ is a given non-negative deterministic vector. Then, as $\eta \to \infty$, the family $\{r^\eta\}$ converges $u.o.c.$ over $[0, \infty)$, $a.s.$, to a deterministic function r. This r is the unique solution to the following differential equation (DE) with reflection:

$$\begin{cases} r_1(t) = r_1(0) - \int_0^t \left[\lambda(u) - \mu_1\left((H_1 + N_1 - r_1(u)) \wedge N_1 \wedge r_2(u)\right)\right]\mathrm{d}u + l(t) \geq 0, \\ r_i(t) = r_i(0) - \int_0^t \left[\mu_{i-1}\left((H_{i-1} + N_{i-1} - r_{i-1}(u)) \wedge N_i \wedge r_i(u)\right) - \mu_i\left((H_i + N_i - r_i(u)) \wedge N_i \wedge r_{i+1}(u)\right)\right]\mathrm{d}u \geq 0, \\ \qquad i = 2, \ldots, k-1; \\ r_k(t) = r_k(0) - \int_0^t \left[\mu_{k-1}\left((H_{k-1} + N_{k-1} - r_{k-1}(u)) \wedge N_{k-1} \wedge r_k(u)\right) - \mu_k\left((H_k + N_k - r_k(u)) \wedge N_k\right)\right]\mathrm{d}u \geq 0, \\ \mathrm{d}l(t) \geq 0, \quad l(0) = 0, \\ \int_0^\infty 1_{\{r_1(t) > 0\}} \mathrm{d}l(t) = 0. \end{cases} \tag{10}$$

The following proposition provides an equivalent representation to (10) in terms of our original formulation (i.e. $q(\cdot)$); see Appendix A for details. Implementing the solution in (11) numerically is straightforward since it is given by a set of differential equations with discontinuous RHS but, notable, without reflection.

Proposition 1. *The stochastic queueing family Q^η, $\eta > 0$ converges u.o.c. over $[0;1)$, a.s., as $\eta \to \infty$ to a deterministic function q. This q is the unique solution to the following differential equation (DE) with refection*

$$\begin{aligned}
q_1(t) &= q_1(0) - \mu_1 \int_0^t [q_1(u) \wedge N_1 \wedge (H_2 + N_2 - q_2(u))]\,du \\
&\quad + \int_0^t \Big[1_{\{q_1(u) < H_1+N_1\}} \cdot \lambda(u) \\
&\quad + 1_{\{q_1(u)=H_1+N_1\}} \cdot [\lambda(u) \wedge \mu_1 [N_1 \wedge (H_2 + N_2 - q_2(u))]]\,du, \\
q_i(t) &= q_i(0) + \mu_{i-1} \int_0^t [q_{i-1}(u) \wedge N_{i-1} \wedge (H_i + N_i - q_i(u))]\,du \\
&\quad - \mu_i \int_0^t [q_i(u) \wedge N_i \wedge (H_{i+1} + N_{i+1} - q_{i+1}(u))]\,du, \\
&\quad i = 2, \ldots, k-1; \\
q_k(t) &= q_k(0) + \mu_{k-1} \int_0^t [q_{k-1}(u) \wedge N_{k-1} \wedge (H_k + N_k - q_k(u))]\,du \\
&\quad - \mu_k \int_0^t [q_k(u) \wedge N_k]\,du. \qquad (11)
\end{aligned}$$

The function q will be referred to as the fluid limit associated with the queueing family Q^η.

Remark 1. The model can easily accommodate Markovian abandonments while being blocked or while waiting. To be more specific, let θ be the individual abandonment rate. Then, the abandonment rate of blocked jobs from each Buffer i, $i = 1, \ldots, k-1$, at time t would be $\theta[N_i - q_i(t) \wedge (H_{i+1} + N_{i+1} - q_{i+1}(t))]^+$; the abandonment rate of waiting jobs from Station i, $i = 1, \ldots, k$, at time t would be $\theta[q_i(t) - N_i]^+$. The mathematical analysis of models with abandonments does not differ from the one without.

Remark 2. The model can also easily accommodate a k-stage blocking mechanism, in which a job begins service at a station only if the next k stations are available. For example, accommodating the case where all downstream stations are required to be available, would be done by replacing the terms $\wedge(H_i + N_i - q_i(u))$, $i = 2, \ldots, k-1$, in (11) with $\wedge \bigwedge_{j=i}^{k} \left(H_j + N_j - q_j(u)\right)$.

In Appendix B we provide numerical examples demonstrating that our proposed fluid model accurately and effectively describes the flow of jobs in the networks, when compared against the average behavior of a stochastic simulation model.

3. Network performance

In this section we focus on steady-state performance, in particular network throughput and job loss rate under BBS and BAS (Section 3.1). The results we present were validated by discrete stochastic simulations. Let s_i and $\bar{q}_i$, $i = 1, \ldots, k$, denote the steady-state number of jobs in service and the steady-state number of jobs (including in the buffer) at Station i, respectively; thus,

$$\begin{aligned}
s_i &= \bar{q}_i \wedge N_i \wedge (H_{i+1} + N_{i+1} - \bar{q}_{i+1}), \qquad i = 1, \ldots, k-1, \qquad (12) \\
s_k &= \bar{q}_k \wedge N_k.
\end{aligned}$$

For calculating steady-state performance, we start with (11), set $\lambda(t) \equiv \lambda$, $t \geq 0$, and $q_i(0) = q_i(t) \equiv \bar{q}_i$, $\forall t \geq 0$, $i = 1, \ldots, k$. We then get that

$$\begin{aligned}
\mu_1 s_1 &= \lambda \cdot 1_{\{\bar{q}_1 < H_1+N_1\}} \\
&\quad + [\lambda \wedge \mu_1 (N_1 \wedge (H_2 + N_2 - \bar{q}_2))] \cdot 1_{\{\bar{q}_1 = H_1+N_1\}}, \\
\mu_{i-1} s_{i-1} &= \mu_i s_i, \qquad i = 2, \ldots, k. \qquad (13)
\end{aligned}$$

The following theorem identifies the "fluid" network throughput and the number of jobs in each station, in steady-state under BBS. The proof of the theorem is provided in Appendix C.

Theorem 1. *Let δ denote the network throughput in the fluid model. Then*

$$\delta = \mu_i s_i = \lambda \wedge \bigwedge_{j=1}^{k} \mu_j N_j \wedge \bigwedge_{j=2}^{k} \frac{H_j + N_j}{1/\mu_{j-1} + 1/\mu_j}, \qquad i = 1, \ldots, k. \qquad (14)$$

When $\delta = \lambda$, then $\bar{q}_j = \lambda/\mu_j$, $j = 1, \ldots, k$. Otherwise (when $\delta < \lambda$),

$$\begin{aligned}
&\bar{q}_1 = H_1 + N_1; \qquad (15) \\
&\bar{q}_j = H_j + N_j - \delta/\mu_{j-1}, \qquad j = 2, \ldots, i; \\
&\bar{q}_j = \delta/\mu_j, \qquad j = i+1, \ldots, k;
\end{aligned}$$

here

$$i = \min \left\{ \arg\min \bigwedge_{j=1}^{k} \mu_j N_j,\ \arg\min \bigwedge_{j=2}^{k} \frac{H_j + N_j}{1/\mu_{j-1} + 1/\mu_j} \right\}. \qquad (16)$$

The interpretation of (14) is that the network throughput is determined according to the minimum among the arrival rate, the processing capacity of the bottleneck (i.e. the slowest station when all servers are occupied) and the processing capacity of a "virtual" bottleneck, formed by two sequential stations. This is similar in spirit to [10], wherein the authors defined a virtual workload condition for the stability of a two-station multi-class fluid network. As in our case, two stations form a "virtual" bottleneck that determines the processing capacity of the entire network.

Note that H_1, the buffer size before the first station, does not affect network throughput. That is because network throughput depends on the arrival rate and the processing capacities of the actual/virtual bottleneck. Increasing only the first buffer, even to infinity, will not affect the network processing capacity.

3.1. Blocking after service

Thus far, we focused on the BBS mechanism. Another common blocking mechanism is BAS (Blocking After Service, also known as manufacturing blocking) [6]. Under BAS, a service begins at Station i when there is an available server there. If upon completion of a service, there is no available capacity (idle server or buffer space) at Station $i+1$, the job is blocked at Station i while occupying a server there. Fig. 2 illustrates the tandem network we analyze under manufacturing blocking. Note that the blocked jobs are placed at the end of each station, rather than at the beginning, as was in Fig. 1. This change seems small but it is not: as shown momentarily, it can significantly affect network performance (see Fig. 3).

The BAS mechanism for time-varying many-server flow lines was analyzed in [35].

We now compare the performance of the two mechanisms. In particular, we are interested in analyzing network throughput. Let δ^x denote the steady-state throughput under mechanism x, $x \in \{\text{BAS}, \text{BBS}\}$ (from now on, δ in (14) will be referred to as δ^{BBS}); s_i^x, $i = 1, \ldots, k$, denotes the steady-state number of jobs in service, at Station i under mechanism x. Applying to BAS the same methodology as we used for BBS (see Eq. (15) in [35], with $\lambda(t) \equiv \lambda$, $\forall t \geq 0$), yields the following BAS throughput:

$$\delta^{\text{BAS}} = \mu_i s_i^{\text{BAS}} = \lambda \wedge \bigwedge_{j=1}^{k} \mu_j N_j, \qquad i = 1, \ldots, k. \qquad (17)$$

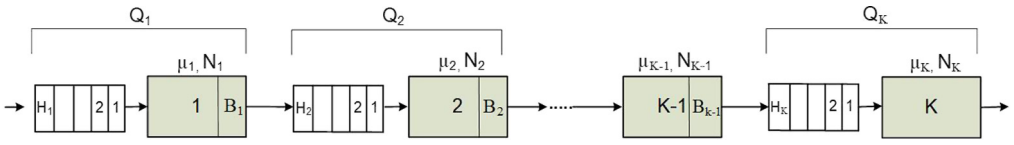

Fig. 2. A network with k stations in tandem under the BAS mechanism.

Remark 3. Note that H_i, $i = 1, \ldots, k$, the buffer sizes throughout the network, do not affect network throughput under BAS, which depends solely on the arrival rate and the bottleneck processing capacity. The intuition behind this phenomenon stems from considering the context in which our fluid models are applicable: networks with many-server stations. In the limiting operational regime we consider, the dependency on buffers in preventing starvation and idleness decreases, since stochastic fluctuations are negligible on the fluid scale. In fact, buffers affect only second-order phenomena (stochastic variability) but not the limiting (fluid) throughput which depends only on the Law of Large Numbers (LLN). Under BBS, however, the internal buffers affect network throughput (14), since they influence the bottleneck processing capacity.

Remark 4. The throughput under BBS, when adding sufficient buffer space after each server, will be equal to the throughput under BAS for the same network without the additional buffer spaces. This follows from our equations: When $H_j \geq N_{j-1}$, then

$$\frac{H_j + N_j}{1/\mu_{j-1} + 1/\mu_j} \geq \frac{\mu_j \mu_{j-1} N_{j-1}}{\mu_{j-1} + \mu_j} + \frac{\mu_{j-1} \mu_j N_j}{\mu_{j-1} + \mu_j} \geq \mu_{j-1} N_{j-1} \wedge \mu_j N_j.$$

Hence, the term that involves buffers (the third term in (14)) does not determine the throughput, and we get that $\delta^{BBS} = \delta^{BAS}$.

Fig. 3 presents the total number of jobs in service at each station under the two mechanisms. In plots A–C the arrival rate function is the sinusoidal function

$$\lambda(t) = \bar{\lambda} + \beta \sin(\gamma t), \qquad t \geq 0, \tag{18}$$

with average arrival rate $\bar{\lambda}$, amplitude β and cycle length $T = 2\pi/\gamma$.

Note the sharp decrease in the number of jobs at Station 1 under BBS (the blue dashed lines) close to the origin. The reason for this is the empty system at the outset. As the two stations begin to fill, that increases the number of blocked jobs at Station 1 and, therefore, the number of jobs in service decreases.

Combining (14) and (17) yields the following:

$$\delta^{\text{BBS}} = \delta^{\text{BAS}} \wedge \bigwedge_{j=2}^{k} \frac{H_j + N_j}{1/\mu_{j-1} + 1/\mu_j},$$

thus, $\delta^{\text{BBS}} \leq \delta^{\text{BAS}}$. The throughputs are equal when $\delta^{BAS} \leq \bigwedge_{j=2}^{k} \frac{H_j+N_j}{1/\mu_{j-1}+1/\mu_j}$; an example for such a case can be seen in Fig. 3, Plot D. The reason why the throughput under BBS is smaller or equal to the throughput under BAS is capacity loss under the former. Capacity loss occurs when servers remain idle, while waiting for service to end at their previous station. This capacity loss also increases the rate of job loss, $\gamma \equiv \lambda - \delta$, which occurs when the first station is full and arriving jobs are forced to leave; thus

$$\gamma^{\text{BBS}} = \left[\lambda - \left[\bigwedge_{i=1}^{k} \mu_i N_i \wedge \bigwedge_{i=2}^{k} \frac{H_i + N_i}{1/\mu_{i-1} + 1/\mu_j}\right]\right]^+ \geq \left[\lambda - \bigwedge_{i=1}^{k} \mu_i N_i\right]^+ = \gamma^{\text{BAS}}.$$

3.2. Example in a surgery-room setting

In this section, we demonstrate how our models can yield design/operational insights in a hospital setting that includes surgery rooms (Station 1) and recovery rooms (Station 2). After a surgery is completed, the patient is transferred to the recovery room. If there are no available beds in the recovery room, the patient is blocked at the surgery room, while preventing it from being cleaned and prepared for the next surgery. To avoid such situations, in some hospitals a surgery begins only when there is an available bed in the recovery room. Is this a worthwhile strategy?

In deciding on the preferable mechanism, we consider two performance measures: throughput and sojourn time. The former is calculated by (14) and (17); the latter is calculated by first calculating the number of patients in the system (Theorem 1) and then, by applying Little's law in steady-state (i.e. dividing the total number of customers by the throughput). Let $\mu_1 = 1/60$, $\mu_2 = 1/60$, $N_1 = 10$, $N_2 = 0$, $H_1 = 10$, $H_2 = 0$ and $\lambda = 1/6$ (time units are measured in minutes). This setting corresponds to cataract surgeries, for example; under it, both BAS and BBS behave the same with average throughput of 10 patients per hour and average sojourn time of 2 h. Now, suppose that recovery takes on average 2 h (instead of one), as in hernia repair for example; then, the throughput under BAS remains 10 patients per hour, but the throughput under BBS is reduced to 6.67 patients per hour. Moreover, while the average sojourn time under BAS is 3 h, under BBS it reaches 5 h. Under this setting, BAS is superior according to both performance measurements.

Acknowledgments

The authors thank Yale T. Herer for valuable discussions and suggesting Remark 4. The work of A.M. has been partially supported by BSF grant 2014180 and ISF grants 357/80 and 1955/15. The work of P.M. has been partially supported by NSF - Division of Civil, Mechanical & Manufacturing Innovation (CMMI) grant 1362630 and BSF grant 2014180. The work of N.Z. has been partially supported by The Israeli Ministry of Science, Technology and Space, and the Technion—Israel Institute of Technology.

Appendix A. Proof of Proposition 1

From (10), we return to our original formulation in terms of $q(\cdot)$ for $t \geq 0$, as follows:

$$\begin{cases} q_1(t) = q_1(0) + \int_0^t [\lambda(u) - \mu_1 (q_1(u) \wedge N_1 \wedge (H_2 + N_2 - q_2(u)))]\, du \\ \quad - l(t) \leq H_1 + N_1, \\ q_i(t) = q_i(0) + \int_0^t \Big[\mu_{i-1} (q_{i-1}(u) \wedge N_{i-1} \wedge (H_i + N_i - q_i(u))) \\ \quad - \mu_i (q_i(u) \wedge N_i \wedge (H_{i+1} + N_{i+1} - q_{i+1}(u)))\Big] du \\ \quad \leq H_i + N_i, \quad i = 2, \ldots, k-1; \\ q_k(t) = q_k(0) + \int_0^t \Big[\mu_{k-1} (q_{k-1}(u) \wedge N_{k-1} \wedge (H_k + N_k - q_k(u))) \\ \quad - \mu_i (q_k(u) \wedge N_k)\Big] du \leq H_k + N_k, \\ dl(t) \geq 0, \quad l(0) = 0, \\ \int_0^\infty 1_{\{q_1(u-) < H_1 + N_1\}} dl(t) = 0. \end{cases} \tag{A.1}$$

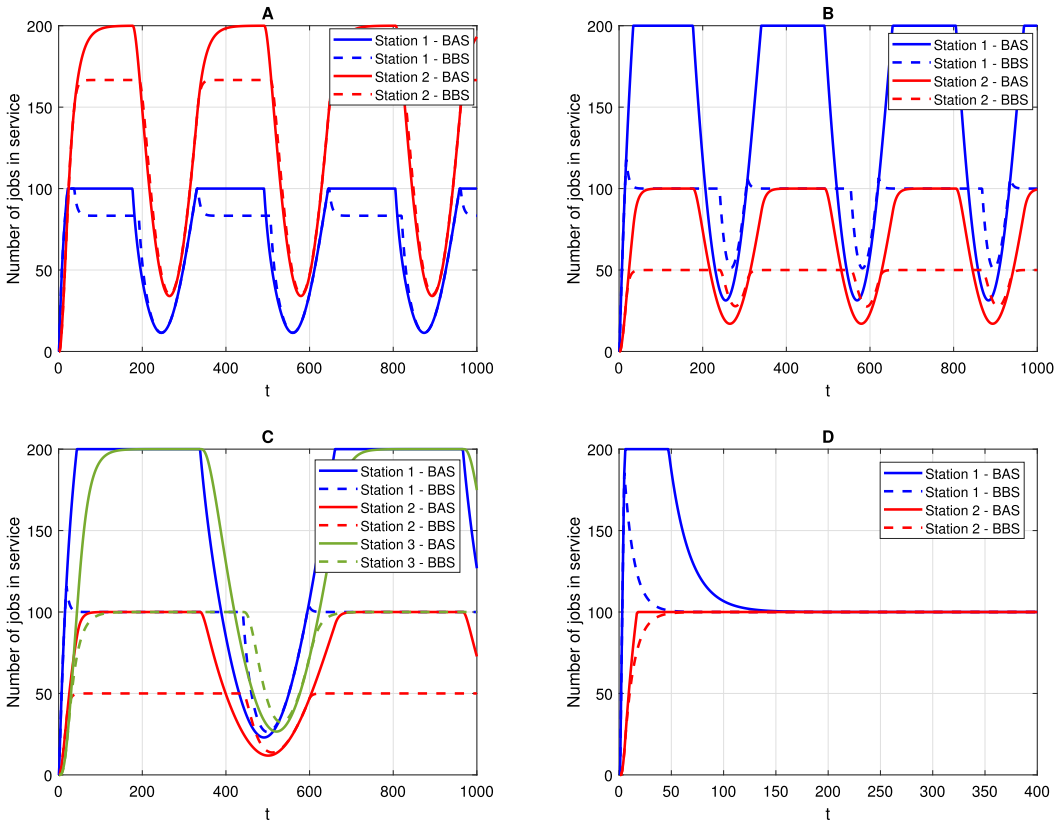

Fig. 3. Total number of jobs in service at each station — BBS vs. BAS with $q(0) = 0$. In Plot A, the sinusoidal arrival rate function in (18) with $\bar{\lambda} = 9$, $\beta = 8$ and $\gamma = 0.02$, $N_1 = 100$, $N_2 = 200$, $H_1 = H_2 = 50$, $\mu_1 = 1/10$, $\mu_2 = 1/20$. In Plot B, the station order was replaced. In Plot C, $\gamma = 0.01$ and a third station is added having $N_3 = 200$, $H_3 = 50$, $\mu_3 = 1/20$. In Plot D, $\lambda(t) = 20$, $t \geq 0$, $N_1 = 200$, $N_2 = 100$ and $\mu_1 = \mu_2 = 1/20$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Now, we prove that the solution for (A.1) satisfies

$$l(t) = \int_0^t 1_{\{q_1(u) \geq H_1 + N_1\}} [\lambda(u) - l_1(u)]^+ \mathrm{d}u, \quad t \geq 0, \tag{A.2}$$

where

$$l_1(u) = \mu_1 \left(q_1(u) \wedge N_1 \wedge (H_2 + N_2 - q_2(u))\right).$$

In order to prove this, we substitute (A.2) in the equation of $q_1(t)$ in (A.1) and show that the properties in (A.1) prevail:

$$\begin{aligned} q_1(t) &= q_1(0) + \int_0^t \left[\lambda(u) - \mu_1 \left(q_1(u) \wedge N_1 \wedge (H_2 + N_2 - q_2(u))\right)\right] \mathrm{d}u \\ &\quad - \int_0^t 1_{\{q_1(u) \geq H_1 + N_1\}} \left[\lambda(u) - \lambda(u) \wedge l_1(u)\right] \mathrm{d}u \\ &= q_1(0) + \int_0^t \Big[1_{\{q_1(u) < H_1 + N_1\}} \cdot \lambda(u) \\ &\quad - \mu_1 \left(q_1(u) \wedge N_1 \wedge (H_2 + N_2 - q_2(u))\right)\Big] \mathrm{d}u \\ &\quad + \int_0^t \left[1_{\{q_1(u) \geq H_1 + N_1\}} \cdot (\lambda(u) \wedge l_1(u))\right] \mathrm{d}u. \end{aligned} \tag{A.3}$$

Clearly, the properties in the last two lines in (A.1) prevail. It is left to verify that the first k conditions prevail. This is done by the following proposition.

Proposition 2. *The functions $q_i(\cdot)$, $i = 1, \ldots, k$, as in (A.3) are bounded by $H_i + N_i$, respectively.*

Proof. First we prove that the function $q_1(\cdot)$, as in (A.3), is bounded by $H_1 + N_1$. Assume that for some t, $q_1(t) > H_1 + N_1$. Since $q_1(0) \leq H_1 + N_1$ and q_1 is continuous (being an integral), there must be a last $\tilde{t}$ in $[0, t]$ such that $q_1(\tilde{t}) = H_1 + N_1$ and $q_1(u) > H_1 + N_1$, for $u \in [\tilde{t}, t]$. Without loss of generality, assume that $\tilde{t} = 0$; thus $q_1(0) = H_1 + N_1$ and $q_1(u) > H_1 + N_1$ for $u \in (0, t]$. From (A.3), we get that

$$\begin{aligned} q_1(t) &= H_1 + N_1 \\ &\quad + \int_0^t \left[(\lambda(u) \wedge l_1(u)) - \mu_1 \left(q_1(u) \wedge N_1 \wedge (H_2 + N_2 - q_2(u))\right)\right] \mathrm{d}u \\ &\leq H_1 + N_1 \\ &\quad + \int_0^t \left[l_1(u) - \mu_1 \left(q_1(u) \wedge N_1 \wedge (H_2 + N_2 - q_2(u))\right)\right] \mathrm{d}u = H_1 + N_1, \end{aligned}$$

which contradicts our assumption and proves that $q_1(\cdot)$ cannot exceed $H_1 + N_1$.

What is left to prove now is that the functions $q_i(\cdot)$, $i = 2, \ldots, k$, are bounded by $H_i + N_i$. Without loss of generality, assume that $q_i(0) = H_i + N_i$ and $q_i(u) > H_i + N_i$ for $u \in (0, t]$. Hence, from (A.1), we get that

$$\begin{aligned} q_i(t) &= H_i + N_i + \int_0^t \Big[\mu_{i-1} \left(q_{i-1}(u) \wedge N_{i-1} \wedge (H_i + N_i - q_i(u))\right) \\ &\quad - \mu_i \left(q_i(u) \wedge N_i \wedge (H_{i+1} + N_{i+1} - q_{i+1}(u))\right)\Big] \mathrm{d}u \\ &\leq H_i + N_i, \end{aligned}$$

which contradicts the assumption that $q_i(t) > H_i + N_i$ and proves that $q_i(\cdot)$, $i = 1, \ldots, k$, are bounded by $H_i + N_i$.

By the solution uniqueness (see Appendix C in [35]), we have established that q, the fluid limit for the stochastic queueing family Q^η in (2), is given by (11). Note that after proving that $q_1(\cdot) \leq H_1 + N_1$ in Proposition 2, the indicators in (A.2) can accommodate only the case when $q_1(\cdot) = H_1 + N_1$.

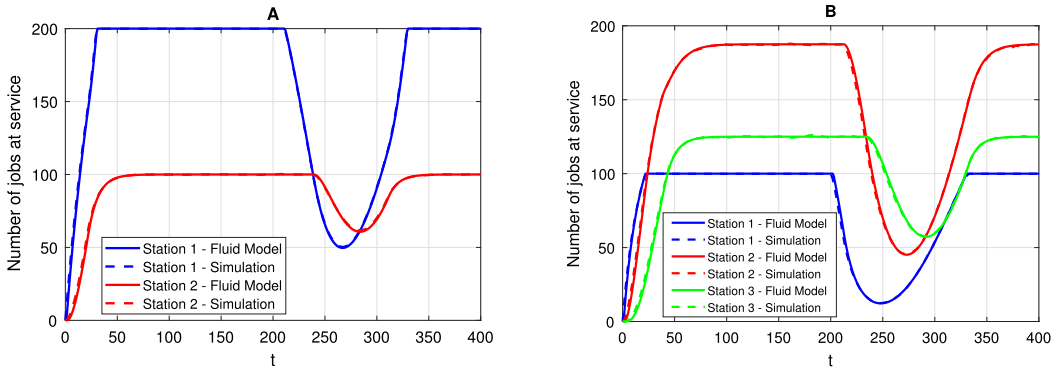

Fig. B.4. Total number of jobs at service — fluid model vs. simulation results, the sinusoidal arrival rate function in (18) with $\bar{\lambda} = 9$, $\beta = 8$ and $\gamma = 0.02$, $q_i(0) = 0$. In Plot A, $\mu_1 = \mu_2 = 1/20$, $H_1 = H_2 = 50$, $N_1 = 200$, $N_2 = 150$; in Plot B, $\mu_1 = 1/10$, $\mu_2 = 1/20$, $\mu_3 = 1/20$, $H_1 = H_2 = H_3 = 50$, $N_1 = 100$, $N_2 = 200$ and $N_3 = 200$.

Appendix B. Numerical examples

To demonstrate that our proposed fluid model accurately describes the flow of jobs in the networks, we compared it to a simulation model. In the simulation model, jobs arrive according to a non-homogeneous Poisson process that was used to represent a process with a general, time-dependent arrival rate. Service treatment was randomly generated from exponential distributions.

Solving the fluid equations in (11) was done by recursion and time discretization. Fig. B.4 shows the comparison between the total number of jobs at each station according to the fluid model (solid lines) and the average simulation results over 500 replications (dashed lines). These four examples, among many others, show that the fluid model accurately describes the underlying average of the stochastic system it approximates.

Appendix C. Proof of Theorem 1

Due to the uniqueness of q (Proposition 1), it suffices to show that δ and $\bar{q}_i$, $i = 1, \ldots, k$, in Eqs. (14)–(16) satisfy the model equations in (11). In particular, it suffices to show that the steady-state equations in (13) are satisfied. Since the second equation in (13) is trivially satisfied, one is left only with the first equation.

When $\delta = \lambda$ and $\bar{q}_j = \lambda/\mu_j, j = 1, \ldots, k$, the first line in (13) yields the following:

$$\begin{aligned}\lambda =& \lambda \cdot 1_{\{\lambda < \mu_1(H_1+N_1)\}} \\ &+ [\lambda \wedge \mu_1 (N_1 \wedge (H_2 + N_2 - \lambda/\mu_2))] \cdot 1_{\{\lambda = \mu_1(H_1+N_1)\}}.\end{aligned} \tag{C.1}$$

The first right-hand side term trivially satisfies the equation. The second right-hand-side term is larger than zero when $\lambda = \mu_1(H_1 + N_1)$. When $\delta = \lambda$, from (14) we know that $\lambda \le \mu_1 N_1$. Therefore, the second indicator in (C.1) equals one when $H_1 = 0$ and $\lambda = \mu_1 N_1$. In this case, the second right-hand side term is $\lambda \wedge \mu_1 N_1 \wedge \mu_1(H_2 + N_2 - \mu_1 N_1/\mu_2) = \mu_1 N_1 = \lambda$. The second equality derives from (14): when $\delta = \lambda$, we get that $\lambda = \mu_1 N_1 \le (H_2 + N_2)/(1/\mu_1 + 1/\mu_2)$, which is equivalent to $N_1 \le H_2 + N_2 - \mu_1 N_1/\mu_1$. Therefore, (C.1) is satisfied. It is easy to show that the second line in (13) is also satisfied by $\bar{q}_j = \lambda/\mu_j, j = 1, \ldots, k$.

Now, when $\delta < \lambda$, from (13) we get that $\bar{q}_1 = H_1 + N_1$ (the first indicator in the first line is zero), and we get that

$$\delta = \lambda \wedge \mu_1 (N_1 \wedge (H_2 + N_2 - \bar{q}_2)) = \mu_1 (N_1 \wedge (H_2 + N_2 - \bar{q}_2)). \tag{C.2}$$

If Station 1 is the first bottleneck ($i = 1$, in (16)) then, from (12) and (14), we get that $\delta = \mu_1 N_1 \le \mu_1(H_2 + N_2 - \mu_1 N_1/\mu_2)$; therefore, (C.2) is satisfied with $\bar{q}_2 = \delta/\mu_2$.

Otherwise, if Station 1 is not the bottleneck then, $\delta < \mu_1 N_1$. Since $\bar{q}_1 = H_1 + N_1$, from (12) we get that $\delta = \mu_1(H_2 + N_2 - \bar{q}_2)$ and therefore, $\bar{q}_2 = H_2 + N_2 - \delta/\mu_1$. We obtain that $\delta = (\mu_1 N_1) \wedge \delta$, which satisfies Eq. (C.2).

For completing the proof for $\bar{q}_i$, $i = 3, \ldots, k$, in (15), we analyze separately the stations before the first bottleneck (inclusive) and the stations after it. We begin with the stations before the bottleneck. Suppose that Station i, $3 \le i \le k$, is the first bottleneck. From (12) we get that $\delta = \mu_2 [\bar{q}_2 \wedge N_2 \wedge (H_3 + N_3 - \bar{q}_3)]$. Since $\delta < \mu_2 N_2$, we get that $\delta = \mu_2 [\bar{q}_2 \wedge (H_3 + N_3 - \bar{q}_3)]$. Assume that $\bar{q}_2$ is the minimum, then $\bar{q}_2 = \delta/\mu_2 = H_2 + N_2 - \delta/\mu_1$ and therefore, $\delta = (H_2 + N_2)/(1/\mu_1 + 1/\mu_2)$, which contradicts the assumption that Station i is the first bottleneck. Hence, $\delta = \mu_2(H_3 + N_3 - \bar{q}_3)$ and $\bar{q}_3 = H_3 + N_3 - \delta/\mu_2$. We iteratively continue this argument up until the first bottleneck.

For the stations after the bottleneck, suppose that Station i, $2 \le i \le k - 1$, is the first bottleneck. From (12) and (13), we get that $\delta = \mu_{i+1} [\bar{q}_{i+1} \wedge N_{i+1} \wedge (H_{i+2} + N_{i+2} - \bar{q}_{i+2})]$. When $\bar{q}_{i+1} = \delta/\mu_{i+1}$ and $\bar{q}_{i+2} = \delta/\mu_{i+2}$, we get that $\delta = \delta \wedge \mu_{i+1} N_{i+1} \wedge \mu_{i+1}(H_{i+2} + N_{i+2} - \delta/\mu_{i+2})$. Since i is the first bottleneck, then $\delta \le \mu_{i+1} N_{i+1}$, as well as $\delta \le (H_{i+2} + N_{i+2})/(1/\mu_{i+1} + 1/\mu_{i+2})$, which is equivalent to $\delta \le \mu_{i+1}(H_{i+2} + N_{i+2} - \delta/\mu_{i+2})$. Hence, (13) is satisfied. We iteratively continue this argument up until Station k.

References

[1] T. Altiok, Approximate analysis of exponential tandem queues with blocking, Eur. J. Oper. Res. 11 (4) (1982) 390–398.
[2] B. Avi-Itzhak, A sequence of service stations with arbitrary input and regular service times, Manag. Sci. 11 (5) (1965) 565–571.
[3] B. Avi-Itzhak, S. Halfin, Servers in tandem with communication and manufacturing blocking, J. Appl. Probab. (1993) 429–437.
[4] B. Avi-Itzhak, H. Levy, A sequence of servers with arbitrary input and regular service times revisited: In memory of Micha Yadin, Manag. Sci. 41 (6) (1995) 1039–1047.
[5] B. Avi-Itzhak, M. Yadin, A sequence of two servers with no intermediate queue, Manag. Sci. 11 (5) (1965) 553–564.
[6] S. Balsamo, V. de Nitto Personé, R. Onvural, Analysis of Queueing Networks with Blocking, Springer, 2001.
[7] P. Bickel, C. Chen, J. Kwon, J. Rice, P. Varaiya, E. van Zwet, Traffic flow on a freeway network, in: Nonlinear Estimation and Classification, Springer, 2003, pp. 63–81.
[8] A. Brandwajn, Y. Jow, An approximation method for tandem queues with blocking, Oper. Res. 36 (1) (1988) 73–83.
[9] C. Daganzo, V. Gayah, E. Gonzales, The potential of parsimonious models for understanding large scale transportation systems and answering big picture questions, EURO J. Transp. Logist. 1 (1–2) (2012) 47–65.
[10] J. Dai, J. Vande Vate, The stability of two-station multitype fluid networks, Oper. Res. 48 (5) (2000) 721–744.
[11] Y. Dallery, Y. Frein, On decomposition methods for tandem queueing networks with blocking, Oper. Res. 41 (2) (1993) 386–399.
[12] A. De Bruin, A. Van Rossum, M. Visser, G. Koole, Modeling the emergency cardiac in-patient flow: An application of queuing theory, Health Care Manag. Sci. 10 (2) (2007) 125–137.
[13] J. Desel, M. Silva, Application and Theory of Petri Nets 1998: 19th International Conference, ICATPN98, Lisbon, Portugal, June 22–26, 1998 Proceedings, Springer, 1998.

[14] E. Dogan-Sahiner, T. Altiok, Blocking policies in pharmaceutical transfer lines, Ann. Oper. Res. 79 (1998) 323–347.
[15] A. Filippov, Differential Equations with Discontinuous Righthand Sides: Control Systems, Springer Science & Business Media, 2013.
[16] Y. Frein, Y. Dallery, Analysis of cyclic queueing networks with finite buffers and blocking before service, Perform. Eval. 10 (3) (1989) 197–210.
[17] S. Gershwin, An efficient decomposition method for the approximate evaluation of tandem queues with finite storage space and blocking, Oper. Res. 35 (2) (1987) 291–305.
[18] P. Guodong, W. Whitt, Heavy-traffic limits for many-server queues with service interruptions, Queueing Syst. 61 (2) (2009) 167–202.
[19] N. Hall, C. Sriskandarajah, A survey of machine scheduling problems with blocking and no-wait in process, Oper.Res. 44 (3) (1996) 510–525.
[20] J. Li, S. Meerkov, Production Systems Engineering, Springer Science & Business Media, 2009.
[21] Y. Liu, W. Whitt, Large-time asymptotics for the $G_t/M_t/s_t + GI_t$ many-server fluid queue with abandonment, Queueing Syst. 67 (2) (2011) 145–182.
[22] Y. Liu, W. Whitt, Many-server heavy-traffic limit for queues with time-varying parameters, Ann. Appl. Probab. 24 (1) (2014) 378–421.
[23] A. Mandelbaum, W. Massey, M. Reiman, Strong approximations for Markovian service networks, Queueing Syst. 30 (1–2) (1998) 149–201.
[24] A. Mandelbaum, W. Massey, M. Reiman, B. Rider, Time varying multiserver queues with abandonment and retrials, in: Proceedings of the 16th International Teletraffic Conference, vol. 4, 1999, pp. 4–7.
[25] A. Mandelbaum, G. Pats, State-dependent queues: Approximations and applications, Stoch. Netw. 71 (1995) 239–282.
[26] S. Meerkov, C.B. Yan, Production lead time in serial lines: Evaluation, analysis, and control, IEEE Trans. Autom. Sci. Eng. 13 (2) (2016) 663–675.
[27] R. Onvural, H. Perros, Approximate throughput analysis of cyclic queueing networks with finite buffers, IEEE Trans. Softw. Eng. 15 (6) (1989) 800–808.
[28] C. Osorio, M. Bierlaire, An analytic finite capacity queueing network model capturing the propagation of congestion and blocking, Eur. J. Oper. Res. 196 (3) (2009) 996–1007.
[29] H. Perros, Queueing Networks with Blocking, Oxford University Press, Inc., 1994.
[30] D.W. Seo, H.C. Lee, S.S. Ko, Stationary waiting times in m-node tandem queues with communication blocking, Manag. Sci. Financial Eng. 14 (1) (2008) 23–34.
[31] R. Suri, G. Diehl, A new 'building block' for performance evaluation of queueing networks with finite buffers, in: ACM SIGMETRICS Performance Evaluation Review, vol. 12, ACM, 1984, pp. 134–142.
[32] M. van Vuuren, I. Adan, S. Resing-Sassen, Performance analysis of multi-server tandem queues with finite buffers and blocking, OR Spect. 27 (2–3) (2005) 315–338.
[33] W. Whitt, Efficiency-driven heavy-traffic approximations for many-server queues with abandonments, Manage. Sci. 50 (10) (2004) 1449–1461.
[34] W. Whitt, Fluid models for multiserver queues with abandonments, Oper. Res. 54 (1) (2006) 37–54.
[35] N. Zychlinski, A. Mandelbaum, P. Momčilović, Time-varying tandem queues with blocking: modeling, analysis and operational insights via fluid models with reflection, Queueing Syst. 89 (1–2) (2018) 15–47.