Queueing Syst (2018) 89:15–47
https://doi.org/10.1007/s11134-018-9578-x

Time-varying tandem queues with blocking: modeling, analysis, and operational insights via fluid models with reflection

Noa Zychlinski[1] · **Avishai Mandelbaum**[1] · **Petar Momčilović**[2]

Received: 23 July 2017 / Revised: 18 January 2018 / Published online: 28 March 2018

Abstract In this paper, we develop time-varying fluid models for tandem networks with blocking. Beyond having their own intrinsic value, these mathematical models are also limits of corresponding many-server stochastic systems. We begin by analyzing a two-station tandem network with a general time-varying arrival rate, a finite waiting room before the first station, and no waiting room between the stations. In this model, customers that are referred from the first station to the second when the latter is saturated (blocked) are forced to wait in the first station while occupying a server there. The finite waiting room before the first station causes customer loss and, therefore, requires reflection analysis. We then specialize our model to a single station (many-server fluid limit of the $G_t/M/N/(N+H)$ queue), generalize it to k stations in tandem, and allow finite internal waiting rooms. Our models yield operational insights into network performance, specifically on the effects of line length, bottleneck location, waiting room size, and the interaction among these effects.

Dedicated to Ward Whitt, on the occasion of his 75th birthday, in gratitude for his inspiring scholarship, and long-lasting leadership, friendship, and mentorship.

✉ Noa Zychlinski
noazy@tx.technion.ac.il

Avishai Mandelbaum
avim@tx.technion.ac.il

Petar Momčilović
petar@ise.ufl.edu

[1] Faculty of Industrial Engineering and Management, Technion – Israel Institute of Technology, Haifa, Israel

[2] Department of Industrial and Systems Engineering, University of Florida, Gainesville, USA

Keywords Fluid models · Tandem queueing networks with blocking · Time-varying queues · Reflection · Flow lines with blocking · Functional Strong Law of Large Numbers

Mathematics Subject Classification 60K25 · 90B22

1 Introduction

Blocking is an important phenomenon in service, computer, communication, and manufacturing systems (for example, [10,62]). This has motivated our paper, in which we analyze several stochastic models of time-varying tandem queues with blocking. For each such model, we develop and prove its fluid limit in the many-server regime: System capacity (number of servers) increases indefinitely jointly with demand (arrival rates). We adopt a fluid framework since it yields accurate approximations for time-varying models, which are otherwise notoriously intractable. In fluid models, entities that flow through the system are animated as continuous fluid, and hence the system dynamics can be captured by differential equations. There is ample literature justifying that fluid models accurately approximate heavily loaded service systems [42,46,48,49,59,75,77].

The models we focus on (flow lines) have been researched for decades [5,6,41,53]; our research takes the analysis to the new territories of time-varying environments and many-server stations. Such general models are also applicable in modeling healthcare environments and the bed-blocking phenomenon [21,36,56,58,65,69,81] in particular. This phenomenon occurs when a patient remains hospitalized after treatment completion due to lack of beds in a more appropriate facility (for example, a rehabilitation or geriatric ward). In that case, the patient occupies/blocks a hospital bed and thus prevents the admittance of another patient from the Emergency Department (ED); this may block the ED as well. Blocking in healthcare systems is pervasive (see [14]) between surgery rooms, recovery rooms, and internal wards.

Our basic model (Sect. 2) is a network with two queues in tandem (Fig. 1), where the arrivals follow a general time-varying counting process. There is a finite waiting room before the first station and no waiting room between the two stations. There are two types of blocking in this network. The first occurs when the first station is saturated (all its servers are occupied and its waiting room is full), and therefore, arriving customers must leave the system (are blocked); such customer loss is mathematically captured by reflection. The second type of blocking occurs when the second station is saturated (all its servers are busy); in this case, customers who complete their service at the first station are forced to wait there while still occupying their server. Such a mechanism is known as *blocking after service* (BAS) or *manufacturing blocking* [10,15]; and here, as it turns out (see [81]), an appropriate state representation renders reflection unnecessary for capturing this type of blocking. A real system that is naturally modeled by such two queues in tandem is an ED feeding a hospital ward; servers here are hospital beds.

Using the Functional Strong Law of Large Numbers for all our stochastic models, we establish the existence and uniqueness of fluid approximations/limits. These are first characterized by differential equations with reflection, which are then transformed into differential equations with no reflection but rather with discontinuous right-hand

side (RHS) [24]; the latter are easier to implement numerically. The accuracy of our fluid models is validated against stochastic simulation, which amplifies the simplicity and flexibility of fluid models in capturing the performance of time-varying overloaded networks.

The two-station network is both specialized and extended. First, we derive a fluid limit for the $G_t/M/N/(N+H)$ queue that seems, to the best of our knowledge, already new. Next, in Sect. 3, we analyze the more general network with k queues in tandem and finite waiting rooms throughout—both before the first station and in-between stations. It is worth noting that our models cover all waiting room options at all locations: finite positive, infinite, or zero (no waiting allowed) and that reflection arises only due to having a finite waiting room before the first station.

Finally, in Sect. 4, we provide operational insights regarding the performance of time-varying tandem queues with finite buffers. Due to space considerations, in this paper we chose to calculate performance measures from the customer viewpoint: throughput, number of customers, waiting times, blocking times, and sojourn times; performance is measured at each station separately as well as overall within the network. (One could also easily accommodate server-oriented metrics, such as occupancy levels or starvation times.) Calculations of the above customer-driven measures provide insights into how network characteristics affect performance: We focus on line length (number of queues in tandem), bottleneck location, size of waiting rooms, and their joint effects.

1.1 Literature review

Despite the fact that time-varying parameters are common in production [38,55] and service systems [23,29], such as in healthcare [4,17,80], research on time-varying models with blocking is scarce. We now review the three research areas most relevant to this work.

Tandem queueing models with blocking Previous research on tandem queueing networks with blocking has focused on steady-state analysis for small networks [2,28,37], steady-state approximations for larger networks [9,13,19,26,58,62,67,68,71], and simulation models [14,18,21,34,54].

Several papers have analyzed tandem queueing networks with an unlimited waiting room before the first station and a blocking-after-service mechanism between the stations. In [7], the steady state of a single-server network with two stations in tandem was analyzed. In this model, the arrival process was Poisson and there was no waiting room between stations. The transient behavior of the same network was analyzed in [63]. The model in [7] was extended in [5] to an ordered sequence of single-server stations with a general arrival process, deterministic service times, and finite waiting room between the stations. The author concluded that the order of stations and the size of the intermediate waiting rooms do not affect the sojourn time in the system. We extend the analysis in [5] to time-varying arrivals, a finite waiting room before the first station, exponential service times, and a different number of servers in each station. We show how the order of stations does affect the sojourn time and how it interacts with the waiting room capacity before the first station.

The system analyzed in [7] was generalized in [6] under *blocking-before-service* (BBS) (or k-stage blocking mechanism) in which a customer enters a station only if the next k stations are available. A tandem queueing network with a single server at each station and no buffers between the stations was analyzed in [35]; the service times for each customer are identical at each station. In [73], heuristics were developed for ordering the stations in a tandem queueing network to minimize the sojourn time in the system. In this setting, each station has a single server and an unlimited waiting room. Simulation was employed in [18] to analyze work in process (WIP) in serial production lines, with and without buffers in balanced and unbalanced lines. The results of [27] were extended in [52] for analyzing tandem queueing networks with finite capacity queues and blocking. In that work, the author estimated the asymptotic behavior of the time customer n finishes service at Station k, as n and k become large together. Single-server flow lines with unlimited waiting rooms between the stations and exponential service times were investigated in [53]. The authors derived formulas for the average sojourn time (waiting and processing times). In our models, in addition to having time-varying arrivals, many-server stations, and finite waiting rooms, the sojourn time also includes blocking time at each station.

Fluid models with time-varying parameters Fluid models were successfully implemented in modeling different types of service systems. These models cover the early applications for post offices [57], claims processing in social security offices [70], call centers [1,29], and healthcare systems [17,80,81]. Fluid models of service systems were extended to include state-dependent arrival rates, and general arrival and service rates [76,77]. Time-varying queueing models were analyzed for setting staffing requirements in service systems with unlimited waiting rooms, by using the offered load heuristics [29,78,79].

Time-varying heavy-traffic fluid limits were developed in [48,49] for queueing systems with exponential service, abandonment, and retrial rates. Accommodating these models for general time-varying arrival rates and a general independent abandonment rate was done in [42] for a single station, and for a network in [43]. These models were extended to general service times in [44–46].

Heavy-traffic approximations for systems with blocking have focused on stationary loss models [11,12,66]. An approximation for the steady-state blocking probability, with service times being dependent and non-exponential, was developed in [39]. A recent work in [40] focused on stabilizing blocking probabilities in time-varying loss models. In our paper, we contribute to this research area by developing a heavy-traffic fluid limit for time-varying models with blocking.

Queueing models with reflection Queueing models with reflection were analyzed in [30] for an assembly operation by developing limit theorems for the associated waiting time process. There it was shown that this process cannot converge in distribution and thus is inherently unstable. This model is generalized in [72] by assuming finite capacities at all stations and developing a conventional heavy-traffic limit theorem for a stochastic model of a production system. The reflection analysis detailed in [16,31] for a single station and for a network is extended in [50,51] for state-dependent queues. Loss systems for one station with reflection were analyzed in [25,74]. More recently, [64] solved a generalized state-dependent drift Skorokhod problem in one dimension,

which is used to approximate the transient distribution of the $M/M/N/N$ queue in the many-server heavy-traffic regime.

1.2 Contributions

As we see it, the main contributions of this paper are the following:

1. *Modeling* We analyze a time-varying model for k many-server stations in tandem, with finite waiting rooms before the first station and between the other stations. This covers, in particular, the case of infinite or no waiting rooms, which includes the $G_t/M/N/(N+H)$ queue. For all these models, we derive a unified fluid model/approximation, which is characterized by a set of differential equations with a discontinuous right-hand side [24].
2. *Analysis of the stochastic model* We introduce a stochastic model for our family of networks in which, as usual, the system state captures station occupancy (for example, (4–5), for $k = 2$). It turns out, however, that a state description in terms of non-utilized servers is more amenable to analysis (7–8). Indeed, it enables a representation of the network in terms of reflection, which yields useful properties of the network reflection operator (for example, Lipschitz continuity).
3. *Analysis of the fluid model* Through the Functional Strong Law of Large Numbers, we derive a fluid limit for the stochastic model with reflection in the many-server regime. Using properties of the reflection operator, we solve for the fluid limit, which allows it to be written as a set of differential equations without reflection. This fluid representation is flexible, accurate and effective, hence, easily implementable for a variety of networks.
4. *Operational insights* Our fluid model yields novel operational insights for time-varying finite-buffer flow lines. Specifically (Sect. 4), via numerical experiments, we analyze the effects on network performance of the following factors: line length, bottleneck location, size of the waiting room, and the interaction among these factors.

2 Two stations in tandem with finite waiting room

We now develop a fluid model with blocking for two stations in tandem, as illustrated in Fig. 1. In Sect. 3, we further extend this model for a network with k stations in tandem and finite internal waiting rooms between the stations.

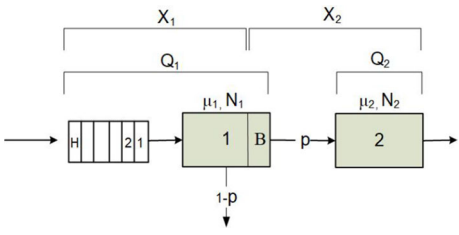

Fig. 1 Two tandem stations with a finite waiting room before the first station

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

1. Arrival rate $\lambda(t)$, $t \geq 0$, to Station 1.
2. Service rate $\mu_i > 0$, $i = 1, 2$.
3. Number of servers N_i, $i = 1, 2$.
4. Transfer probability p from Station 1 to Station 2, $0 \leq p \leq 1$ (i.e., with probability p, a customer will be referred to Station 2 upon completion of service at Station 1);
5. Finite waiting room H at Station 1; there is no waiting room at Station 2 ($H = 0$ is allowed; in this case, customers join the system only if there is an idle server in Station 1).

The stochastic model is created from the following stochastic building blocks, all of which are assumed to be independent:

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; here

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

   A special case is the non-homogeneous Poisson process, for which

$$A(t) = A_0\left(\int_0^t \lambda(u)\,\mathrm{d}u\right), \quad t \geq 0,$$

   where $A_0(\cdot)$ is a standard Poisson process (unit arrival rate).
2. "Basic" nominal service processes $D_i = \{D_i(t), t \geq 0\}$, $i = 1, 2, 3$, where $D_i(t)$ are standard Poisson processes.
3. The stochastic process $X_1 = \{X_1(t), t \geq 0\}$, which denotes the number of customers present at Station 1 that have *not* completed their service at Station 1 at time t.
4. The stochastic process $X_2 = \{X_2(t), t \geq 0\}$, which denotes the number of customers present at Station 1 or 2 that have completed service at Station 1, but not at Station 2, at time t.
5. Initial number of customers in each state, denoted by $X_1(0)$ and $X_2(0)$.

A customer is forced to leave the system if Station 1 is saturated (the waiting room is full, if a waiting room is allowed) upon its arrival. We assume that the blocking mechanism between Station 1 and Station 2 is blocking *after* service (BAS) [10]. Thus, if upon service completion at Station 1, Station 2 is saturated, the customer will be forced to stay in Station 1, occupying a server there until a server at Station 2 becomes available. This mechanism was modeled in [81] for a network with an infinite waiting room before Station 1. In our case, however, to accommodate customer loss, we must use reflection in our modeling and analysis.

Let $Q = \{Q_1(t), Q_2(t), t \geq 0\}$ denote a stochastic queueing process in which $Q_1(t)$ represents the number of customers at Station 1 (including the waiting room)

and $Q_2(t)$ represents the number of customers in service at Station 2 at time t. The process Q is characterized by the following equations:

$$\begin{aligned} Q_1(t) &= X_1(t) + B(t), \\ Q_2(t) &= X_2(t) \wedge N_2, \end{aligned}$$

where $B(t) = (X_2(t) - N_2)^+$ represents the number of blocked customers in Station 1, and

$$\begin{aligned} X_1(t) &= X_1(0) + \int_0^t 1_{\{X_1(u-)+(X_2(u-)-N_2)^+ < N_1+H\}} \, \mathrm{d}A(u) \\ &\quad - D_1\left(p\mu_1 \int_0^t [X_1(u) \wedge (N_1 - B(u))] \, \mathrm{d}u\right) \\ &\quad - D_3\left((1-p)\mu_1 \int_0^t [X_1(u) \wedge (N_1 - B(u))] \, \mathrm{d}u\right), \\ X_2(t) &= X_2(0) + D_1\left(p\mu_1 \int_0^t [X_1(u) \wedge (N_1 - B(u))] \, \mathrm{d}u\right) \\ &\quad - D_2\left(\mu_2 \int_0^t [X_2(u) \wedge N_2] \, \mathrm{d}u\right); \quad t \geq 0. \end{aligned} \tag{2}$$

Here, $1_{\{x\}}$ is an indicator function that equals 1 when x holds and 0 otherwise. The second right-hand term in the first equation of (2) represents the number of arrivals that entered service up to time t. As noted in [51], an inductive construction over time shows that (2) uniquely determines the process X. Observe that $X_1(t) + (X_2(t) - N_2)^+ = N_1 + H$ implies that the first station is blocked until the next departure.

2.1 Representation in terms of reflection

First, we rewrite (2) by using the fact that

$$\begin{aligned} &\int_0^t 1_{\{X_1(u-)+(X_2(u-)-N_2)^+ < N_1+H\}} \, \mathrm{d}A(u) \\ &\qquad = A(t) - \int_0^t 1_{\{X_1(u-)+(X_2(u-)-N_2)^+ = N_1+H\}} \, \mathrm{d}A(u); \end{aligned} \tag{3}$$

here, the last right-hand term represents the cumulative number of arrivals to Station 1 that were blocked because all N_1 servers were busy and the waiting room was full.

Now, we rewrite (2) and (3):

$$\begin{cases} \begin{bmatrix} X_1(t) \\ X_1(t) + X_2(t) \end{bmatrix} = \begin{bmatrix} Y_1(t) - L(t) \\ Y_2(t) - L(t) \end{bmatrix} \leq \begin{bmatrix} N_1 + H \\ N_1 + N_2 + H \end{bmatrix}, \quad t \geq 0, \\ \mathrm{d}L(t) \geq 0, \, L(0) = 0, \\ \int_0^\infty 1_{\{X_1(t)+(X_2(t)-N_2)^+ < N_1+H\}} \, \mathrm{d}L(t) = 0, \end{cases} \tag{4}$$

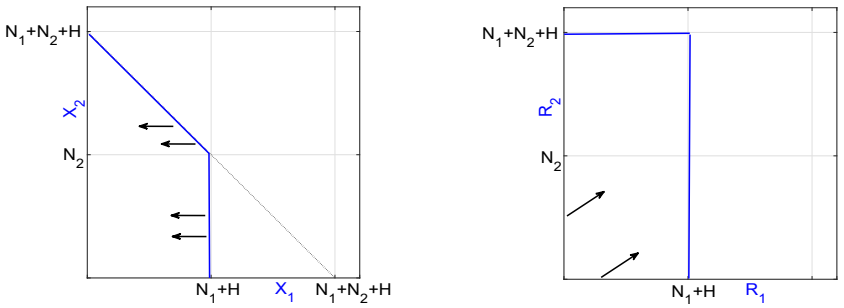

Fig. 2 Geometrical representation of the reflection. On the left—in terms of X, and on the right—in terms of R

where

$$
\begin{aligned}
Y_1(t) &= X_1(0) + A(t) - D_1\left(p\mu_1 \int_0^t [X_1(u) \wedge (N_1 - B(u))]\,\mathrm{d}u\right) \\
&\quad - D_3\left((1-p)\mu_1 \int_0^t [X_1(u) \wedge (N_1 - B(u))]\,\mathrm{d}u\right), \\
Y_2(t) &= X_1(0) + X_2(0) + A(t) - D_3\left((1-p)\mu_1 \int_0^t [X_1(u) \wedge (N_1 - B(u))]\,\mathrm{d}u\right) \\
&\quad - D_2\left(\mu_2 \int_0^t [X_2(u) \wedge N_2]\,\mathrm{d}u\right), \\
L(t) &= \int_0^t 1_{\{X_1(u-)+(X_2(u-)-N_2)^+ = N_1+H\}}\,\mathrm{d}A(u). \qquad (5)
\end{aligned}
$$

Figure 2 (left) geometrically illustrates the reflection in (4). The region for X_1 and X_2 is limited by the two blue lines. Arrivals are lost when the system is on the blue lines. The system leaves the state $X_1 = N_1 + H$ when a service is completed at Station 1. The system leaves the state $X_1 + X_2 = N_1 + N_2 + H$ when a service is completed at Station 2.

The last equation of (4) is a complementary relation between L and X: $L(\cdot)$ increases at time t only if $X_1(t) + (X_2(t) - N_2)^+ = N_1 + H$. We justify this by first substituting the last equation of (5) in the last equation for $L(t)$ of (4), which yields the following:

$$
\int_0^\infty 1_{\{X_1(t)+(X_2(t)-N_2)^+ < N_1+H\}} \cdot 1_{\{X_1(t-)+(X_2(t-)-N_2)^+ = N_1+H\}}\,\mathrm{d}A(t) = 0. \qquad (6)
$$

Now, if (6) does not hold, there must be a time when, at state N_1, a service completion and an arrival occur simultaneously. However, when $X_1 + (X_2 - N_2)^+ = N_1 + H$, the next departure will occur according to an exponential random variable; hence, by the independence of the building blocks, an arrival occurs simultaneously with a departure with probability 0.

We simplify (4), so that the reflection will occur on the axes, by letting

$$R_1(t) = N_1 + H - X_1(t),$$
$$R_2(t) = N_1 + N_2 + H - (X_1(t) + X_2(t)) = R_1(t) + N_2 - X_2(t), \quad t \geq 0.$$

Note that $R_1(t)$ represents the non-utilized space in Station 1 at time t, namely the blocked servers, the idle servers, and the available waiting room space. When all N_1 servers are occupied and the waiting room is full, $R_1(t)$ includes the blocked servers at Station 1. When all N_1 servers are occupied but the waiting room is not full, $R_1(t)$ includes the blocked servers and the available waiting room space. When some of the N_1 servers are idle, R_1 includes the sum of the idle servers, the blocked servers, and the available waiting room space. The function $R_2(t)$ represents the available space in the system at time t. Hence, when the $N_1 + N_2$ servers are occupied, $R_2(t)$ includes the available waiting room space. When only the N_2 servers are occupied but not all N_1 servers are occupied, $R_2(t)$ includes the idle servers in Station 1 and the available waiting room space. Finally, when Station 2 is not full, $R_2(t)$ includes the idle servers in Stations 1 and 2 and the available waiting room space.

The functions R_1 and R_2 give rise to the following, equivalent to (4):

$$\begin{cases} \begin{bmatrix} R_1(t) \\ R_2(t) \end{bmatrix} = \begin{bmatrix} \tilde{Y}_1(t) + L(t) \\ \tilde{Y}_2(t) + L(t) \end{bmatrix} \geq 0, \quad t \geq 0, \\ \mathrm{d}L(t) \geq 0, \, L(0) = 0, \\ \int_0^\infty 1_{\{R_1(t) \wedge R_2(t) > 0\}} \, \mathrm{d}L(t) = 0, \end{cases} \tag{7}$$

where

$$\tilde{Y}(t) = \begin{bmatrix} \tilde{Y}_1(t) \\ \tilde{Y}_2(t) \end{bmatrix} = \begin{bmatrix} N_1 + H - Y_1(t) \\ N_1 + N_2 + H - Y_2(t) \end{bmatrix}; \tag{8}$$

the last line in (7) is derived from

$$\int_0^t 1_{\{X_1(t) + (X_2(t) - N_2)^+ < N_1 + H\}} \, \mathrm{d}L(t) = \int_0^t 1_{\{N_1 + H - X_1(t) > (X_2(t) - N_2)^+\}} \, \mathrm{d}L(t)$$
$$= \int_0^t 1_{\{R_1(t) - (R_1(t) - R_2(t))^+ > 0\}} \, \mathrm{d}L(t) = \int_0^t 1_{\{R_1(t) \wedge R_2(t) > 0\}} \, \mathrm{d}L(t).$$

The processes $\tilde{Y}_1$, $\tilde{Y}_2$, and L (see (7)) can be stated in the "language" of R:

$$\begin{cases} \tilde{Y}_1(t) & = R_1(0) - A(t) + D_1 \left(p\mu_1 \int_0^t [(N_1 + H - R_1(u)) \wedge (N_1 - B(u))] \, \mathrm{d}u \right) \\ & \quad + D_3 \left((1-p)\mu_1 \int_0^t [(N_1 + H - R_1(u)) \wedge (N_1 - B(u))] \, \mathrm{d}u \right), \\ \tilde{Y}_2(t) & = R_2(0) - A(t) + D_3 \left((1-p)\mu_1 \int_0^t \left[\left(N_1 + H - R_1(u)\right) \wedge \left(N_1 - B(u)\right) \right] \mathrm{d}u \right) \\ & \quad + D_2 \left(\mu_2 \int_0^t \left[N_2 \wedge \left(R_1(u) - R_2(u) + N_2\right) \right] \mathrm{d}u \right), \\ L(t) & = \int_0^t 1_{\{R_1(u-) \wedge R_2(u-) = 0\}} \, \mathrm{d}A(u). \end{cases}$$

Here, $B(u) = \big(R_1(u) - R_2(u)\big)^+$ in terms of R.

Figure 2 (right) presents the direction of reflection in terms of R. When the process hits the boundary of the positive quadrant, L increases. This increase causes equal positive displacements in both R_1 and R_2 as necessary to keep $R_1 \geq 0$ and $R_2 \geq 0$, which drives L in the diagonal direction, presented in Fig. 2.

From (7), we see that $L(t) \geq -\tilde{Y}_1(t)$ and $L(t) \geq -\tilde{Y}_2(t)$. Therefore, $L(t) \geq \big(-\tilde{Y}_1(t) \vee -\tilde{Y}_2(t)\big) = -\big(\tilde{Y}_1(s) \wedge \tilde{Y}_2(s)\big)$, and

$$L(t) = \sup_{0 \leq s \leq t} \Big(-\Big(\tilde{Y}_1(s) \wedge \tilde{Y}_2(s)\Big)\Big)^+.$$

Note that this solution is applicable even though $\tilde{Y}$ depends on R (see [50] for details, though recall that they do not cover blocking).

2.2 Fluid approximation

We now develop a fluid limit for our queueing model through the Functional Strong Law of Large Numbers (FSLLN). We begin with (7) and scale up the arrival rate and the size of the system (servers and waiting room) by $\eta > 0$, $\eta \to \infty$. This parameter η will serve as an index of a corresponding queueing process R^η, which is the unique solution to the following Skorokhod representation:

$$\begin{cases} R_1^\eta(t) = \tilde{Y}_1^\eta(t) + L^\eta(t), \\ R_2^\eta(t) = \tilde{Y}_2^\eta(t) + L^\eta(t), \end{cases} \quad t \geq 0,$$

where

$$\begin{bmatrix} \tilde{Y}_1^\eta(\cdot) \\ \tilde{Y}_2^\eta(\cdot) \end{bmatrix} = \begin{bmatrix} R_1^\eta(0) - A^\eta(\cdot) + D_1\left(p\mu_1 \int_0^\cdot \left[(\eta N_1 + \eta H - R_1^\eta(u)) \wedge (\eta N_1 - B^\eta(u))\right] \mathrm{d}u\right) \\ \quad + D_3\left((1-p)\mu_1 \int_0^\cdot \left[(\eta N_1 + \eta H - R_1^\eta(u)) \wedge (\eta N_1 - B^\eta(u))\right] \mathrm{d}u\right) \\ R_2^\eta(0) - A^\eta(\cdot) + D_3\left((1-p)\mu_1 \int_0^\cdot \left[(\eta N_1 + \eta H - R_1^\eta(u)) \wedge (\eta N_1 - B^\eta(u))\right] \mathrm{d}u\right) \\ \quad + D_2\left(\mu_2 \int_0^\cdot \left[\eta N_2 \wedge (R_1^\eta(u) - R_2^\eta(u) + \eta N_2)\right] \mathrm{d}u\right) \end{bmatrix}.$$

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

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

We now introduce the scaled processes $r^\eta = \{r^\eta(t), t \geq 0\}$, $l^\eta = \{l^\eta(t), t \geq 0\}$ and $b^\eta = \{b^\eta(t), t \geq 0\}$ by

$$r^\eta(t) = \eta^{-1} R^\eta(t), \quad l^\eta(t) = \eta^{-1} L^\eta(t) \quad \text{and} \quad b^\eta(t) = \eta^{-1} B^\eta(t),$$

respectively; similarly, $\tilde{y}_1^\eta = N_1 + H - y_1^\eta$ and $\tilde{y}_2^\eta = N_1 + H + N_2 - y_2^\eta$. Then, we get that

$$\begin{bmatrix} \tilde{y}_1^\eta(\cdot) \\ \tilde{y}_2^\eta(\cdot) \end{bmatrix} = \begin{bmatrix} r_1^\eta(0) - \eta^{-1}A^\eta(\cdot) + \eta^{-1}D_1\left(\eta p\mu_1 \int_0^\cdot \left[(N_1 + H - r_1^\eta(u)) \wedge (N_1 - b^\eta(u))\right] \mathrm{d}u\right) \\ \quad + \eta^{-1}D_3\left(\eta(1-p)\mu_1 \int_0^\cdot \left[(N_1 + H - r_1^\eta(u)) \wedge (N_1 - b^\eta(u))\right] \mathrm{d}u\right) \\ r_2^\eta(0) - \eta^{-1}A^\eta(\cdot) + \eta^{-1}D_3\left(\eta(1-p)\mu_1 \int_0^\cdot \left[(N_1 + H - r_1^\eta(u)) \wedge (N_1 - b^\eta(u))\right] \mathrm{d}u\right) \\ \quad + \eta^{-1}D_2\left(\eta\mu_2 \int_0^\cdot \left[N_2 \wedge (r_1^\eta(u) - r_2^\eta(u) + N_2)\right] \mathrm{d}u\right) \end{bmatrix}. \tag{9}$$

The asymptotic behavior of r^η is described in the following theorem, which we prove in Appendix A.

Theorem 1 *Suppose that*

$$\left\{\eta^{-1}A^\eta(t),\ t \geq 0\right\} \to \left\{\int_0^t \lambda(u)du,\ t \geq 0\right\} \text{ u.o.c. as } \eta \to \infty,$$

and $r^\eta(0) \to r(0)$ a.s., as $\eta \to \infty$, where $r(0)$ is a given nonnegative 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:

$$\begin{cases} r_1(t) = r_1(0) - \int_0^t [\lambda(u) - \mu_1((N_1 + H - r_1(u)) \wedge (N_1 - b(u)))]\ \mathrm{d}u + l(t) \geq 0, \\ r_2(t) = r_2(0) - \int_0^t [\lambda(u) - (1-p)\mu_1((N_1 + H - r_1(u)) \wedge (N_1 - b(u)))]\ \mathrm{d}u \\ \qquad + \int_0^t [\mu_2(N_2 \wedge (r_1(u) - r_2(u) + N_2))]\ \mathrm{d}u + l(t) \geq 0, \\ \mathrm{d}l(t) \geq 0, l(0) = 0, \\ \int_0^\infty 1_{\{r_1(t) \wedge r_2(t) > 0\}}\ \mathrm{d}l(t) = 0; \end{cases} \tag{10}$$

where $b(t) = \left(r_1(t) - r_2(t)\right)^+$, $t \geq 0$.

Returning to our original formulation (4), (10) can in fact be written in terms of $x(\cdot)$ for $t \geq 0$ as follows:

$$\begin{cases} x_1(t) = x_1(0) + \int_0^t [\lambda(u) - \mu_1(x_1(u) \wedge (N_1 - b(u)))]\ \mathrm{d}u - l(t) \leq N_1 + H, \\ x_1(t) + x_2(t) = x_1(t) + x_2(0) + \int_0^t [p\mu_1(x_1(u) \wedge (N_1 - b(u))) - \mu_2(N_2 \wedge x_2(u))]\ \mathrm{d}u \\ \qquad \leq N_1 + N_2 + H, \\ \mathrm{d}l(t) \geq 0, l(0) = 0, \\ \int_0^\infty 1_{\{x_1(t) + (x_2(t) - N_2)^+ < N_1 + H\}}\ \mathrm{d}l(t) = 0. \end{cases} \tag{11}$$

The function x will be referred to as the *fluid limit* associated with the queueing family X^η, where $X^\eta = (X_1^\eta,\ X_2^\eta) = (\eta N_1 + \eta H - R_1^\eta,\ R_1^\eta - R_2^\eta + \eta N_2)$.

The following proposition provides a solution to (11); see Appendix B for details. As opposed to (11), this solution (12) is given by a set of differential equations with discontinuous RHS but without reflection. Thus, implementing (12) numerically is straightforward via recursion, which would not be the case with (11).

Proposition 1 *The fluid limit approximation for X in (2) is given by*

$$\begin{aligned}
x_1(t) &= x_1(0) - \mu_1 \int_0^t [x_1(u) \wedge (N_1 - b(u))]\, \mathrm{d}u \\
&\quad + \int_0^t \left[1_{\{x_1(u) < N_1+H\}} \cdot 1_{\{x_1(u)+x_2(u) < N_1+N_2+H\}} \cdot \lambda(u)\right] \mathrm{d}u \\
&\quad + \int_0^t \left[1_{\{x_1(u) = N_1+H\}} \cdot 1_{\{x_1(u)+x_2(u) < N_1+N_2+H\}} \cdot \left[\lambda(u) \wedge l_1^*(u)\right]\right] \mathrm{d}u \\
&\quad + \int_0^t \left[1_{\{x_1(u) < N_1+H\}} \cdot 1_{\{x_1(u)+x_2(u) = N_1+N_2+H\}} \cdot \left[\lambda(u) \wedge l_2^*(u)\right]\right] \mathrm{d}u \\
&\quad + \int_0^t \left[1_{\{x_1(u) = N_1+H\}} \cdot 1_{\{x_1(u)+x_2(u) = N_1+N_2+H\}} \cdot \left[\lambda(u) \wedge l_1^*(u) \wedge l_2^*(u)\right]\right] \mathrm{d}u, \\
x_2(t) &= x_2(0) + \int_0^t [p\mu_1(x_1(u) \wedge (N_1 - b(u))) - \mu_2(x_2(u) \wedge N_2)]\, \mathrm{d}u, \qquad (12)
\end{aligned}$$

where

$$\begin{aligned}
l_1^*(u) &= \mu_1 N_1, \\
l_2^*(u) &= \mu_2 N_2 + (1-p)\mu_1 (x_1(u) \wedge (N_1 - b(u))), \\
b(u) &= (x_2(u) - N_2)^+.
\end{aligned}$$

We now introduce the functions q_1 and q_2 that denote the number of customers at Station 1 (including the waiting room) and the number of customers in service at Station 2, respectively:

$$\begin{aligned}
q_1(t) &= x_1(t) + b(t); \\
q_2(t) &= x_2 \wedge N_2.
\end{aligned}$$

Remark 1 Our model can be used to analyze the $G_t/M/N/(N+H)$ queueing system. By assuming $N_2 = \infty$ and $b = 0$, the network can be reduced to a single station ($N_1 = N$ and $\mu_1 = \mu$). In that case, the fluid limit q for the number of customers in the system is given by

$$q(t) = q(0) + \int_0^t \left[\lambda(u) - (\lambda(u) - \mu N)^+ \cdot 1_{\{q(u)=N+H\}} - \mu(q(u) \wedge N)\right] \mathrm{d}u.$$

Remark 2 Abandonments from the waiting room can occur when customers have finite patience. This is a prevalent phenomenon in service systems and healthcare in particular (for example, customers that abandon the Emergency Department are categorized as Left Without Being Seen (LWBS) [3,8]). Such abandonments can be added to our model by following [49,60]. In particular, let θ denote the individual abandonment rate from the waiting room. Thus, the term $\theta \int_0^t [x_1(u) + b(u) - N_1]^+ \mathrm{d}u$ should be subtracted from the right-hand side of $x_1(t)$ in (12); here, $[x_1(t) + b(t) - N_1]^+$ represents the number of waiting customers at Station 1 at time t.

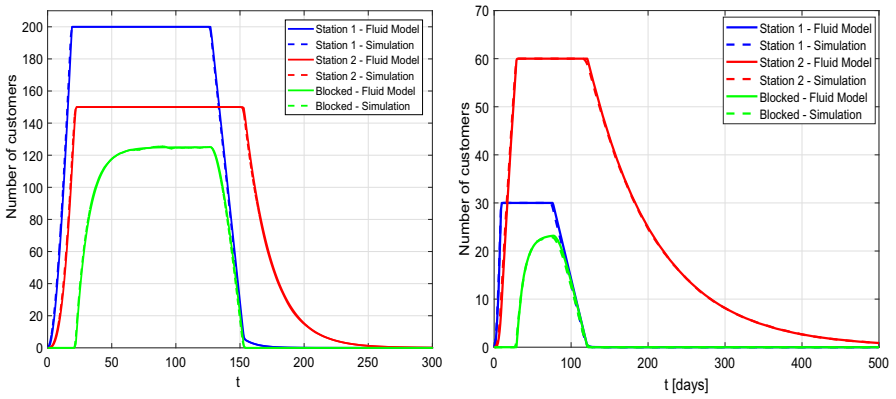

Fig. 3 Total number in each station–fluid formulation versus simulation for two scenarios. The fluid model curves overlap the simulation curves

2.3 Numerical examples

To demonstrate that our proposed fluid model accurately describes the flow of customers, we compared it to a discrete stochastic simulation model. In that model, service durations were randomly generated from exponential distributions. Customers arrive according to a non-homogeneous Poisson process that was used to represent a process with a general, time-dependent arrival rate. We note that simulating a general time-varying arrival process (G_t) is not trivial [32,47]. In [44], the authors introduce an algorithm that is based on the standard equilibrium renewal process (SERP). This algorithm is implemented in [61] to approximate the general inter-arrival times for the phase-type distribution.

The fluid equations in (12) were solved recursively, by discretizing time. Figure 3 shows the comparison between the proposed fluid model and the average simulation results for two scenarios. In the first (left plot), $N_1 = 200$, $N_2 = 150$, $H = 50$, $\mu_1 = 1/10$, $\mu_2 = 1/20$, $p = 1$, $q_1(0) = q_2(0) = 0$ and $\lambda(t) = 2t$, $0 \le t \le 120$. In the second (right plot), $N_1 = 30$, $N_2 = 60$, $H = 10$, $\mu_1 = 1/10$, $\mu_2 = 1/90$, $p = 1$, $q_1(0) = q_2(0) = 0$, and $\lambda(t) = t$, $0 \le t \le 60$.

We calculated the simulation standard deviations, averaged over time and over 500 replications. For the first scenario, the standard deviations were 0.657 for the number of customers in Station 1 with a maximal value of 4.4, 0.558 for the number in Station 2 with a maximal value 4.2, and 0.585 for the number of blocked customers with a maximal value of 4.462. To conclude, the average difference between the simulation replications and their average is less than one customer.

3 Multiple stations in tandem with finite internal waiting rooms

We now extend our model to a network with k stations in tandem and finite internal waiting rooms, as presented in Fig. 4. The notation remains as before, only with a subscript i, $i = 1, \ldots, k$, indicating Station i. Moreover, we denote the transfer

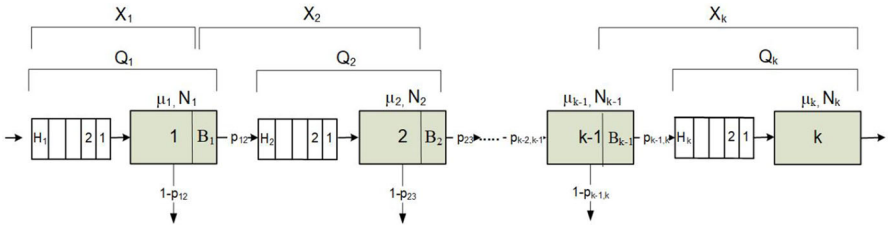

Fig. 4 Multiple stations in tandem with finite internal waiting rooms

probability from Station i to Station $i+1$ as $p_{i,i+1}$. Before each station i, there is Waiting Room i of size H_i. The parameter H_i can vary from 0 to ∞, inclusive. A customer that is referred to Station $i, i > 1$, when it is saturated waits in Waiting Room i. If the latter is full, then the customer is blocked in Station $i-1$ while occupying a server there, until space becomes available in Waiting Room i.

The stochastic model is created from the following stochastic building blocks, which are assumed to be independent: the external arrival process $A = \{A(t), t \geq 0\}$, as was defined in (2), processes $D_i = \{D_i(t), t \geq 0\}$, $i = 1, \ldots, 2k-1$, where $D_i(t)$ are standard Poisson processes, and $X_i(0)$, $i = 1, \ldots, k$, the initial number of customers in each state.

As before, the above building blocks will yield a k-dimensional stochastic process, which captures the state of our system. The stochastic process $X_1 = \{X_1(t), t \geq 0\}$ denotes the number of arrivals to Station 1 that have not completed their service at Station 1 at time t, and the stochastic process $X_i = \{X_i(t), t \geq 0\}$, $i = 2, \ldots, k$, denotes the number of customers that have completed service at Station $i-1$, but not at Station i at time t. The stochastic process $B_i = \{B_i(t), t \geq 0\}$, $i = 1, \ldots, k-1$, denotes the number of blocked customers at Station i waiting for an available server in Station $i+1$.

Let $Q = \{Q_1(t), Q_2(t), \ldots, Q_k(t), t \geq 0\}$ denote the stochastic queueing process in which $Q_i(t)$ represents the number of customers at Station i (including the waiting customers) at time t. The process Q is characterized by the following equations:

$$\begin{aligned} Q_1(t) &= X_1(t) + B_1(t); \\ Q_i(t) &= \big[X_i(t) + B_i(t)\big] \wedge (N_i + H_i), \quad i = 2, \ldots, k-1; \\ Q_k(t) &= X_k(t) \wedge (N_k + H_k), \quad t \geq 0. \end{aligned} \tag{13}$$

Here,

$$\begin{aligned} X_1(t) = X_1(0) + A(t) &- D_1\left(p_{12} \cdot \mu_1 \int_0^t [X_1(u) \wedge (N_1 - B_1(u))]\,\mathrm{d}u\right) \\ &- D_{k+1}\left((1-p_{12}) \cdot \mu_1 \int_0^t \big[X_1(u) \wedge \big(N_1 - B_1(u)\big)\big]\,\mathrm{d}u\right) \\ &- \int_0^t 1_{\{X_1(u-)+B_1(u-)=N_1+H_1\}}\,\mathrm{d}A(u), \end{aligned}$$

$$
\begin{aligned}
X_i(t) &= X_i(0) + D_{i-1}\left(p_{i-1,i}\cdot\mu_{i-1}\int_0^t \left[X_{i-1}(u)\wedge\left(N_{i-1}-B_{i-1}(u)\right)\right]\mathrm{d}u\right)\\
&\quad - D_i\left(p_{i,i+1}\cdot\mu_i\int_0^t \left[X_i(u)\wedge(N_i-B_i(u))\right]\mathrm{d}u\right)\\
&\quad - D_{k+i}\left((1-p_{i,i+1})\cdot\mu_i\int_0^t \left[X_i(u)\wedge(N_i-B_i(u))\right]\mathrm{d}u\right), i=2,\ldots,k-1,\\
X_k(t) &= X_k(0) + D_{k-1}\left(p_{k-1,k}\cdot\mu_{k-1}\int_0^t \left[X_{k-1}(u)\wedge\left(N_{k-1}-B_{k-1}(u)\right)\right]\mathrm{d}u\right)\\
&\quad - D_k\left(\mu_k\int_0^t \left[X_k(u)\wedge N_k\right]\mathrm{d}u\right),\\
B_i(t) &= \left[X_{i+1}(t)+B_{i+1}(t)-N_{i+1}-H_{i+1}\right]^+, i=1,\ldots,k-2,\\
B_{k-1}(t) &= \left[X_k(t)-N_k-H_k\right]^+.
\end{aligned}
\tag{14}
$$

Note that although $B_i(t)$, $i=1,\ldots,k-1$, is defined recursively by $B_{i+1}(t)$, it can be written explicitly for every i. For example, when $k=3$, we get that $B_1(t) = [X_2(t)+[X_3(t)-N_3-H_3]^+-N_2-H_2]^+$. An inductive construction over time shows that (14) uniquely determines the processes X and B.

By using similar methods as for the two-station network in Sect. 2, with more cumbersome algebra and notation, we establish that x, the fluid limit for the stochastic queueing family X^η, is given, for $t\geq 0$, by

$$
\begin{aligned}
x_1(t) &= x_1(0) - \mu_1\int_0^t \left[x_1(u)\wedge(N_1-b_1(u))\right]\mathrm{d}u\\
&\quad + \sum_{m=0}^{k}\sum_{\substack{A\subset\{1,\ldots,k\}:\\ |A|=m}}\int_0^t\left[\prod_{j\in A}1_{\left\{\sum_{i=1}^j x_i(u)=\sum_{i=1}^j(N_i+H_i)\right\}}\right.\\
&\quad \left.\times\prod_{j\in\{1,\ldots,k\}\cap\bar{A}}1_{\left\{\sum_{i=1}^j x_i(u)<\sum_{i=1}^j(N_i+H_i)\right\}}\left[\lambda(u)\wedge\bigwedge_{y\in A}l_y^*(u)\right]\right]\mathrm{d}u,\\
x_i(t) &= x_i(0) + \int_0^t\left[p_{i-1,i}\cdot\mu_{i-1}\left(x_{i-1}(u)\wedge(N_{i-1}-b_{i-1}(u))\right)\right.\\
&\quad \left.-\mu_i\left(x_i(u)\wedge(N_i-b_i(u))\right)\right]\mathrm{d}u,\quad i=2,\ldots,k-1,\\
x_k(t) &= x_k(0) + \int_0^t\left[p_{k-1,k}\cdot\mu_{k-1}\left(x_{k-1}(u)\wedge(N_{k-1}-b_{k-1}(u))\right)\right.\\
&\quad \left.-\mu_k\left(x_k(u)\wedge N_k\right)\right]\mathrm{d}u,
\end{aligned}
\tag{15}
$$

where

$$
\begin{aligned}
l_1^*(u) &= \mu_1 N_1,\\
l_n^*(u) &= \mu_n N_n + \sum_{j=1}^{n-1}(1-p_{j,j+1})\mu_j\left(x_j(u)\wedge\left(N_j-b_j(u)\right)\right), n=2,\ldots,k,
\end{aligned}
$$

$$b_i(t) = \left[x_{i+1}(t) + b_{i+1}(t) - N_{i+1} - H_{i+1}\right]^+, i = 1, \ldots, k-2,$$
$$b_{k-1}(t) = \left[x_k(t) - N_k - H_k\right]^+.$$

The term in the second line of (15) is a generalization of the last four terms in the expression for $x_1(t)$ in (12), when $k = 2$.

For each summand and j, if $\sum_{i=1}^{j} x_i(u) = \sum_{i=1}^{j} N_i + H_i$, the corresponding $l_j(u)$ will appear in the product. The term $l_j(u)$ represents the departure rate from Station j when the waiting room and Stations $1, \ldots, j$ are full (i.e., $\sum_{i=1}^{j} x_i(u) = \sum_{i=1}^{j}(N_i + H_i)$). The two first summations account for all combinations of $l_j(u)$, $j \in \{1, \ldots, k\}$.

We now introduce the functions $q_i(t)$, $i = 1, \ldots, k$, which denote the number of customers at Station i at time t and are given by

$$\begin{aligned} q_1(t) &= x_1(t) + b_1(t); \\ q_i(t) &= [x_i(t) + b_i(t)] \wedge (N_i + H_i), \quad i = 2, \ldots, k-1; \\ q_k(t) &= x_k(t) \wedge (N_k + H_k). \end{aligned}$$

Remark 3 A special case for the model analyzed in Sect. 3 is a model with an infinite sized waiting room before Station 1 ($H = \infty$). In this case, since customers are not lost and no reflection occurs, both the stochastic model and the fluid limit are simplified. This special case is in fact an extension of the two-station model developed in [81].

4 Numerical experiments and operational insights

In this section, we demonstrate how our models yield operational insights into time-varying tandem networks with finite capacities. To this end, we implement our models by conducting numerical experiments and parametric performance analysis. Specifically, we analyze the effects of line length, bottleneck location, and size of the waiting room on network output rate, number of customers in process, as well as sojourn, waiting, and blocking times. The phenomena presented were validated by discrete stochastic simulations.

In Sects. 4.1, 4.2, we focus on and compare two types of networks. The first has no waiting room before Station 1 ($H = 0$), and in the second there is an infinite sized waiting room before Station 1 ($H = \infty$). Sects. 4.3, 4.4 are dedicated to buffer-size effects (H varies).

The model we provide here is a tool for analyzing tandem networks with blocking. Some observations we present are intuitive and can easily be explained; others, less trivial and possibly challenging, are left for future research.

4.1 Line length

We now analyze the line length effect on network performance. We start with the case where all stations are statistically identical and their primitives independent (i.i.d.

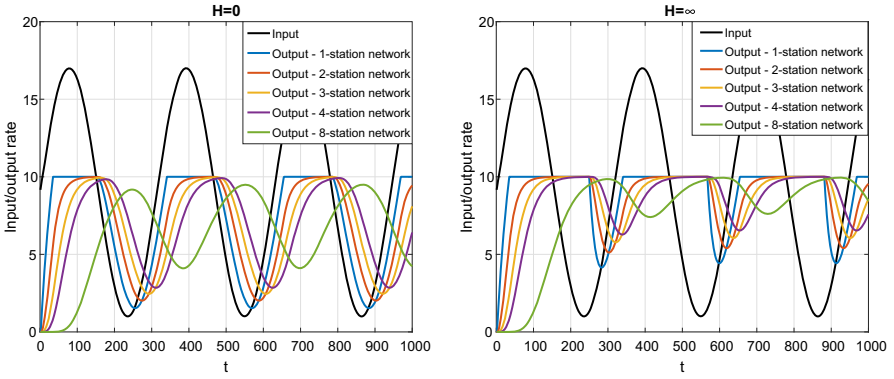

Fig. 5 Line length effect on the network output rate with k i.i.d. stations, the sinusoidal arrival rate function in (16) with $\bar{\lambda} = 9$, $\beta = 8$ and $\gamma = 0.02$, $N_i = 200$, $\mu_i = 1/20$ and $q_i(0) = 0$, $\forall i \in \{1, \ldots, k\}$. Five networks of different length are considered

stations). This implies that the stations are identical in the fluid model; in Sect. 4.2, we relax this assumption.

The arrival rate function in the following examples is the sinusoidal function

$$\lambda(t) = \bar{\lambda} + \beta \sin(\gamma t), \quad t \geq 0, \tag{16}$$

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

Figure 5 presents the time-varying input and output rates from the network, as the number of stations increases from one to eight. In both types of networks ($H = 0$ and $H = \infty$), the variation in the output rate diminishes and the average output rate (over time) decreases as the line becomes longer. When $H = 0$, due to customer loss and blocking, the variation is larger and the average output rate is smaller.

Figure 6 shows the time-varying number of customers in each station in a network with eight stations in tandem. When $H = 0$ (left plot), due to customer loss, the average number of customers is smaller, while the variation is larger, compared to the case when $H = \infty$. In fact, only about 70% of arriving customers were served when $H = 0$, compared to the obvious 100% when $H = \infty$.

Observe that the same phenomenon of the variation and average output rate decreasing as the line becomes longer (Fig. 5) also occurs when stations have ample capacities to eliminate blocking and customer loss. In these cases, system performance reaches its upper bound. Here, the output from one station is the input for the next one. In [20], an analytic expression was developed for the number of customers in the $M_t/G/\infty$ queue with a sinusoidal arrival rate as in (16). In particular, the output rate from Station 1 is given by

$$\delta_1(t) = \bar{\lambda} + \beta \left(\frac{\mu^2}{\mu^2 + \gamma^2} \sin(\gamma t) - \frac{\gamma \mu}{\mu^2 + \gamma^2} \cos(\gamma t) \right), \quad t \geq 0. \tag{17}$$

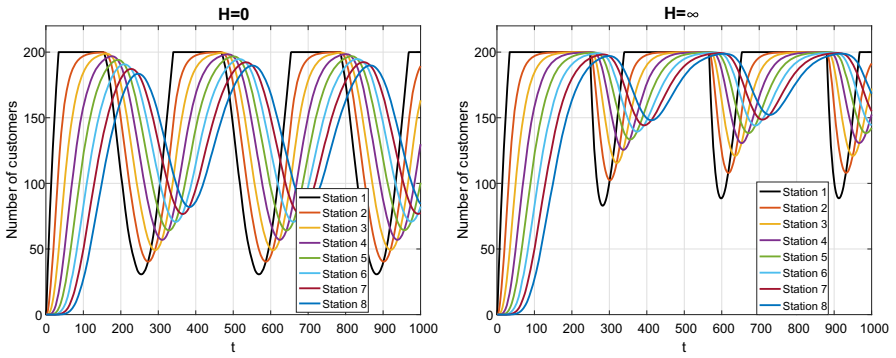

Fig. 6 Total number of customers in each station in a network with eight i.i.d. stations and the sinusoidal arrival rate function in (16) with $\bar{\lambda} = 9$, $\beta = 8$ and $\gamma = 0.02$, $N_i = 200$, $\mu_i = 1/20$, and $q_i(0) = 0$, $i = 1, \ldots, 8$

We now extend this analysis to tandem networks with ample capacity and hence no blocking (tandem networks with an infinite number of servers). Specifically, we consider (17) as the input rate for the second station and calculate the output rate from it, and so on for the rest of the stations. Consequently, the output rate from a network with i, $i = 1, 2, \ldots$, i.i.d. stations in tandem and exponential service times is given by the following expression:

$$\delta_i(t) = \bar{\lambda} + \beta \left(C_1^{(i)} \sin(\gamma t) - C_2^{(i)} \cos(\gamma t) \right), \quad t \geq 0, \tag{18}$$

where

$$\begin{aligned} &C_1^{(1)} = A_1, \quad C_2^{(1)} = B_1, \\ &A_i = \frac{\mu_i^2}{\mu_i^2 + \gamma^2}, B_i = \frac{\gamma \mu_i}{\mu_i^2 + \gamma^2}, \quad i = 1, \ldots, k, \\ &C_1^{(i)} = C_1^{(i-1)} A_i - C_2^{(i-1)} B_i, C_2^{(i)} = C_1^{(i-1)} B_i + C_2^{(i-1)} A_i, \quad i = 2, \ldots, k. \end{aligned} \tag{19}$$

Figure 7 demonstrates that, in the special case of no blocking and sinusoidal arrival rate, our results are consistent with those derived in [20]. Using (18) and (19), one can verify that the amplitude of the output rate decreases as the line becomes longer.

When capacity is lacking, blocking and customer loss prevail. Analytical expressions such as (18) do not exist for stochastic models with blocking, which renders our fluid model essential for analyzing system dynamics.

4.2 Bottleneck location

In networks where stations are not identical, the location of the bottleneck in the line has a significant effect on network performance. In our experiments, we analyzed two types of networks ($H = 0$ and $H = \infty$), each with eight stations in tandem. In

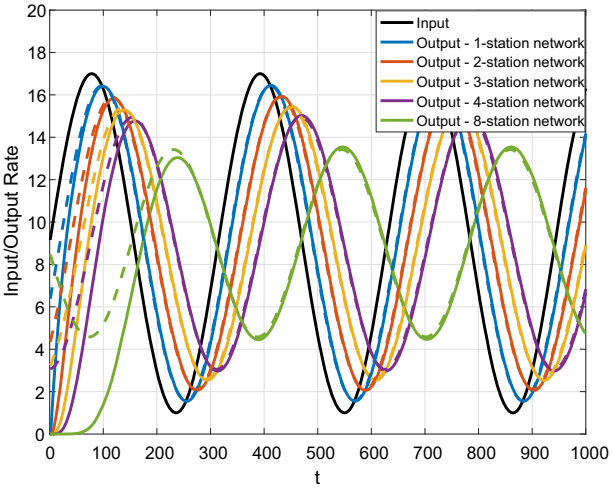

Fig. 7 Input and output rates from networks with k i.i.d. stations—fluid model (solid lines) versus values from (18) (dashed lines). The sinusoidal arrival rate function in (16) with $\bar{\lambda} = 9$, $\beta = 8$ and $\gamma = 0.02$, $N = 200$, $\mu = 1/20$, and $q_i(0) = 0$, $\forall i \in \{1, \ldots, k\}$. Five networks of different length are considered. Once the system reaches steady state, the curves from the fluid model and the analytic formula overlap

each experiment, a different station is the bottleneck; thus, it has the least processing capacity $0.3\,\mu$N, while the other stations are i.i.d. with processing capacity μN. Figure 8 presents the total number of customers in each station when the bottleneck is located first or last. In both types of networks, the bottleneck location affects the entire network.

Figure 9 presents the total number of blocked customers in each station when the last station is the bottleneck. When $H = \infty$, blocking begins at Station 7 and surges backward to the other stations. Then, the blocking is released in reversed order: First in Station 1 and then in the other stations until Station 7 is freed up. In contrast, when $H = 0$, blocking occurs only at Station 8. The blocking does not affect the other stations since Station 7 is not saturated, due to customer loss.

4.3 Waiting room size

We now examine the effect of waiting room size before the first station. Figure 10 presents this effect on a network with four i.i.d. stations in tandem, as the size of the waiting room before the first station increases from zero to infinity. The left plot in Fig. 10 presents the total number of customers in the network, and the right plot presents the network output rate. The effect of the waiting room size on these two performances is similar. As the waiting room becomes larger, fewer customers are lost, and therefore the total number of customers in the network and the output rate increase.

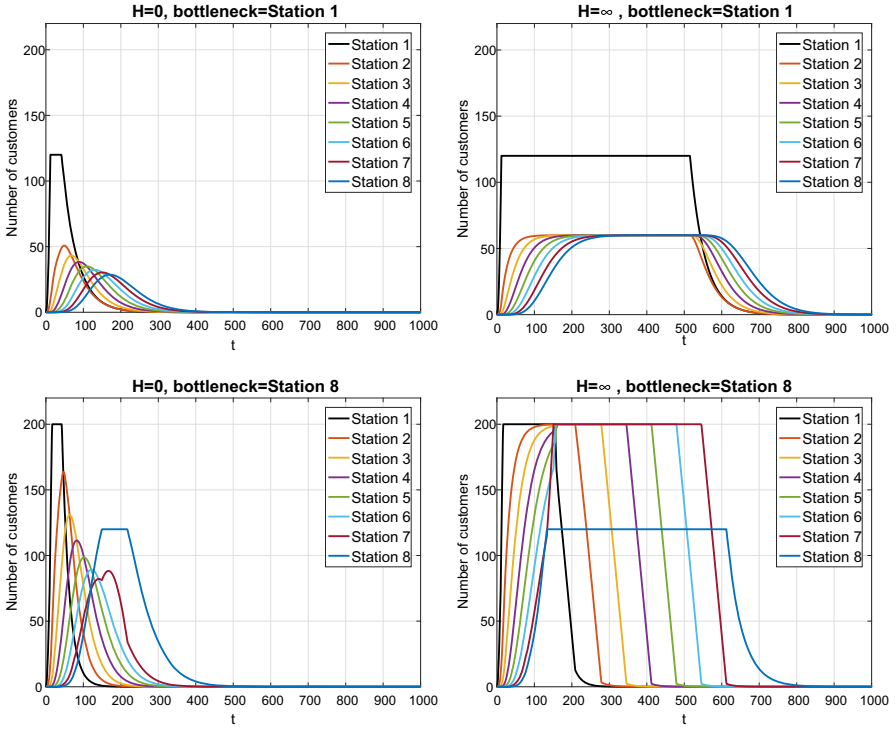

Fig. 8 The bottleneck location effect on the total number of customers in each station. For the bottleneck station, j, $N_j = 120$, $\mu_j = 1/40$. For the other stations, $i = 1, \ldots, 8$, $i \neq j$ $N_i = 200$, $\mu_i = 1/20$, $q_m(0) = 0$, $m = 1, 2, \ldots, 8$, and $\lambda(t) = 2t$, $0 \leq t \leq 40$

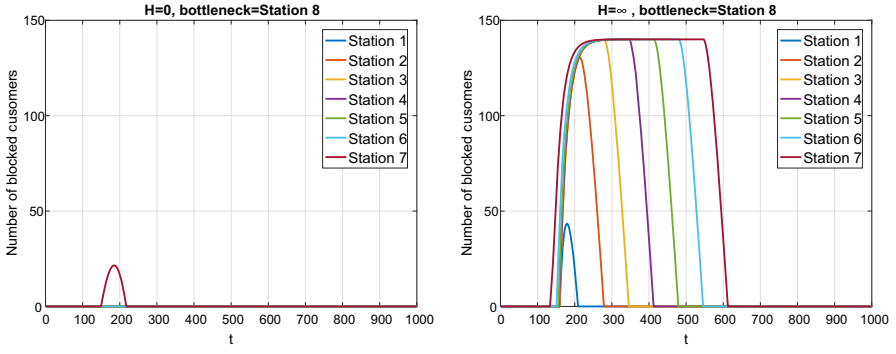

Fig. 9 Number of blocked customers in each station when the last station (Station 8) is the bottleneck. $N_i = 200$, $\mu_i = 1/20$, $i = 1, \ldots, 7$, $N_8 = 120$, $\mu_8 = 1/40$. $q_m(0) = 0$, $m = 1, \ldots, 8$, and $\lambda(t) = 2t$, $0 \leq t \leq 40$. On the left, the curves for Stations 1–6 are zero and overlap

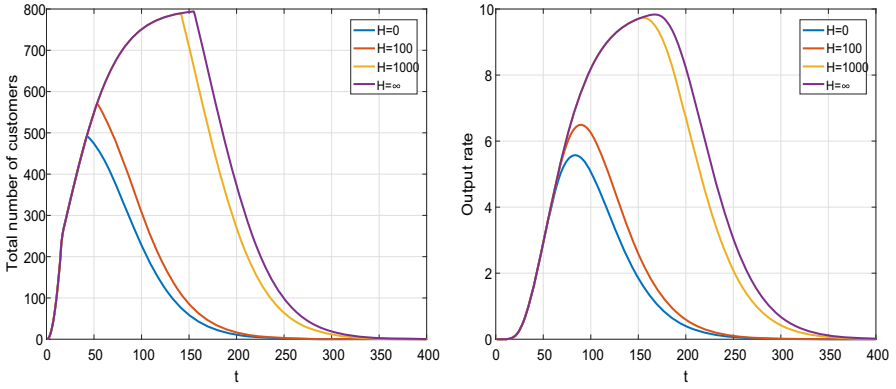

Fig. 10 Waiting room size effect on the total number of customers (left plot) and on the output rate (right plot) in a network with four i.i.d. stations, where $N_i = 200$, $\mu_i = 1/20$, $q_i(0) = 0$, $i = 1, 2, 3, 4$, and $\lambda(t) = 2t$, $0 \le t \le 40$

4.4 Sojourn time in the system

It is of interest to analyze system sojourn time and the factors that affect it. We begin by analyzing a network with two stations in tandem. Figure 11 presents the effect of the waiting room size and the bottleneck location on average sojourn time and customer loss. When there is enough waiting room to eliminate customer loss, the minimal sojourn time is achieved when the bottleneck is located at Station 2. This adds to [5] and [7], who found that the order of stations does not affect the sojourn time when service durations are deterministic and the number of servers in each station is equal. When the waiting room is not large enough to prevent customer loss, there exists a trade-off between average sojourn time and customer loss. The average sojourn time is shorter when the bottleneck is located first; however, customer loss, in this case, is greater. Explaining in detail this phenomenon requires further research.

We conclude with some observations on networks with k stations in tandem. Figure 12 presents the average sojourn time for different bottleneck locations and waiting room sizes. When the waiting room size is unlimited, the shortest sojourn time is achieved when the bottleneck is located at the end of the line. Conversely, when the waiting room is finite, the shortest sojourn time is achieved when the bottleneck is in the first station. Moreover, when the waiting room is finite, the sojourn time, as a function of the bottleneck location, increases up to a certain point and then begins to decrease. This is another way of looking at the *bowl-shaped* phenomenon [18,33] of production line capacity. In the recent example, the maximal sojourn time is achieved when the bottleneck is located at Station 6; however, other examples show that it can happen at other stations as well. To better understand this, one must analyze the components of the sojourn time—namely, the waiting time before Station 1, the blocking time at Stations $1, \ldots, 7$, and the service time at Stations $1, \ldots, 8$. Since the total service time was the same in all the networks, we examined the pattern of the sojourn time is governed by the sum of the blocking and waiting times. Figure 13 presents each of these two components. The average waiting time (right plot) decreases as

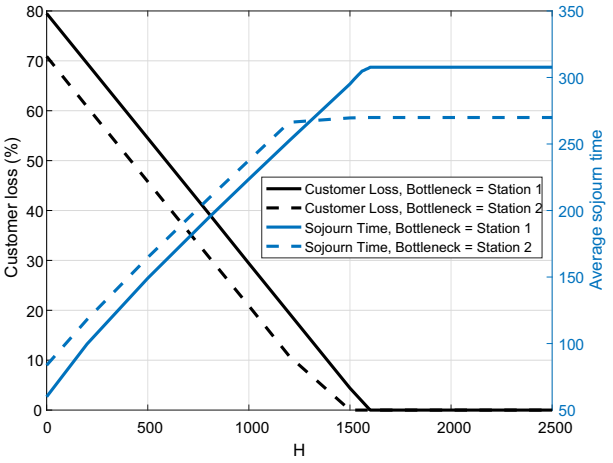

Fig. 11 The effects of waiting room size and bottleneck location on sojourn time and customer loss in a tandem network with two stations, where $q_m(0) = 0$, $m = 1, 2$, and $\lambda(t) = 20$, $0 \le t \le 100$. In the bottleneck station, j, $N_j = 120$, and $\mu_j = 1/40$; in the other station, i, $N_i = 200$, and $\mu_i = 1/20$

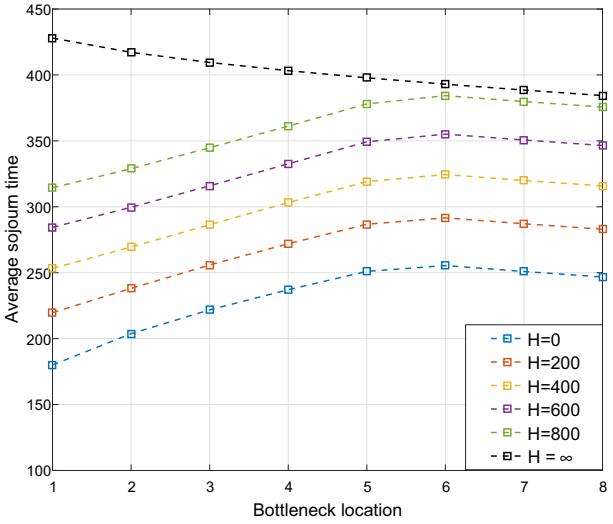

Fig. 12 The effects of waiting room size and bottleneck location on the average sojourn time in a tandem network with eight stations. Here, $q_m(0) = 0$, $m = 1, \ldots, 8$, and $\lambda(t) = 20$, $0 \le t \le 100$. In the bottleneck station, j, $N_j = 120$, and $\mu_j = 1/40$; in all other stations, $i = 1, 2, \ldots, 8$, $i \ne j$, $N_i = 200$, and $\mu_i = 1/20$

the bottleneck is located farther down the line. However, the blocking time (left plot) increases up to a certain point and then starts to decrease. To better understand the non-intuitive pattern of the average blocking time, one must analyze the components of the blocking time. In this case, it is the sum of the blocking time in Stations $1, \ldots, 7$. Figure 14 presents the blocking time in each station and overall when $H = 0$. The

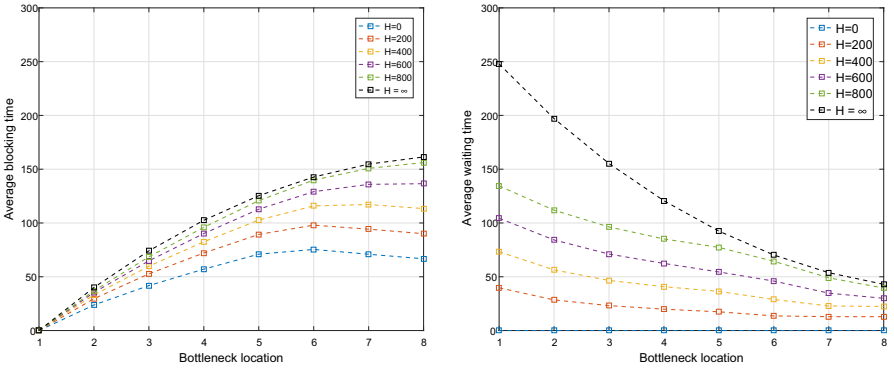

Fig. 13 The effects of waiting room size and bottleneck location on the average blocking time (left plot) and the average waiting time (right plot). The summation of the waiting time, blocking time, and service time yields the sojourn times presented in Fig. 12

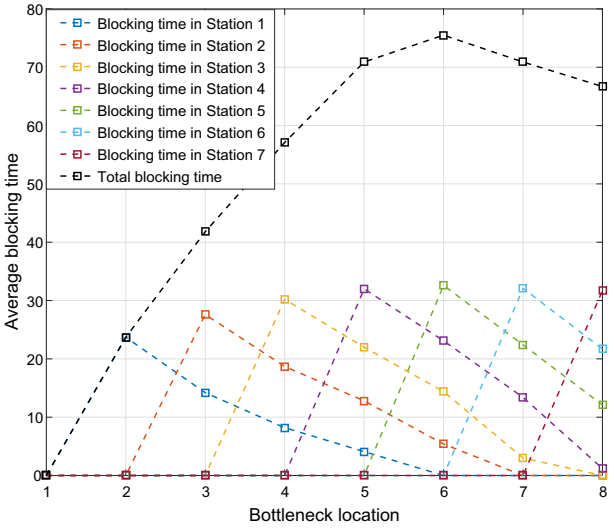

Fig. 14 Average blocking time in each station and overall when $H = 0$

blocking time in Station i, $i = 1, \ldots, 7$, equals zero when Station i is the bottleneck, since its exit is not blocked. Further, it reaches its maximum when Station $i + 1$ is the bottleneck. The sum of the average blocking time in each station yields the total blocking time and its increasing–decreasing pattern.

Acknowledgements The authors thank Junfei Huang for valuable discussions. 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 Grant CMMI-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 Theorem 1

Let T be an arbitrary positive constant. Using the Lipschitz property (Appendix C) and subtracting the equation for r in (10) from the equation for r^η in (9) yields that

$$
\begin{aligned}
&\left\|r_1^\eta - r_1\right\|_T \vee \left\|r_2^\eta - r_2\right\|_T \le G\Bigg[\left|r_1^\eta(0) - r_1(0)\right| + \left\|\int_0^\cdot \lambda(u)\,\mathrm{d}u - \eta^{-1}A^\eta(\cdot)\right\|_T \\
&+ \left\|\eta^{-1}D_1\left(\eta p\mu_1 \int_0^\cdot \left[\left(N_1 + H - r_1^\eta(u)\right) \wedge \left(N_1 - b^\eta(u)\right)\right]\mathrm{d}u\right)\right. \\
&\left. - p\mu_1 \int_0^\cdot \left[\left(N_1 + H - r_1^\eta(u)\right) \wedge \left(N_1 - b^\eta(u)\right)\right]\,\mathrm{d}u\right\|_T \\
&+ \left\|\eta^{-1}D_3\left(\eta(1-p)\mu_1 \int_0^\cdot \left[\left(N_1 + H - r_1^\eta(u)\right) \wedge \left(N_1 - b^\eta(u)\right)\right]\mathrm{d}u\right)\right. \\
&\left. - (1-p)\mu_1 \int_0^\cdot \left[\left(N_1 + H - r_1^\eta(u)\right) \wedge \left(N_1 - b^\eta(u)\right)\right]\,\mathrm{d}u\right\|_T \\
&+ \Bigg\|\mu_1 \int_0^\cdot \left[\left(N_1 + H - r_1^\eta(u)\right) \wedge \left(N_1 - b^\eta(u)\right)\right. \\
&\left. - \left(N_1 + H - r_1(u)\right) \wedge \left(N_1 - b(u)\right)\right]\,\mathrm{d}u\Bigg\|_T\Bigg] \vee \\
&G\Bigg[\left|r_2^\eta(0) - r_2(0)\right| + \left\|\int_0^\cdot \lambda(u)\,\mathrm{d}u - \eta^{-1}A^\eta(\cdot)\right\|_T \\
&+ \left\|\eta^{-1}D_3\left(\eta(1-p)\mu_1 \int_0^\cdot \left[\left(N_1 + H - r_1^\eta(u)\right) \wedge \left(N_1 - b^\eta(u)\right)\right]\mathrm{d}u\right)\right. \\
&\left. - (1-p)\mu_1 \int_0^\cdot \left[\left(N_1 + H - r_1^\eta(u)\right) \wedge \left(N_1 - b^\eta(u)\right)\right]\,\mathrm{d}u\right\|_T \\
&+ \left\|\eta^{-1}D_2\left(\eta\mu_2 \int_0^\cdot \left[N_2 \wedge \left(r_1^\eta(u) - r_2^\eta(u) + N_2\right)\right]\mathrm{d}u\right)\right. \\
&\left. - \mu_2 \int_0^\cdot \left[N_2 \wedge \left(r_1^\eta(u) - r_2^\eta(u) + N_2\right)\right]\,\mathrm{d}u\right\|_T \\
&+ \Bigg\|(1-p)\mu_1 \int_0^\cdot \left[\left(N_1 + H - r_1^\eta(u)\right) \wedge \left(N_1 - b^\eta(u)\right)\right. \\
&\left. - \left(N_1 + H - r_1(u)\right) \wedge \left(N_1 - b(u)\right)\right]\,\mathrm{d}u\Bigg\|_T \\
&+ \Bigg\|\mu_2 \int_0^\cdot \left[\left(N_2 \wedge \left(r_1^\eta(u) - r_2^\eta(u) + N_2\right)\right)\right. \\
&\left. - \left(N_2 \wedge \left(r_1(u) - r_2(u) + N_2\right)\right)\right]\,\mathrm{d}u\Bigg\|_T\Bigg], \qquad (20)
\end{aligned}
$$

where G is the Lipschitz constant.

The first, second, sixth, and seventh terms on the right-hand side converge to zero by the conditions of the theorem. For proving convergence to zero of the third, fourth, eighth, and ninth terms, we use Lemma 1 in Appendix D. By the FSLLN for Poisson processes,

$$\sup_{0\le u\le t}\left|\eta^{-1}D(\eta u)-u\right|\to 0,\quad \forall t\ge 0\quad a.s.$$

Note that the functions $p\mu_1\int_0^t\left[\left(N_1+H-r_1^\eta(u)\right)\wedge(N_1-b^\eta(u))\right]\mathrm{d}u$ and $\mu_2\int_0^t\left[N_2\wedge\left(r_1^\eta(u)-r_2^\eta(u)+N_2\right)\right]\mathrm{d}u$ are bounded by $p\mu_1\cdot(N_1+H)\cdot T$ and $\mu_2\cdot N_2\cdot T$, respectively, for $0\le p\le 1$ and $t\in[0,T]$. This, together with Lemma 1, implies that the third, fourth, eighth, and ninth terms in (20) converge to 0.

We get that

$$\begin{aligned}
&\left\|r_1^\eta-r_1\right\|_T\vee\left\|r_2^\eta-r_2\right\|_T\\
&\le\left[\epsilon_1^\eta(T)+G\mu_1\left\|\int_0^\cdot\left[\left(N_1+H-r_1^\eta(u)\right)\wedge\left(N_1-b^\eta(u)\right)-\left(N_1+H-r_1(u)\right)\wedge\left(N_1-b(u)\right)\right]\mathrm{d}u\right\|_T\right]\vee\\
&\left[\epsilon_2^\eta(T)+G(1-p)\mu_1\left\|\int_0^\cdot\left[\left(N_1+H-r_1^\eta(u)\right)\wedge\left(N_1-b^\eta(u)\right)-\left(N_1+H-r_1(u)\right)\wedge\left(N_1-b(u)\right)\right]\mathrm{d}u\right\|_T\right.\\
&\quad\left.+G\mu_2\left\|\int_0^\cdot\left[N_2\wedge\left(r_1^\eta(u)-r_2^\eta(u)+N_2\right)\right]-\left[N_2\wedge\left(r_1(u)-r_2(u)+N_2\right)\right]\mathrm{d}u\right\|_T\right]\\
&\le\left[\epsilon_1^\eta(T)+G\mu_1\left\|\int_0^\cdot\left[r_1^\eta(u)-r_1(u)\right]\mathrm{d}u\right\|_T+G\mu_1\left\|\int_0^\cdot\left[b^\eta(u)-b(u)\right]\mathrm{d}u\right\|_T\right]\vee\\
&\quad\left[\epsilon_2^\eta(T)+G(1-p)\mu_1\left\|\int_0^\cdot\left[r_1^\eta(u)-r_1(u)\right]\mathrm{d}u\right\|_T\right.\\
&\quad+G(1-p)\mu_1\left\|\int_0^\cdot\left[b^\eta(u)-b(u)\right]\mathrm{d}u\right\|_T\\
&\quad\left.+G\mu_2\left\|\int_0^\cdot\left[r_1^\eta(u)-r_1(u)\right]\mathrm{d}u\right\|_T+G\mu_2\left\|\int_0^\cdot\left[r_2^\eta(u)-r_2(u)\right]\mathrm{d}u\right\|_T\right]\\
&\le\left[\epsilon_1^\eta(T)+G\mu_1\int_0^T\left\|r_1^\eta-r_1\right\|_u\mathrm{d}u+G\mu_1\int_0^T\left\|b^\eta-b\right\|_u\mathrm{d}u\right]\vee\\
&\quad\left[\epsilon_2^\eta(T)+G\mu_1\int_0^T\left\|r_1^\eta-r_1\right\|_u\mathrm{d}u+G\mu_1\int_0^T\left\|b^\eta-b\right\|_u\mathrm{d}u\right.\\
&\quad\left.+G\mu_2\int_0^T\left\|r_1^\eta-r_1\right\|_u\mathrm{d}u+G\mu_2\int_0^T\left\|r_2^\eta-r_2\right\|_u\mathrm{d}u\right],
\end{aligned}\tag{21}$$

where $\epsilon_1^\eta(T)$ bounds the sum of the first four terms on the right-hand side of (20), and $\epsilon_2^\eta(T)$ bounds the sum of the sixth to ninth terms; these two quantities $\epsilon_1^\eta(T)$ and $\epsilon_2^\eta(T)$ converge to zero, as $\eta\to\infty$. The second inequality in (21) is obtained by using the inequalities $|a\wedge b-a\wedge c|\le|b-c|$ and $|a\wedge b-c\wedge d|\le|a-c|+|b-d|$ for any a, b, c, and d. The third equality in (21) is because $0\le p\le 1$.

We now use

$$\begin{aligned}\int_0^T \left\|b^\eta - b\right\|_u \,\mathrm{d}u &= \int_0^T \left\|\left(r_1^\eta - r_2^\eta\right)^+ - (r_1 - r_2)^+\right\|_u \,\mathrm{d}u \\ &= \int_0^T \left\|r_1^\eta - r_1^\eta \wedge r_2^\eta - r_1 + r_1 \wedge r_2\right\|_u \,\mathrm{d}u \\ &\le \int_0^T \left[\left\|r_1^\eta - r_1\right\|_u + \left\|r_1^\eta \wedge r_2^\eta - r_1 \wedge r_2\right\|_u\right]\mathrm{d}u \\ &\le \int_0^T \left[2\left\|r_1^\eta - r_1\right\|_u + \left\|r_2^\eta - r_2\right\|_u\right]\mathrm{d}u \\ &= 2\int_0^T \left\|r_1^\eta - r_1\right\|_u \,\mathrm{d}u + \int_0^T \left\|r_2^\eta - r_2\right\|_u \,\mathrm{d}u. \end{aligned} \tag{22}$$

From (21) and (22), we get that

$$\begin{aligned}&\left\|r_1^\eta - r_1\right\|_T \vee \left\|r_2^\eta - r_2\right\|_T \\ &\le \left[\epsilon_1^\eta(T) \vee \epsilon_2^\eta(T)\right] + G\left(3\mu_1 + \mu_2\right)\int_0^T \left\|r_1^\eta - r_1\right\|_u \,\mathrm{d}u + G\left(\mu_1 \vee \mu_2\right)\int_0^T \left\|r_2^\eta - r_2\right\|_u \,\mathrm{d}u \\ &\le \left[\epsilon_1^\eta(T) \vee \epsilon_2^\eta(T)\right] + 2G\left(3\mu_1 \vee \mu_2\right)\left[\int_0^T \left\|r_1^\eta - r_1\right\|_u \,\mathrm{d}u + \int_0^T \left\|r_2^\eta - r_2\right\|_u \,\mathrm{d}u\right] \\ &\le \left[\epsilon_1^\eta(T) \vee \epsilon_2^\eta(T)\right] + 4G\left(3\mu_1 \vee \mu_2\right)\left[\int_0^T \left\|r_1^\eta - r_1\right\|_u \,\mathrm{d}u \vee \int_0^T \left\|r_2^\eta - r_2\right\|_u \,\mathrm{d}u\right] \\ &\le \left[\epsilon_1^\eta(T) \vee \epsilon_2^\eta(T)\right] + 4G\left(3\mu_1 \vee \mu_2\right)\left[\int_0^T \left\|r_1^\eta - r_1\right\|_u \vee \left\|r_2^\eta - r_2\right\|_u \,\mathrm{d}u\right]. \end{aligned} \tag{23}$$

The first equality in (23) is obtained by using the inequality $(a + b) \vee (c + d) \le a \vee c + b \vee d$, for any a, b, c, and d. Applying Gronwall's inequality [22] to (23) completes the proof for both the existence and uniqueness of r.

Appendix B: Proof of Proposition 1

We begin by proving that the solution for (11) satisfies, for $t \ge 0$,

$$\begin{aligned} l(t) = &\int_0^t 1_{\{x_1(u) \ge N_1 + H\}} \cdot 1_{\{x_1(u) + x_2(u) < N_1 + N_2 + H\}} \left[\lambda(u) - l_1(u)\right]^+ \mathrm{d}u \\ &+ \int_0^t 1_{\{x_1(u) < N_1 + H\}} \cdot 1_{\{x_1(u) + x_2(u) \ge N_1 + N_2 + H\}} \left[\lambda(u) - l_2(u)\right]^+ \mathrm{d}u \\ &+ \int_0^t 1_{\left\{x_1(u) \ge N_1 + H\right\}} \cdot 1_{\left\{x_1(u) + x_2(u) \ge N_1 + N_2 + H\right\}} \left[\lambda(u) - l_1(u) \wedge l_2(u)\right]^+ \mathrm{d}u, \end{aligned} \tag{24}$$

where

$$l_1(u) = \mu_1\left(x_1(u) \wedge \left(N_1 - b(u)\right)\right);$$

$$l_2(u) = \mu_2 \left(x_2(u) \wedge N_2\right) + (1-p)\mu_1 \left(x_1(u) \wedge (N_1 - b(u))\right).$$

In order to prove this, we substitute (24) in (11) and show that the properties in (11) prevail. We begin by substituting (24) in the first line of (11). Using $(a-b)^+ = [a - a \wedge b]$, for any a, b, we obtain

$$\begin{aligned}
x_1(t) = x_1(0) &+ \int_0^t \left[\lambda(u) - \mu_1 \left[x_1(u) \wedge (N_1 - b(u))\right]\right] \mathrm{d}u \\
&- \int_0^t 1_{\{x_1(u) \geq N_1+H\}} \cdot 1_{\{x_1(u)+x_2(u) < N_1+N_2+H\}} \left[\lambda(u) - \lambda(u) \wedge l_1(u)\right] \mathrm{d}u \\
&- \int_0^t 1_{\{x_1(u) < N_1+H\}} \cdot 1_{\{x_1(u)+x_2(u) \geq N_1+N_2+H\}} \left[\lambda(u) - \lambda(u) \wedge l_2(u)\right] \mathrm{d}u \\
&- \int_0^t 1_{\{x_1(u) \geq N_1+H\}} \cdot 1_{\{x_1(u)+x_2(u) \geq N_1+N_2+H\}} \left[\lambda(u) - \lambda(u) \wedge l_1(u) \wedge l_2(u)\right] \mathrm{d}u,
\end{aligned}$$

and therefore,

$$\begin{aligned}
x_1(t) = x_1(0) &+ \int_0^t \Big[1_{\{x_1(u) < N_1+H\}} \cdot 1_{\{x_1(u)+x_2(u) < N_1+N_2+H\}} \cdot \lambda(u) \\
&- \mu_1 \left[x_1(u) \wedge (N_1 - b(u))\right]\Big] \mathrm{d}u \\
&+ \int_0^t \left[1_{\{x_1(u) \geq N_1+H\}} \cdot 1_{\{x_1(u)+x_2(u) < N_1+N_2+H\}} \cdot (\lambda(u) \wedge l_1(u))\right] \mathrm{d}u \\
&+ \int_0^t \left[1_{\{x_1(u) < N_1+H\}} \cdot 1_{\{x_1(u)+x_2(u) \geq N_1+N_2+H\}} \cdot (\lambda(u) \wedge l_2(u))\right] \mathrm{d}u \\
&+ \int_0^t \left[1_{\{x_1(u) \geq N_1+H\}} \cdot 1_{\{x_1(u)+x_2(u) \geq N_1+N_2+H\}} \cdot (\lambda(u) \wedge l_1(u) \wedge l_2(u))\right] \mathrm{d}u; \\
x_2(t) = x_2(0) &+ \int_0^t \left[p\mu_1 \left[x_1(u) \wedge (N_1 - b(u))\right] - \mu_2 \left(x_2(u) \wedge N_2\right)\right] \mathrm{d}u. \qquad (25)
\end{aligned}$$

Clearly, the properties in the third and fourth lines in (11) prevail. It is left to verify that the first and second conditions prevail. This is done by the following proposition.

Proposition 2 *The functions $x_1(\cdot)$ and $x_1(\cdot)+x_2(\cdot)$ as in (25) are bounded by N_1+H and N_1+N_2+H, respectively.*

Proof First, we prove that the function $x_1(\cdot)$, as in (25), is bounded by N_1+H. Assume that, for some t, $x_1(t) > N_1+H$. Since $x_1(0) \leq N_1+H$ and x_1 is continuous (being an integral), there must be a last $\tilde{t}$ in $[0,t]$, such that $x_1(\tilde{t}) = N_1+H$ and $x_1(u) > N_1+H$, for $u \in [\tilde{t}, t]$. Without loss of generality, assume that $\tilde{t} = 0$; thus $x_1(0) = N_1 + H$ and $x_1(u) > N_1 + H$ for $u \in (0, t]$. From (25), we get that

$$\begin{aligned}
x_1(t) = N_1 + H &+ \int_0^t \left[1_{\{x_1(u)+x_2(u) < N_1+N_2+H\}} \cdot (\lambda(u) \wedge l_1(u))\right] \mathrm{d}u \\
&+ \int_0^t \left[1_{\{x_1(u)+x_2(u) \geq N_1+N_2+H\}} \cdot (\lambda(u) \wedge l_1(u) \wedge l_2(u))\right] \mathrm{d}u \\
&- \mu_1 \int_0^t \left[x_1(u) \wedge (N_1 - b(u))\right] \mathrm{d}u
\end{aligned}$$

$$\leq N_1 + H + \int_0^t [l_1(u) - \mu_1 [x_1(u) \wedge (N_1 - b(u))]] \, \mathrm{d}u = N_1 + H,$$

which contradicts our assumption and proves that $x_1(\cdot)$ cannot exceed $H_1 + N_1$.

What is left to prove now is that the function $x_1(\cdot) + x_2(\cdot)$ is bounded by $N_1 + N_2$. Without loss of generality, assume that $x_1(0) + x_2(0) = N_1 + N_2 + H$ and $x_1(u) + x_2(u) > N_1 + N_2 + H$ for $u \in (0, t]$. This assumption, together with $x_1 \leq N_1 + H$, yields that $x_2 > N_2$; hence, from (25), we get that

$$\begin{aligned}
&x_1(t) + x_2(t) \\
&\quad = N_1 + N_2 + H \int_0^t \left[1_{\{x_1(u) \geq N_1 + N_1\}} \cdot (\lambda(u) \wedge l_1(u) \wedge l_2(u))\right] \mathrm{d}u \\
&\qquad + \int_0^t \left[1_{\{x_1(u) < N_1 + H\}} \cdot (\lambda(u) \wedge l_2(u))\right] \mathrm{d}u \\
&\qquad - \int_0^t [(1-p)\mu_1 (x_1(u) \wedge (N_1 - b(u))) + \mu_2 (x_2(u) \wedge N_2)] \, \mathrm{d}u \\
&\quad \leq N_1 + N_2 + H + \int_0^t [l_2(u) - (1-p)\mu_1 (x_1(u) \wedge (N_1 - b(u))) \\
&\qquad - \mu_2 (x_2(u) \wedge N_2)] \, \mathrm{d}u \\
&\quad = N_1 + N_2 + H,
\end{aligned}$$

which contradicts the assumption that $x_1(t) + x_2(t) > N_1 + N_2 + H$ and proves that $x_1(\cdot) + x_2(\cdot)$ is bounded by $N_1 + N_2 + H$. □

By the solution uniqueness (Proposition 3), we have established that x, the fluid limit for the stochastic queueing family X^η in (2), is given by (25).

The following two remarks explain why (25) is equivalent to (12):

1. After proving that $x_1(\cdot) \leq N_1 + H$ and $x_1(\cdot) + x_2(\cdot) \leq N_1 + N_2 + H$ in Proposition 2, the indicators in (24) can accommodate only the cases when $x_1(\cdot) = N_1 + H$ and $x_1(\cdot) + x_2(\cdot) = N_1 + N_2 + H$.
2. When $x_1(u) = N_1 + H$ and $x_1(u) + x_2(u) < N_1 + N_2 + H$, $x_2(u) < N_2$, and hence $b(u) = 0$ and $l_1(u) = l_1^*(u)$. Alternatively, when $x_1(u) < N_1 + H$ and $x_1(u) + x_2(u) = N_1 + N_2 + H$, $x_2(u) > N_2$, and therefore $l_2(u) = l_2^*(u)$.

Appendix C: Uniqueness and Lipschitz property

Let $C \equiv C[0, \infty]$. We now define mappings $\psi : C^2 \to C$ and $\phi : C^2 \to C^2$ for $m \in C^2$ by setting

$$\psi(m)(t) = \sup_{0 \leq s \leq t} \left(-\Big(m_1(s) \wedge m_2(s)\Big)\right)^+;$$

$$\phi(m)(t) = m(t) + \psi(m)(t) \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad t \geq 0.$$

Proposition 3 *Suppose that $m \in C^2$ and $m(0) \geq 0$. Then, $\psi(m)$ is the unique function l, such that*

1. *l is continuous and non-decreasing with $l(0) = 0$,*
2. *$r(t) = m(t) + l(t) \geq 0$ for all $t \geq 0$,*
3. *l increases only when $r_1 = 0$ or $r_2 = 0$.*

Proof Let l^* be any other solution. We set $y = r_1^* - r_1 = r_2^* - r_2 = l^* - l$. Using the Riemann–Stieltjes chain rule [31, Ch. 2.2]

$$f(y_t) = f(y_0) + \int_0^t f'(y)\,\mathrm{d}y,$$

for any continuously differentiable $f : R \to R$. Taking $f(y) = y^2/2$, we get that

$$\frac{1}{2}\big(r_i^*(t) - r_i(t)\big)^2 = \int_0^t (r_i^* - r_i)\,\mathrm{d}l^* + \int_0^t (r_i - r_i^*)\,\mathrm{d}l. \tag{26}$$

The function l^* increases when either $r_1^* = 0$ or $r_2^* = 0$. In addition, $r_1 \geq 0$ and $r_2 \geq 0$. Thus, either $(r_1^* - r_1)\,\mathrm{d}l^* \leq 0$ or $(r_2^* - r_2)\,\mathrm{d}l^* \leq 0$. Since $r_1^* - r_1 = r_2^* - r_2$, both terms are non-positive. The same principles yield that the second terms in both lines on the right-hand side of (26) are non-positive. Since the left-hand side ≥ 0, both sides must be zero; thus, $r_1^* = r_1$, $r_2^* = r_2$, and $l^* = l$. □

Proposition 4 *The mappings ψ and ϕ are Lipschitz continuous on $D_o[0, t]$ under the uniform topology for any fixed t.*

Proof We begin by proving the Lipschitz continuity of ψ. For this, we show that for any $T > 0$ there exists $C \in R$ such that

$$\big\|\psi(m) - \psi(m')\big\|_T \leq C\Big[\,\big\|m_1 - m_1'\big\|_T \vee \big\|m_2 - m_2'\big\|_T\,\Big],$$

for all $m, m' \in D_0^2$.

$$\begin{aligned}
&\big\|\psi(m) - \psi(m')\big\|_T \\
&= \left\| \sup_{0\leq s\leq \cdot} \Big(-\big(m_1(s) \wedge m_2(s)\big)\Big)^+ - \sup_{0\leq s\leq t} \Big(-\big(m_1'(s) \wedge m_2'(s)\big)\Big)^+ \right\|_T \\
&\leq \left\| \sup_{0\leq s\leq \cdot} \big|\big(m_1(s) \wedge m_2(s)\big) - \big(m_1'(s) \wedge m_2'(s)\big)\big| \right\|_T \\
&= \big\|\big(m_1 \wedge m_2\big) - \big(m_1' \wedge m_2'\big)\big\|_T \leq 2\Big[\,\big\|m_1 - m_1'\big\|_T \vee \big\|m_2 - m_2'\big\|_T\,\Big].
\end{aligned} \tag{27}$$

The last inequality derives from

$$m_1(t) \wedge m_2(t) = \big(m_1(t) - m_1'(t) + m_1'(t)\big) \wedge \big(m_2(t) - m_2'(t) + m_2'(t)\big);$$

therefore,

$$m_1(t) \wedge m_2(t) \le m_1'(t) \wedge m_2'(t) + \left\|m_1 - m_1'\right\|_T + \left\|m_2 - m_2'\right\|_T,$$
$$m_1(t) \wedge m_2(t) \ge m_1'(t) \wedge m_2'(t) - \left\|m_1 - m_1'\right\|_T - \left\|m_2 - m_2'\right\|_T,$$

and

$$\left|m_1(t) \wedge m_2(t) - m_1'(t) \wedge m_2'(t)\right| \le \left\|m_1 - m_1'\right\|_T + \left\|m_2 - m_2'\right\|_T,$$

which yields

$$\begin{aligned}\left\|m_1(t) \wedge m_2(t) - m_1'(t) \wedge m_2'(t)\right\|_T &\le \left\|m_1 - m_1'\right\|_T + \left\|m_2 - m_2'\right\|_T \\ &\le 2\left(\left\|m_1 - m_1'\right\|_T \vee \left\|m_2 - m_2'\right\|_T\right).\end{aligned}$$

Our next step is proving the Lipschitz continuity of ϕ. For this, we show that for any $T > 0$ there exists $C \in R$ such that

$$\left\|\phi_1(m) - \phi_1(m')\right\|_T \vee \left\|\phi_2(m) - \phi_2(m')\right\|_T \le C\Big[\left\|m_1 - m_1'\right\|_T \vee \left\|m_2 - m_2'\right\|_T\Big],$$

for all $m, m' \in D_0^2$.

We begin with the left-hand side:

$$\begin{aligned}&\left\|\phi_1(m) - \phi_1(m')\right\|_T \vee \left\|\phi_2(m) - \phi_2(m')\right\|_T \\ &= \left\|m_1(t) + \psi(m)(t) - m_1'(t) - \psi(m')(t)\right\|_T \vee \\ &\quad \left\|m_2(t) + \psi(m)(t) - m_2'(t) - \psi(m')(t)\right\|_T \\ &= \left\|m_1(t) - m_1'(t) + \psi(m)(t) - \psi(m')(t)\right\|_T \vee \\ &\quad \left\|m_2(t) - m_2'(t) + \psi(m)(t) - \psi(m')(t)\right\|_T \\ &\le \left\|m_1(t) - m_1'(t)\right\|_T + \left\|\psi(m)(t) - \psi(m')(t)\right\|_T \vee \\ &\quad \left\|m_2(t) - m_2'(t)\right\|_T + \left\|\psi(m)(t) - \psi(m')(t)\right\|_T \\ &\le \left\|m_1 - m_1'\right\|_T \vee \left\|m_2 - m_2'\right\|_T + \left\|\psi(m)(t) - \psi(m')(t)\right\|_T \\ &\le 3\left(\left\|m_1 - m_1'\right\|_T \vee \left\|m_2 - m_2'\right\|_T\right),\end{aligned}$$

where the last inequality is derived from (27). □

Appendix D: Lemma 1

Lemma 1 *Let the function $f_\eta(\cdot) \to 0$, u.o.c. as $\eta \to \infty$. Then, $f_\eta(g_\eta(\cdot)) \to 0$, u.o.c. as $\eta \to \infty$, for any $g_\eta(\cdot)$ that are locally bounded uniformly in η.*

Proof Choose $T > 0$, and let C_T be a constant such that $\left|g_\eta(t)\right| \le C_T$, for all $t \in [0, T]$. By the assumption on $f_\eta(\cdot)$, we have $\|f_\eta\|_{C_T} \to 0$ as $\eta \to \infty$. It follows that $\|f_\eta(g_\eta(\cdot))\|_T \to 0$ as $\eta \to \infty$, which completes the proof. □

References

1. Afèche, P., Araghi, M., Baron, O.: Customer acquisition, retention, and queueing-related service quality: optimal advertising, staffing, and priorities for a call center. Manuf. Serv. Oper. Manag. **19**(4), 674–691 (2017)
2. Akyildiz, I., von Brand, H.: Exact solutions for networks of queues with blocking-after-service. Theor. Comput. Sci. **125**(1), 111–130 (1994)
3. Arendt, K., Sadosty, A., Weaver, A., Brent, C., Boie, E.: The left-without-being-seen patients: what would keep them from leaving? Ann. Emerg. Med. **42**(3), 317–IN2 (2003)
4. Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y., Tseytlin, Y., Yom-Tov, G.: On patient flow in hospitals: a data-based queueing-science perspective. Stoch. Syst. **5**(1), 146–194 (2015)
5. Avi-Itzhak, B.: A sequence of service stations with arbitrary input and regular service times. Manag. Sci. **11**(5), 565–571 (1965)
6. Avi-Itzhak, B., Levy, H.: A sequence of servers with arbitrary input and regular service times revisited: in memory of Micha Yadin. Manag. Sci. **41**(6), 1039–1047 (1995)
7. Avi-Itzhak, B., Yadin, M.: A sequence of two servers with no intermediate queue. Manag. Sci. **11**(5), 553–564 (1965)
8. Baker, D., Stevens, C., Brook, R.: Patients who leave a public hospital emergency department without being seen by a physician: causes and consequences. JAMA **266**(8), 1085–1090 (1991)
9. Balsamo, S., de Nitto Personè, V.: A survey of product form queueing networks with blocking and their equivalences. Ann. Oper. Res. **48**(1), 31–61 (1994)
10. Balsamo, S., de Nitto Personé, V., Onvural, R.: Analysis of Queueing Networks with Blocking. Springer, Berlin (2001)
11. Borisov, I., Borovkov, A.: Asymptotic behavior of the number of free servers for systems with refusals. Theory Probab. Appl. **25**(3), 439–453 (1981)
12. Borovkov, A.: Stochastic Processes in Queueing Theory. Springer, Berlin (2012)
13. Brandwajn, A., Jow, Y.: An approximation method for tandem queues with blocking. Oper. Res. **36**(1), 73–83 (1988)
14. Bretthauer, K., Heese, H., Pun, H., Coe, E.: Blocking in healthcare operations: a new heuristic and an application. Prod. Oper. Manag. **20**(3), 375–391 (2011)
15. Buzacott, J., Shanthikumar, J.: Stochastic Models of Manufacturing Systems. Prentice Hall, Englewood Cliffs (1993)
16. Chen, H., Yao, D.: Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization. Springer, Berlin (2013)
17. Cohen, I., Mandelbaum, A., Zychlinski, N.: Minimizing mortality in a mass casualty event: fluid networks in support of modeling and staffing. IIE Trans. **46**(7), 728–741 (2014)
18. Conway, R., Maxwell, W., McClain, J., Thomas, L.: The role of work-in-process inventory in serial production lines. Oper. Res. **36**(2), 229–241 (1988)
19. Dallery, Y., Gershwin, S.: Manufacturing flow line systems: a review of models and analytical results. Queueing Syst. **12**(1–2), 3–94 (1992)
20. Eick, S., Massey, W., Whitt, W.: $M_t/G/\infty$ queues with sinusoidal arrival rates. Manag. Sci. **39**(2), 241–252 (1993)
21. El-Darzi, E., Vasilakis, C., Chaussalet, T., Millard, P.: A simulation modelling approach to evaluating length of stay, occupancy, emptiness and bed blocking in a hospital geriatric department. Health Care Manag. Sci. **1**(2), 143–149 (1998)
22. Ethier, S., Kurtz, T.: Markov Processes: Characterization and Convergence. Wiley, New York (2009)
23. Feldman, Z., Mandelbaum, A., Massey, W., Whitt, W.: Staffing of time-varying queues to achieve time-stable performance. Manag. Sci. **54**(2), 324–338 (2008)
24. Filippov, A.: Differential Equations with Discontinuous Righthand Sides: Control Systems. Springer, Berlin (2013)
25. Garnett, O., Mandelbaum, A., Reiman, M.: Designing a call center with impatient customers. Manuf. Serv. Oper. Manag. **4**(3), 208–227 (2002)
26. Gershwin, S.: An efficient decomposition method for the approximate evaluation of tandem queues with finite storage space and blocking. Oper. Res. **35**(2), 291–305 (1987)
27. Glynn, P., Whitt, W.: Departures from many queues in series. Ann. Appl. Probab. **1**(4), 546–572 (1991)
28. Grassmann, W., Drekic, S.: An analytical solution for a tandem queue with blocking. Queueing Syst. **36**(1–3), 221–235 (2000)

29. Green, L., Kolesar, P., Whitt, W.: Coping with time-varying demand when setting staffing requirements for a service system. Prod. Oper. Manag. **16**(1), 13–39 (2007)
30. Harrison, J.: Assembly-like queues. J. Appl. Probab. **10**(02), 354–367 (1973)
31. Harrison, J.: Brownian Motion and Stochastic Flow Systems. Wiley, New York (1985)
32. He, B., Liu, Y., Whitt, W.: Staffing a service system with non-Poisson non-stationary arrivals. Probab. Eng. Inf. Sci. **30**(4), 593–621 (2016)
33. Hillier, F., Boling, R.: Finite queues in series with exponential or Erlang service times—a numerical approach. Oper. Res. **15**(2), 286–303 (1967)
34. Katsaliaki, K., Brailsford, S., Browning, D., Knight, P.: Mapping care pathways for the elderly. J. Health Organ. Manag. **19**(1), 57–72 (2005)
35. Kelly, F.: Blocking, reordering, and the throughput of a series of servers. Stoch. Process. Appl. **17**(2), 327–336 (1984)
36. Koizumi, N., Kuno, E., Smith, T.: Modeling patient flows using a queuing network with blocking. Health Care Manag. Sci. **8**(1), 49–60 (2005)
37. Langaris, C., Conolly, B.: On the waiting time of a two-stage queueing system with blocking. J. Appl. Probab. **21**(03), 628–638 (1984)
38. Leachman, R., Gascon, A.: A heuristic scheduling policy for multi-item, single-machine production systems with time-varying, stochastic demands. Manag. Sci. **34**(3), 377–390 (1988)
39. Li, A., Whitt, W.: Approximate blocking probabilities in loss models with independence and distribution assumptions relaxed. Perform. Eval. **80**, 82–101 (2014)
40. Li, A., Whitt, W., Zhao, J.: Staffing to stabilize blocking in loss models with time-varying arrival rates. Probab. Eng. Inf. Sci. **30**(02), 185–211 (2016)
41. Li, J., Meerkov, S.: Production Systems Engineering. Springer, Berlin (2009)
42. Liu, Y., Whitt, W.: Large-time asymptotics for the $G_t/M_t/s_t + GI_t$ many-server fluid queue with abandonment. Queueing Syst. **67**(2), 145–182 (2011)
43. Liu, Y., Whitt, W.: A network of time-varying many-server fluid queues with customer abandonment. Oper. Res. **59**(4), 835–846 (2011)
44. Liu, Y., Whitt, W.: The $G_t/GI/s_t + GI$ many-server fluid queue. Queueing Syst. **71**(4), 405–444 (2012)
45. Liu, Y., Whitt, W.: A many-server fluid limit for the $G_t/GI/s_t + GI$ queueing model experiencing periods of overloading. Oper. Res. Lett. **40**(5), 307–312 (2012)
46. Liu, Y., Whitt, W.: Many-server heavy-traffic limit for queues with time-varying parameters. Ann. Appl. Probab. **24**(1), 378–421 (2014)
47. Ma, N., Whitt, W.: Efficient simulation of non-Poisson non-stationary point processes to study queueing approximations. Stat. Probab. Lett. **109**, 202–207 (2016)
48. Mandelbaum, A., Massey, W., Reiman, M.: Strong approximations for Markovian service networks. Queueing Syst. **30**(1–2), 149–201 (1998)
49. Mandelbaum, A., Massey, W., Reiman, M., Rider, B.: Time varying multiserver queues with abandonment and retrials. In: Proceedings of the 16th International Teletraffic Conference (1999)
50. Mandelbaum, A., Pats, G.: State-dependent queues: approximations and applications. Stoch. Netw. **71**, 239–282 (1995)
51. Mandelbaum, A., Pats, G.: State-dependent stochastic networks. Part I. Approximations and applications with continuous diffusion limits. Ann. Appl. Probab. **8**(2), 569–646 (1998)
52. Martin, J.: Large tandem queueing networks with blocking. Queueing Syst. **41**(1–2), 45–72 (2002)
53. Meerkov, S., Yan, C.B.: Production lead time in serial lines: evaluation, analysis, and control. IEEE Trans. Autom. Sci. Eng. **13**(2), 663–675 (2016)
54. Millhiser, W., Burnetas, A.: Optimal admission control in series production systems with blocking. IIE Trans. **45**(10), 1035–1047 (2013)
55. Nahmias, S., Cheng, Y.: Production and Operations Analysis, vol. 5. McGraw-Hill, New York (2009)
56. Namdaran, F., Burnet, C., Munroe, S.: Bed blocking in Edinburgh hospitals. Health Bull. **50**(3), 223–227 (1992)
57. Oliver, R., Samuel, A.: Reducing letter delays in post offices. Oper. Res. **10**(6), 839–892 (1962)
58. Osorio, C., Bierlaire, M.: An analytic finite capacity queueing network model capturing the propagation of congestion and blocking. Eur. J. Oper. Res. **196**(3), 996–1007 (2009)
59. Pang, G., Whitt, W.: Heavy-traffic limits for many-server queues with service interruptions. Queueing Syst. **61**(2), 167–202 (2009)

60. Pender, J.: Nonstationary loss queues via cumulant moment approximations. Probab. Eng. Inf. Sci. **29**(1), 27–49 (2015)
61. Pender, J., Ko, Y.: Approximations for the queue length distributions of time-varying many-server queues. INFORMS J. Comput. **29**(4), 688–704 (2017)
62. Perros, H.: Queueing Networks with Blocking. Oxford University Press Inc, Oxford (1994)
63. Prabhu, N.: Transient behaviour of a tandem queue. Manag. Sci. **13**(9), 631–639 (1967)
64. Reed, J., Ward, A., Zhan, D.: On the generalized drift Skorokhod problem in one dimension. J. Appl. Probab. **50**(1), 16–28 (2013)
65. Rubin, S., Davies, G.: Bed blocking by elderly patients in general-hospital wards. Age Ageing **4**(3), 142–147 (1975)
66. Srikant, R., Whitt, W.: Simulation run lengths to estimate blocking probabilities. ACM Trans. Model. Comput. Simul. (TOMACS) **6**(1), 7–52 (1996)
67. Takahashi, Y., Miyahara, H., Hasegawa, T.: An approximation method for open restricted queueing networks. Oper. Res. **28**(3–part–i), 594–602 (1980)
68. Tolio, T., Gershwin, S.: Throughput estimation in cyclic queueing networks with blocking. Ann. Oper. Res. **79**, 207–229 (1998)
69. Travers, C., McDonnell, G., Broe, G., Anderson, P., Karmel, R., Duckett, S., Gray, L.: The acute-aged care interface: exploring the dynamics of bed blocking. Aust. J. Ageing **27**(3), 116–120 (2008)
70. Vandergraft, J.: A fluid flow model of networks of queues. Manag. Sci. **29**(10), 1198–1208 (1983)
71. van Vuuren, M., Adan, I., Resing-Sassen, S.: Performance analysis of multi-server tandem queues with finite buffers and blocking. OR Spectr. **27**(2–3), 315–338 (2005)
72. Wenocur, M.: A production network model and its diffusion approximation. Technical report, DTIC Document (1982)
73. Whitt, W.: The best order for queues in series. Manag. Sci. **31**(4), 475–487 (1985)
74. Whitt, W.: Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and their Application to Queues. Springer, Berlin (2002)
75. Whitt, W.: Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. Manag. Sci. **50**(10), 1449–1461 (2004)
76. Whitt, W.: Two fluid approximations for multi-server queues with abandonments. Oper. Res. Lett. **33**(4), 363–372 (2005)
77. Whitt, W.: Fluid models for multiserver queues with abandonments. Oper. Res. **54**(1), 37–54 (2006)
78. Whitt, W.: What you should know about queueing models to set staffing requirements in service systems. Nav. Res. Logist. (NRL) **54**(5), 476–484 (2007)
79. Whitt, W.: OM forum—offered load analysis for staffing. Manuf. Serv. Oper. Manag. **15**(2), 166–169 (2013)
80. Yom-Tov, G., Mandelbaum, A.: Erlang-R: a time-varying queue with reentrant customers, in support of healthcare staffing. Manuf. Serv. Oper. Manag. **16**(2), 283–299 (2014)
81. Zychlinski, N., Mandelbaum, A., Momčilović, P., Cohen, I.: Bed blocking in hospitals due to scarce capacity in geriatric institutions—cost minimization via fluid models. Working paper (2017)