Time-Varying Fluid Networks with Blocking: Models Supporting Patient Flow Analysis in Hospitals

Noa Zychlinski

Time-Varying Fluid Networks with Blocking: Models Supporting Patient Flow Analysis in Hospitals

Research Thesis

In Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

Noa Zychlinski

Submitted to the Senate of the Technion - Israel Institute of Technology

Tamuz, 5778, Haifa, June, 2018

The Research Thesis Was Done Under The Supervision of Prof. Avishai Mandelbaum and Dr. Izack Cohen in The Faculty of Industrial Engineering and Management Technion – Israel Institution of Technology

The Generous Financial Help of The Technion and The Israeli Ministry of Science, Technology and Space are Gratefully Acknowledged

I would like to express my deep appreciation and gratitude to my advisers Prof. Avishai Mandelbaum and Dr. Izack Cohen, as well as to Prof. Petar Momčilović for their endless encouragement, advice and guidance throughout my studies. Finally, I would like to thank my family for their unconditional love and support.

Publications:

1. Zychlinski, N., Mandelbaum, A., Momčilović, P., and Cohen, I. (2018). Bed blocking in hospitals due to scarce capacity in geriatric institutions – cost minimization via fluid models. Under revision in Manufacturing and Service Operations Management (MSOM).

2. Zychlinski, N., Mandelbaum, A., and Momčilović, P. (2018). Time-varying tandem queues with blocking: Modeling, analysis and operational insights via fluid models with reflection. Queueing Systems, 89(1), 15-47.

3. Zychlinski, N., Mandelbaum, A., and Momčilović, P. (2018). Time-varying many-server finite-queues in tandem: Comparing blocking mechanisms via fluid models. Under revision in Operations Research Letters.

Contents

List of Figures

List of Tables

Abstract

This thesis was motivated by the bed blocking problem, which occurs when elderly hospital patients are ready to be discharged, but must remain in the hospital until a bed in a geriatric institution becomes available. Bed blocking has become a challenge to healthcare operators due to its economic implications and quality-of-life effect on patients. Indeed, hospital-delayed patients, who cannot access their most appropriate treatment (e.g. rehabilitation), prevent new admissions. Moreover, bed blocking is costly since a hospital bed is more expensive to operate than a geriatric bed.

The first part of this thesis (Section 2) focuses on analyzing the bed blocking problem, in order to improve the joint operation of hospitals and geriatric institutions. To this end, we develop a mathematical fluid model, which accounts for blocking, mortality and readmission—all significant features of the discussed environment. The comparison between our fluid model, a two-year data set from a hospital chain and simulation results shows that our model is accurate and useful. Then, for bed allocation decisions, the fluid model and especially its offered-load counterpart turn out insightful and easy to implement. Our analysis yields a closed-form expression for bed allocation decisions, which minimizes the sum of underage and overage costs. The proposed solution demonstrates that significant reductions in cost and waiting list length are achievable, as compared to current operations.

A more comprehensive view of the system analyzed in Section 2 can be achieved by including Emergency Department (ED) boarded patients, waiting for admission to hospital wards. This analysis should also include finite waiting rooms and customer loss when they are full. Accordingly, we set out to model and analyze time-varying tandem networks with blocking and finite waiting rooms throughout the network (Section 3). These models capture the essential characteristics of our first model–namely, time-variation and blocking; in this case, however, accommodating customer loss requires reflection analysis. We conclude this section by providing operational insights on network performance of tandem flow lines, in a broader perspective that goes beyond hospital networks.

Sections 2 and 3 focus on Blocking After Service (BAS). Section 4, however, focuses on the Blocking Before Service (BBS) mechanism. BBS arises in telecommunication networks, production lines and healthcare systems. We begin by modeling the stochastic queueing network of time-varying tandem networks with finite buffers throughout the network; then, we develop its corresponding fluid limit and provide design/operational insights regarding BAS/BBS mechanisms; in particular, on network throughput and job loss rate.

List of Abbreviations and Notation

Abbreviations

ED	Emergency Department
LOS	Length of Stay
BAS	Blocking After Service
BBS	Blocking Before Service
FCFS	First Come First Served
i.i.d.	independent and identically distributed
DE	Differential Equation
LWBS	Left Without Being Seen
MSHT	Many-server heavy-traffic
FSLLN	Functional Strong Law of Large Numbers
RMSE	Root Mean Square Error
u.o.c.	uniformly on compact
a.s.	almost surely

Notation

$\lambda(t)$	External arrival rate to Station 1 at time t
μ_i	Service rate at Station i
N_i	Number of servers/beds at Station i
$p_{ij}(t)$	Routing probability from Station i to j at time t
$X_1(t)/x_1(t)$	Number of arrivals to Station 1 that have not completed their service at Station 1 at time t (stochastic process/fluid limit)
$X_i/x_i(t)$	Number of customers that have completed service at Station 1, require service at Station i, but have not yet completed their service at Station i at time t (stochastic process/fluid limit)
$Q_i(t)/q_i(t)$	Number of customers in Station i at time t (stochastic process/fluid limit)
$B(t)/b(t)$	Number of blocked customers at time t (stochastic process/fluid limit)

Specific Notations for Section 2

θ_i	Individual mortality rate at Station i
β_i	Readmission rate from Station i back to hospital
$\delta_r(t)$	Treatment completion rate at Station 1 at time t
$\delta_{total}(t)$	Total departure (mortality and treatment completion) rate from Station 1
$r_i(t)$	Offered load in Station i at time t
T	Planning horizon
C_{o_i}	Overage cost per day per bed at Station i
C_{u_i}	Underage cost per day per bed at Station i
N^*	Optimal number of beds
I	The fraction of time during which underage costs were incurred
$\bar{I}$	An estimator for I
K	Fixed setup cost associated with the introduction of each new geriatric bed
B	The current bed capacity
N_K^*	Optimal number of beds when including setup cost for new beds
C_r	Reallocating cost associated with adding and removing a geriatric bed
$N_{\mathcal{I}}^*$	Optimal number of beds for a fixed period $\mathcal{I}$

Specific Notations for Sections 3 and 4

k	Number of stations in the network
H_i	Waiting room/buffer before Station i
$\bar{q}_i$	Steady-state number of jobs in Station i
$s_i^{\mathrm{BBS}}/s_i^{\mathrm{BAS}}$	Steady-state number of jobs in service at Station i under BBS/BAS
$\delta^{\mathrm{BBS}}/\delta^{\mathrm{BAS}}$	Steady-state throughput of the network under BBS/BAS
$\gamma^{\mathrm{BBS}}/\gamma^{\mathrm{BAS}}$	Steady-state rate of loss jobs under BBS/BAS

1 Introduction

Providing high quality healthcare services for the ageing population is becoming a major challenge in developed countries. This challenge is amplified by the fact that the number of elderly people, aged 65 and over who today account for 10% of the population, will double within two decades (World Health Organization, 2014; United Nations Population Fund, 2014). Moreover, elderly patients are often frail and undergo frequent hospitalizations. These facts are and will increasingly be major contributors to the high occupancy levels in inpatient wards and EDs. For example, in the last several years, some OECD countries reported averages of over 90% occupancy levels in hospital inpatient wards (OECD iLibrary - Health at a Glance, 2013; NHS England - Bed Availability and Occupancy Data, 2015); and these yearly averages hardly reveal the hour-by-hour reality of the busiest periods (e.g. winters).

The bed blocking problem occurs when hospital patients are ready to be discharged, but must remain in the hospital until a bed in a more appropriate geriatric facility (a nursing home or a geriatric institution) becomes available. Research about the bed blocking problem (e.g. Rubin and Davies, 1975; Namdaran et al., 1992; El-Darzi et al., 1998; Koizumi et al., 2005; Cochran and Bharti, 2006; Travers et al., 2008; Osorio and Bierlaire, 2009; Shi et al., 2015) is important since it can potentially improve the quality of patient care and reduce the mounting costs associated with bed blocking (Cochran and Bharti, 2006). For example, the estimated cost of bed blocking in the UK alone exceeds 1.2 billion dollars per year (BBC News, 2016). In contrast to previous models, which relied on simulations for modeling bed blocking, our research offers an analytical model for minimizing the overage and underage costs of a system consisting of hospitals and geriatric institutions; the model yields a tractable solution by determining the optimal number of beds for each geriatric ward.

Patient flow (Figure 1) begins when elderly people turn to the ED due to a clinical deterioration or a health crisis. After stabilizing their condition, doctors decide on discharge or hospitalization. Patients can also be hospitalized without going through the ED in cases of elective procedures. Upon treatment completion, hospital doctors decide whether the patient is capable of returning to the community, needs to be admitted to a nursing home, or requires further treatment in a geriatric institution. We subdivide the latter option into the three most common geriatric wards: reha-

bilitation, mechanical ventilation and skilled nursing care. In Section 2 we focus on these three wards together with the hospital inpatient wards (i.e. the four framed stations in Figure 1) since, in our setting and according to the data we analyze, the problem in geriatric institutions is much more severe than in regular nursing homes. Having said that, our modeling framework accommodates any environment, in which the phenomenon of blocking is severe and gives rise to operational challenges.

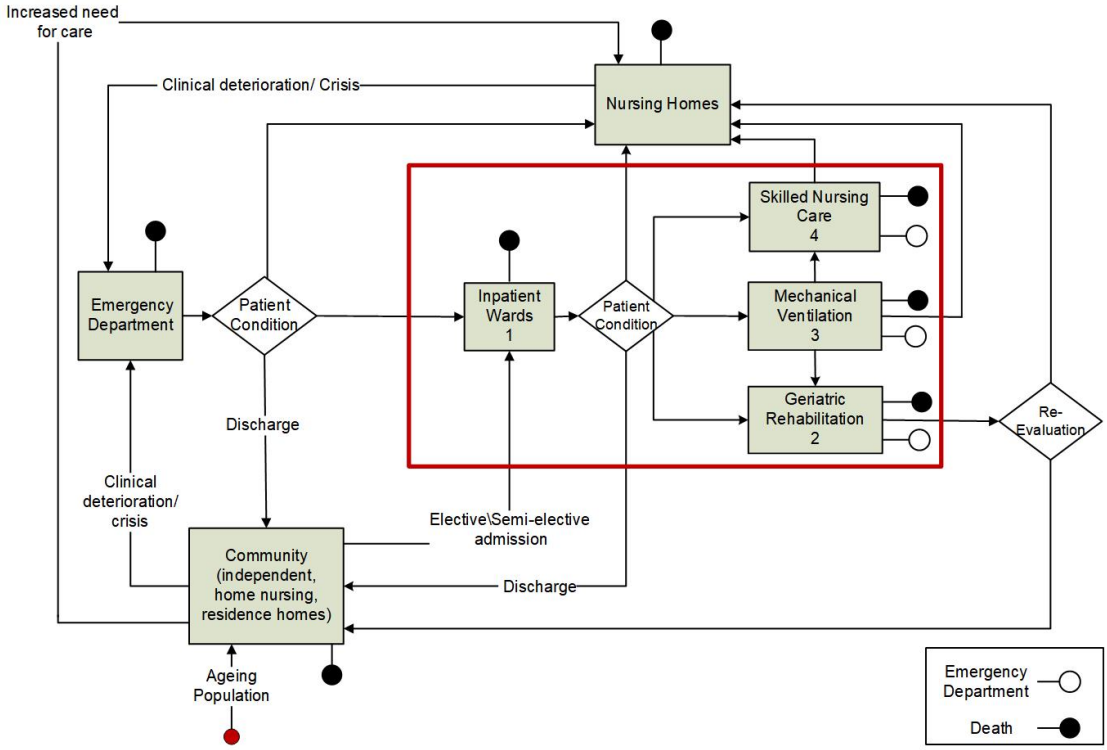


Figure 1: Network of patient flow through the community, inpatient wards, nursing homes and geriatric institutions. The readmission sign substitutes for an arrow from Station 2,3 or 4 back to Station 1.

In Section 2 we develop a mathematical fluid model, which accounts for blocking, mortality and readmission—all significant features of the discussed environment. Then, for bed allocation decisions, the fluid model and especially its offered-load counterpart turn out insightful and easy to implement. We compare our fluid model with a two-year data set from a hospital chain and simulation results. These comparisons show that our model is accurate and useful. Moreover, our analysis yields a closed-form expression for bed allocation decisions, which minimizes the sum of underage and overage costs. Solving for the optimal number of geriatric beds in our system demonstrates that significant reductions in cost and waiting list length are achievable, as compared to current operations. In addition, we propose two feasible extensions for capacity

allocation problems with time-varying demand of beds: a periodic reallocation of beds and the incorporation of setup costs into bed allocation decisions.

Achieving a more comprehensive view of the system analyzed in Section 2 can be done by including ED boarded patients waiting for admission to hospital wards. This analysis should also include finite waiting room before the first station and customer loss when this waiting room is full. Accordingly, in Section 3, we model and analyze time-varying multi-server tandem networks with blocking and finite waiting rooms throughout the network – before the first station and between the stations. These models capture the essential characteristics of the model analyzed in Section 2 – namely, time-variation and blocking; in these models, however, accommodating customer loss requires reflection analysis.

In order to analyze these networks, we begin with the stochastic queueing model of time-varying multi-server flow-lines with finite buffers throughout. Then, we develop fluid models for these networks and justify them by establishing many-server heavy-traffic (MSHT) functional strong law of large numbers (FSLLNs). We conclude Section 3 by providing operational insights on network performance derived from our models; specifically the effects of line length, bottleneck location, waiting room size, and the interaction among these effects.

The models analyzed in Sections 2 and 3 focus on the Blocking After Service (BAS) mechanism. Section 4, however, focuses on Blocking Before Service (BBS). Under the latter, a service can begin at Station i, only when there is available capacity (buffer space/server) at Station $i+1$. As in Section 3, we begin by modeling the stochastic queueing networks and then, by establishing a many-server heavy-traffic (MSHT) functional strong law of large numbers (FSLLNs), we develop fluid models for these networks. Finally, we analytically compare and provide design/operational insights regarding the two blocking mechanisms; in particular, on network throughput and job loss rate.

Each of the three main sections in this thesis is based on a research paper; namely: Section 2 is based on Zychlinski et al. (2018c), Section 3 on Zychlinski et al. (2018b) and Section 4 on Zychlinski et al. (2018a).

2 Bed Blocking in Hospitals

2.1 Introduction

Congestion problems and their highly significant effect, both medically and financially, motivated us to model and analyze the system, depicted schematically in Figure 2 (which is the framed sub-system in Figure 1). Patient flow begins when people of all ages are admitted to hospital inpatient wards. Upon treatment completion, and focusing on geriatric patients, hospital doctors decide whether the patient is capable of returning to the community or requires further care in a geriatric institution. We subdivide the latter option into the three most common long-term care geriatric wards: rehabilitation, mechanical ventilation and skilled nursing care.

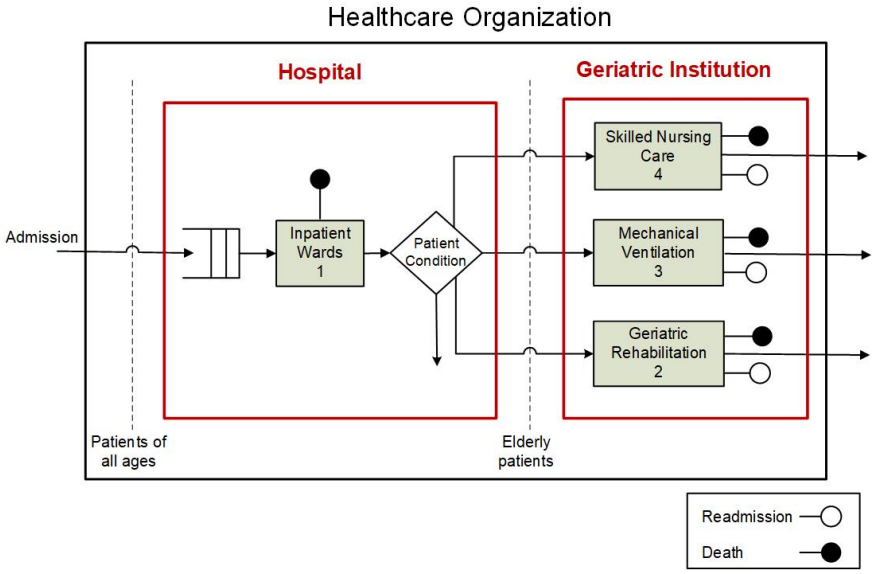


Figure 2: Network of patient flow through inpatient wards and geriatric institutions. The readmission sign substitutes for an arrow from Station 2,3 or 4 back to Station 1.

Patients who are sent to a geriatric rehabilitation ward stay there one month on average, before they are able to return to full or partial functioning. Mechanical ventilation wards treat patients who cannot breathe on their own, typically after three unsuccessful weaning attempts in a hospital; the average stay in a mechanical ventilation ward is 5–6 months. Unfortunately, only a minority of these patients are discharged; most die or are readmitted to hospitals. Skilled nursing wards treat patients who, in addition to functional dependency, suffer from active diseases that require close medical supervision, for example due to bedsores or chemotherapy; the average stay there is 1–1.5 months. Some patients are discharged to nursing homes but, again, most either

die or are readmitted to hospitals.

In our setting, the central decision maker is a large healthcare organization that operates several hospitals and several geriatric institutions. In some countries (e.g. Singapore and Israel), the government functions as this organization. In England, the NHS, an arm of the government, is the central decision maker; in Australia it is the Medicare Healthcare System; and in the U.S., it can be the Veterans Administration (VA) with its 500+ hospitals.

The methodology we propose is rather general and can accommodate other settings, with a different number or type of wards. Since the system we analyze and the data we use are for three types of geriatric wards, in the empirical part of the paper, we focus on the four stations depicted in Figure 2: Inpatient wards (Station 1), Geriatric Rehabilitation (Station 2), Mechanical Ventilation (Station 3) and Skilled Nursing Care (Station 4). Applying our general methodology to analyzing these stations, for which there are long waiting lists, will yield policies that significantly reduce total operational costs.

To this end, we develop a mathematical fluid model that accounts for blocking, mortality and readmission—all significant features of the discussed environment. Then, we use our fluid model and its time-varying offered-load counterpart to formulate and solve bed allocation problems for geriatric wards. Our goal is to find the optimal number of geriatric beds, in order to minimize the total overage plus underage costs of the system. Moreover, we propose two feasible extensions for capacity allocation problems with time-varying demand of beds: a periodic reallocation of beds and the incorporation of setup costs into bed allocation decisions.

In our analysis we use two data sets, over a period of two years. The first covers the patient flow in a hospital chain comprising four hospitals and three geriatric institutions (three rehabilitation wards, two mechanical ventilation wards and three skilled nursing wards). The second data set includes individual in-hospital waiting lists for each geriatric ward. (Details about our data are provided in Appendix A.) These data indicate that the average in-hospital waiting times are 28 days for mechanical ventilation, 17 days for skilled nursing care and 3.5 days for rehabilitation wards. Although the average waiting time for rehabilitation seems relatively short, this is definitely not the case when considering the fact that these are elderly patients, waiting unnecessarily for their rehabilitation care, while occupying a bed that could have been used for

newly admitted acute patients. Moreover, the number of patients who are referred to a rehabilitation ward is 5 and 9 times that of the corresponding numbers for skilled nursing care and mechanical ventilation, respectively; this implies (Section 2.6) that the overall demand they generate exceeds that of the other patients.

Figure 3 presents the waiting list lengths (daily resolution) within the hospital, for each geriatric ward over one calendar year. The dotted lines represent length according to our data, while the solid lines represent our fluid model (Equations (6)–(7) in the sequel). According to this plot, all three geriatric wards work at full capacity throughout the year (long waiting lists); furthermore, in the winter, the demand for beds increases.

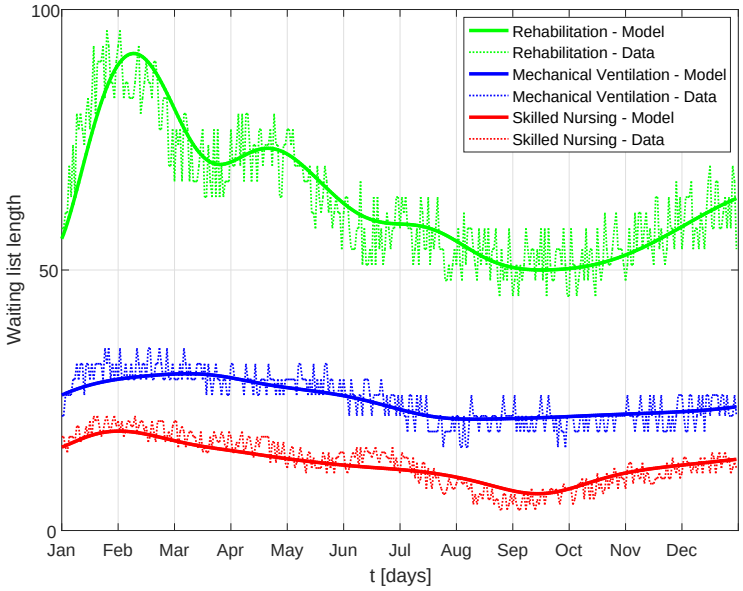


Figure 3: Waiting list length in hospital for each geriatric ward - model (solid lines) vs. data (dashed lines). The X axis is one calendar year in units of days. (We are plotting here the 2nd year of our data. The 1st year was used to fit the parameters of our model.)

The fit between our model and the data is excellent. In fact, in Appendix A we demonstrate, via multiple scenarios with various treatment distributions, that our continuous, deterministic fluid model approximates well and usefully its underlying stochastic environment.

The long waiting lists, and the fact that hospitalization costs are much higher in hospitals than in geriatric institutions, indicate that the system is operated inefficiently; this leads to excessive costs that can be reduced by adopting our solution. Moreover, in Sections 2.7.1 and 2.8.3 we demonstrate how the constant and periodic

allocations we suggest can reduce costs and shorten waiting lists. (The latter is illustrated in Figure 4 (right) and Figure 5 (bottom left); this is relative to the current waiting list lengths presented in Figure 3.)

2.2 Literature Review

The review covers the main areas that are relevant to this research: high-level modeling of healthcare systems, queuing networks with blocking, time-varying queueing networks and bed planning in long-term care facilities.

2.2.1 High-level Modeling of Healthcare Systems

The three main approaches used for modeling healthcare systems with elderly patients have been Markov models, system dynamics and discrete event simulation.

For tractability reasons, Markov models have been applied to networks with a limited number of stations, typically 2–3, in order to characterize steady-state performance such as length of stay (LOS) at each station. For example, Harrison and Millard (1991) analyze the empirical distribution of patient LOS in geriatric wards by fitting a sum of two exponentials to a data set: most patients are discharged or die shortly after admission, while some stay hospitalized for months. Several papers use Markov models to describe the flow of geriatric patients between hospitals and community-based care (Taylor et al., 1997, 2000; Xie et al., 2005; Faddy and McClean, 2005; McClean and Millard, 2006). In general, these models, which include short-stay and long-stay states in each facility, distinguish between the movement of patients within and between facilities. Differently from these papers, our approach emphasizes station capacity and time-varying parameters.

Another common approach for modeling healthcare systems is system dynamics. It is used to analyze patient flow through healthcare services by focusing on the need to coordinate capacity levels across all health services. Wolstenholme (1999) develops a patient flow model for the UK National Health Service and uses it to analyze alternatives for shortening waiting times of community care patients. According to the author, reducing total waiting times can be achieved by adding 'intermediate care' facilities, which are aimed at preventing elderly medical patients from hospitalization and community care. Our approach contributes to this line of research by considering

the dependency between capacity allocation and waiting time.

System dynamics is also used to analyze the bed blocking problem (Gray et al., 2006; Travers et al., 2008; Rohleder et al., 2013). These papers demonstrate the importance of coordinating capacity levels across different health services. Desai et al. (2008) use system dynamics to forecast the future demand for social care services by elderly people. While our proposed fluid model is also deterministic, we are able to justify it as the fluid limit of an underlying stochastic model/system.

Discrete event simulation is another popular approach for analyzing complex systems and phenomena such as bed blocking. El-Darzi et al. (1998) describe patient flow through geriatric wards, by examining the impact of bed blocking and occupancy on patient flow. They show that the availability of acute beds is strongly connected to referral rates for long-stay care facilities. Katsaliaki et al. (2005) build a simulation model of elderly patient flow between the community, hospitals and geriatric institutions. They approximate the delay in discharge from hospital and the relevant costs. Shi et al. (2015) and Armony et al. (2015) discuss a two-time-scale (days and hours) service time in hospital wards. Shi et al. (2015) investigate ED boarding times (waiting for admission to hospital wards) at a Singaporean hospital. Via simulation studies, they examine the effects of various discharge policies on admission waiting times. The two-time-scale service time captures both treatment time and additional service time caused by operational factors, such as discharge schedule. In our research, we develop a time-varying analytical model, for setting bed capacities in geriatric institutions. Our model evolves on a single time-scale – it is days since, for the decisions we are interested in (and the data we have), days are natural and adequate.

2.2.2 Queueing Networks with Blocking

Several blocking mechanisms are acknowledged in the literature (Perros, 1994; Balsamo et al., 2001). We focus on the *blocking-after-service* (BAS) mechanism, which happens when a patient attempts to enter a fully-capacitated Station j upon completion of treatment at Station i. Since it is not possible to queue in front of Station j, the patient must wait in Station i and therefore, blocks a bed there until a departure occurs at Station j.

Healthcare systems usually have complex network topologies, multiple-server queues and time-varying dynamics. In contrast, closed-form solutions of queueing models with

blocking exist only for steady-state, single-server networks with two or three tandem queues or with two cyclic queues (Osorio and Bierlaire, 2009). The solutions for more complex networks are based on approximations, which are typically derived via decomposition methods (Hillier and Boling, 1967; Takahashi et al., 1980; Gershwin, 1987; Koizumi et al., 2005; Osorio and Bierlaire, 2009) and expansion methods (Kerbache and MacGregor Smith, 1987, 1988; Cheah and Smith, 1994). Koizumi et al. (2005) use a decomposition method to analyze a healthcare system with mentally disabled patients as a multiple-server queueing network with blocking, while Osorio and Bierlaire (2009) develop an analytic finite capacity queueing network that enables the analysis of patient flow and bed blocking in a network of hospital operative and post-operative units.

Bretthauer et al. (2011) offer a heuristic method, for estimating the waiting time for each station in a tandem queueing network with blocking, by adjusting the per-server service rate to account for blocking effects. Bekker and de Bruin (2010) analyze the effect of a predictable patient arrival pattern, to a clinical ward, on its performance and bed capacity requirements. In particular, the authors use the offered-load approximation and the square-root staffing formula for calculating the required beds for each day of the week. Although we also use the offered-load approximation for the time-varying demand, our approach is different, since it goes beyond a single-station analysis and takes into account blocking effects by minimizing overage and underage costs. Moreover, the periodic reallocation we suggest takes into account a reallocation cost that is associated with adding and removing a bed.

Capturing blocking in stochastic systems with a single-station in steady-state has been done via reflection. Specifically, reflection is a mathematical mechanism that has been found necessary to capture customer loss (see Whitt, 2002, Chapter 5.2 and Garnett et al., 2002). Reflection modeling, however, requires the use of indicators, which cause technical continuity problems when calculating approximating limits. We circumvent this challenge by developing a fluid model with blocking yet without reflection, which enables us to prove convergence of our stochastic model without reflection. Our simple and intuitive model, compared to models with reflection, enables us to model, successfully and insightfully, time-varying networks.

2.2.3 Queueing Networks with Time-Varying Parameters

Time-varying queueing networks have been analyzed by McCalla and Whitt (2002), who focused on long service lifetimes, measured in years, in private-line telecommunication services. Liu and Whitt (2011b) analyze time-varying networks with many-server fluid queues and customer abandonment. In addition, time-varying queueing models have been analyzed for setting staffing requirements in service systems with unlimited queue capacity, by using the offered-load analysis (Whitt, 2013). The methods for coping with time-varying demand when setting staffing levels are reviewed in Green et al. (2007a) and Whitt (2007). A recent work of Li et al. (2015) focuses on stabilizing blocking probabilities in loss models with a time-varying Poisson arrival process, by using a variant of the modified-offered-load (MOL) approximation.

Fluid frameworks are well adapted to large, time-varying overloaded systems (Mandelbaum et al., 1998, 1999), which is the case here. Previous research shows that fluid models have been successfully implemented in modeling healthcare systems (Ata et al., 2013; Yom-Tov and Mandelbaum, 2014; Cohen et al., 2014). Moreover, fluid models yield analytical insights, which typically cannot be obtained using their alternatives (e.g. simulation, time-varying stochastic queueing networks).

2.2.4 Bed Planning for Long-term Care Facilities

Most research on bed planning in healthcare systems focuses on short-term facilities, such as hospitals (Green, 2004; Akcali et al., 2006). Research about bed planning for long-term care facilities is scarce. We now review the existing literature.

Future demand for long-term care has a strong impact on capacity setting decisions. Hare et al. (2009) develop a deterministic model for predicting future long-term care needs in home and community care services in Canada. Zhang et al. (2012) develop a simulation-based approach to find the minimal number of nursing home beds in order to achieve a target waiting time. The model we suggest considers time-varying demand for beds throughout the year, as well as mortality and readmission rates which are all significant in the context of geriatric patients. In addition, we analyze a network capacity problem of several geriatric wards by taking into account blocking effects in hospitals.

De Vries and Beekman (1998) present a deterministic dynamic model for expressing

waiting lists and waiting times of psycho-geriatric patients for nursing homes, based on data from the previous year. Ata et al. (2013) analyze the expected profit of hospice care. They propose an alternative reimbursement policy for the United States Medicare and determine the recruiting rates of short and long stay patients to maximize profitability of the hospice. Kao and Tung (1981) consider the monthly fluctuation in demand for hospital services, yet the bed allocation they allow is constant throughout the year. In particular, they try to minimize the hospital yearly average overflow probability. To accommodate for the seasonal demand, we suggest a periodic reallocation of beds, which takes into account a reallocation cost that is associated with adding and removing each bed.

Harrison and Zeevi (2005) develop a method, which was extended in Bassamboo et al. (2006), for staffing large call centers with multiple customer classes and multiple server pools; they deploy stochastic fluid models to minimize the sum of personnel costs and abandonment penalties. The method they suggest reduces the staffing problem to a multidimensional Newsvendor problem and hence, the critical fractile solution they suggest is distribution dependent. In Remark 2.3, we further elaborate on the relation of Harrison and Zeevi (2005) to the present work.

Afèche et al. (2017) develop a fluid model for maximizing the profit of service firms by determining customer acquisition investment as well as capacity allocation. Our research includes finite capacities and time-variation; we also go beyond a single-station analysis to a network analysis. This allows us to consider the blocking customers, occupying servers in the first station, and explicitly accommodate the blocking costs when calculating the optimal number of beds. Moreover, we justify the fluid model by proving convergence of the corresponding stochastic model.

2.3 Contributions

The main contributions of this section are:

1. Modeling: We develop and analyze an analytical model comprising both long-term care geriatric wards and their feeding hospitals. This joint modeling is necessary in order to capture blocking effects (while previous research was restricted to a single-station utility maximization; e.g. Jennings et al. (1997)). This is done by explicitly considering geriatric ward blocking costs and minimizing the overall

underage and overage cost within the system.

2. Methodology: Our work contributes to the literature on queueing (fluid) networks with blocking. In particular, our proposed fluid model captures blocking without the need for reflection (see Section 2.2.2), and it applies to general networks (for example, networks with multiple stations in tandem). We use our model to derive analytical solutions and insights about cost minimization and bed allocation policies. The modeling approach accommodates time-varying systems, jointly with finite capacity considerations, patient mortality and readmissions—all of these are prevalent features in healthcare.

3. Practice: This research gives rise to new capacity allocation strategies. Specifically, we offer a closed-form solution for periodic reallocation of beds that accounts for seasonal demand, and an analytical model that incorporates setup costs. This is but two examples, made analyzable by our model, that demonstrates how our framework would yield managerial recommendations for healthcare managers in allocating geriatric beds.

2.4 The Model

In this section, we describe our environment and its dynamics. We then formally introduce model notations and equations.

2.4.1 Environment, Dynamics and Notations

Consider the four stations in Figure 2: hospital wards (Station 1) and long-term care geriatric wards—rehabilitation (Station 2), mechanical ventilation (Station 3), and skilled nursing care (Station 4). Station 1 includes all ward patients, while Stations 2–4 include only geriatric patients that need long-term care beyond hospitalization.

Our model is at the macro level; thus the capacity of each station is an aggregation of the individual capacities of all stations of this type in the discussed geographical area (e.g. assume that a district includes three rehabilitation wards; then the capacity of the modeled rehabilitation station is the sum of all three individual capacities). Such aggregated capacities are justified since, in practice, patients can be sent from any individual hospital to any individual geriatric ward and vice versa, especially if they are all within the same geographic area (a city or a district).

We model the exogenous arrival rate to hospital wards as a continuous time-varying function $\lambda(t)$ (see Mandelbaum et al., 1999). Internal arrivals are patients returning from geriatric wards back to the hospital. Hospital wards include N_1 beds. If there are available beds, arriving patients are admitted and hospitalized; otherwise, they wait in the queue. We assume that hospital wards have an unlimited queue capacity, since the ED serves as a queue buffer for them (our model does, nevertheless, accommodate blocking of the first station). Patients leave the queue either when a bed becomes available or if they, unfortunately, die. Medical treatment is performed at a known service rate μ_1. Upon treatment completion, patients are discharged back to the community, admitted to nursing homes, or referred to a geriatric ward (2, 3 or 4) with routing probabilities $p_{1i}(t)$, $i = 2, 3, 4$, respectively. The number of beds in each geriatric ward i, $i = 2, 3, 4$, is N_i. If there are no available beds in the requested geriatric ward, its referred patients must wait in the hospital while blocking their current bed. This blocking mechanism is known as *blocking-after-service* (Balsamo et al., 2001). The treatment rates in Stations i, $i = 2, 3, 4$, are μ_i. Frequently, the clinical condition of patients deteriorates while hospitalized in a geriatric ward, and they are hence readmitted to the hospital according to rate β_i, $i = 2, 3, 4$.

As mentioned, patients do die during their stay in a station, which we assume occurs at individual mortality rates θ_i, $i = 1, 2, 3, 4$, for Stations 1–4. These mortality rates are significant and cannot be ignored. We follow the modeling of mortality as in Cohen et al. (2014) and, in queueing theory parlance, refer to it as "abandonments" that can occur while waiting or while being treated. Although we use the same mortality rates while waiting and while being treated, if data prevail, our model can easily accommodate two different mortality rates per station.

2.4.2 Model Equations

We now introduce the functions $q_i(t)$, $i = 1, 2, 3, 4$, which denote the number of patients at Station i at time t. The standard fluid modeling approach defines differential equations describing the rate of change for each q_i. This direct approach has led to analytically intractable models that could not be justified as fluid limits of their corresponding stochastic counterparts. Moreover, these direct descriptions based on q_i included indicator functions which are harder to analyze due to their discontinuity. Hence, we propose a new modeling approach, in which we introduce alternative

functions $x_i(t)$, $i = 1,..4$, that suffice to capture the state of the system. Then, we develop differential equations for x_i, which are tractable, and ultimately deduce q_i from x_i. This novel modeling approach also simplifies the convergence proof of the corresponding stochastic model, which is provided in Appendix B.

The value $x_1(t)$ denotes the number of arrivals to Station 1 that have not completed their treatment at Station 1 at time t. The values $x_i(t)$, $i = 2, 3, 4$, denote the number of patients that have completed treatment at Station 1, require treatment at Station i, but have not yet completed their treatment at Station i at time t (these patients may still be blocked in Station 1). The dynamics of the system is captured through a set of differential equations (DEs); each characterizes the rate of change in the number of patients at each state at time t. Let $\lambda_{total}(t)$ denote the arrival rate to Station 1 at time t and $\delta_{total}(t)$ denote its departure rate. The DE for x_1 is, therefore

$$\dot{x}_1(t) \triangleq \frac{dx_1(t)}{dt} = \lambda_{total}(t) - \delta_{total}(t). \tag{1}$$

Patients arrive to Station 1 from two sources: externally, according to rate $\lambda(t)$, and internally from Stations 2, 3 and 4. Since β_i is the readmission rate from Station i back to Station 1, the internal arrival rate to Station 1 is $\sum_{i=2}^{4} \beta_i\big(x_i(t) \wedge N_i\big)$, where $x \wedge y = \min(x, y)$; here $\big(x_i(t) \wedge N_i\big)$ denotes the number of patients in treatment at Station i. The total arrival rate to Station 1 at time t is, therefore,

$$\lambda_{total}(t) = \lambda(t) + \sum_{i=2}^{4} \beta_i\big(x_i(t) \wedge N_i\big). \tag{2}$$

The total departure rate, $\delta_{total}(t)$, consists of two types. The first is due to patients who die at an individual mortality rate θ_1. Since patients might die while being hospitalized or waiting in queue, the rate at which patients die is $\theta_1 x_1(t)$. If data regarding different mortality rates while waiting (θ_{1q}) and while being treatment (θ_{1t}) prevail, then the total mortality from Station 1 would be

$$\theta_{1q}\Big[x_1(t) - \Big(N_1 - \sum_{i=2}^{4} \big(x_i(t) - N_i\big)^+\Big)\Big]^+ + \theta_{1t}\Big[x_1(t) \wedge \Big(N_1 - \sum_{i=2}^{4} \big(x_i(t) - N_i\big)^+\Big)\Big], \tag{3}$$

where the number of blocked patients waiting in Station 1 for a transfer to Station

i is $\big(x_i(t) - N_i\big)^+$. Therefore, the number of unblocked beds at Station 1 is $\Big(N_1 - \sum_{i=2}^{4}\big(x_i(t) - N_i\big)^+\Big)$, which can vary from 0 to N_1.

The second departure type, $\delta_r(t)$, is of patients who complete their treatment at Station 1. The rate at which patients complete their treatment in Station 1 is

$$\delta_r(t) = \mu_1\Big[x_1(t) \wedge \Big(N_1 - \sum_{i=2}^{4}\big(x_i(t) - N_i\big)^+\Big)\Big], \tag{4}$$

where the expression in the rectangular brackets indicates the number of occupied unblocked beds at Station 1. Thus, the total departure rate at time t is

$$\delta_{total}(t) = \theta_1 x_1(t) + \delta_r(t). \tag{5}$$

Using similar principles, we construct the DEs for the rate of change in x_i, $i = 2, 3, 4$. The referral rate to Station i is $p_{1i}(t)$ multiplied by $\delta_r(t)$, the rate at which patients complete their treatment at Station 1. The departure rate of patients who have completed service at Station 1, but not at Station i at time t consists of the mortality rate, $\theta_i x_i(t)$, readmission rate back to the hospital, $\beta_i\big(x_i(t) \wedge N_i\big)$ and treatment completion rate $\mu_i\big(x_i(t) \wedge N_i\big)$.

The set of DEs for x_i, $i = 1, 2, 3, 4$, is, therefore,

$$\begin{aligned}
&\dot{x}_1(t) = \lambda_{total}(t) - \delta_{total}(t),\\
&\dot{x}_i(t) = p_{1i}(t)\cdot\delta_r(t) - \beta_i\big(x_i(t) \wedge N_i\big) - \theta_i x_i(t) - \mu_i\big(x_i(t) \wedge N_i\big), \qquad i = 2, 3, 4.
\end{aligned} \tag{6}$$

The functions $q_i(t)$, $i = 1, 2, 3, 4$, which denote the number of patients at Station i at time t, are

$$\begin{aligned}
q_1(t) &= x_1(t) + \sum_{i=2}^{4}\big(x_i(t) - N_i\big)^+;\\
q_i(t) &= x_i(t) \wedge N_i, \qquad i = 2, 3, 4.
\end{aligned} \tag{7}$$

Note that $b_i(t)$, the number of blocked patients at Station 1 at time t, waiting for an available bed at Station i, $i = 2, 3, 4$, is given by $b_i(t) = \big(x_i(t) - N_i\big)^+$.

The validation of the model, both against data and a discrete event stochastic simulation with different treatment distributions, is detailed in Appendix A. It shows that there is an excellent fit between the fluid model, the actual data, and the corresponding simulation results.

2.5 The Bed-Allocation Model

The decision maker in our analysis is an organization that operates both hospitals and geriatric institutions. The objective is to find the optimal number of beds for each geriatric ward, so as to minimize overall long-term underage and overage cost of care (beds) in the system.

Minimizing overage and underage costs is a typical objective in resource allocation problems (Porteus, 2002). In our context, overage costs are incurred when geriatric beds remain empty while medical equipment, supply and labor costs are still being paid. We denote by C_o the per bed per day overage cost: this is the amount that could have been saved if the level of geriatric beds had been reduced by one unit in the event of an overage. This cost includes the per day labor, medical equipment and supply costs required for operating a geriatric bed. Underage cost, C_u, is incurred when patients are delayed in the hospital due to lack of availability in the geriatric wards. Thus, it is the amount that could have been saved if the level of geriatric beds had been increased by one unit in the event of an underage; C_u is hence the per bed per day cost of hospitalization in hospitals minus the per bed per day cost in geriatric institutions. To elaborate, hospitalization costs also include risk costs, which are incurred when a patient is required to remain hospitalized. These costs include expected costs of patient medical deterioration by not providing the proper medical treatment, and by exposing the patient to diseases and contaminations prevalent in hospitals. The sum of C_o and C_u, which will later on appear in the optimal solution in (16), amounts to the per bed per day hospitalization cost in hospitals. Excluding or underestimating the cost of risk will yield a lower bound for the required number of beds. Since our solution serves as a guide for thinking, meaningful insights can be derived already from such a lower bound.

We denote by C_{o_i} and C_{u_i} the overage and underage costs, respectively, for Stations i, $i = 2, 3, 4$. The resulting overall cost for Stations 2, 3 and 4 over a planning horizon T, is

$$C^{(0)}(N_2, N_3, N_4) = \sum_{i=2}^{4} C^{(0)}(N_i), \tag{8}$$

where $C^{(0)}(N_i)$ is the total overage and underage costs for each Station i, given by:

$$C^{(0)}(N_i) = \int_0^T \left[C_{u_i} \cdot b_i(t) + C_{o_i} \cdot \left(N_i - q_i(t) \right)^+ \right] \mathrm{d}t, \quad i = 2, 3, 4. \tag{9}$$

The first integrand is the underage cost, calculated by adding up the number of blocked patients, and the second integrand is the overage cost calculated via the total number of vacant beds. Minimizing (8) will yield a constant capacity level, for each geriatric ward, over the whole planning horizon. In Section 2.8.2 we introduce a periodic reallocation of beds, which yields several capacity levels for each ward during the planning horizon.

Remark 2.1. *Calculating the cost from (8) and (9) requires forecasting the arrival rate $\lambda(t)$, for the planning horizon $[0, T]$. This is done by using historical data: it shows that there is an annual arrival rate pattern that repeats itself, while the volume increases at a rather constant rate each year. Hence, our healthcare partners can accurately predict the arrival rate over the planning horizon.*

Minimizing (8), subject to (2)–(7), is analytically intractable, since $q_i(t)$ and $b_i(t)$ are solutions of a complex system of differential equations. To estimate the total cost, we use an offered-load approximation to the time-varying demand for beds (see Jennings et al., 1997; Whitt, 2007). Thus, in Section 2.6.1 we present a closed-form solution for minimizing the total underage and overage cost based on the offered load. Then, in Section 2.7.2 we compare our closed-form solution with a numerical solution of the original problem.

2.6 Offered Loads in Our System

Given a resource, its *offered load* $r = \{r(t), t \geq 0\}$ represents the average amount of work being processed by that resource at time t, under the assumption that waiting and processing capacity are ample (no one queues up prior to service). In our context, offered-load analysis is important for understanding demand. Indeed, we express demand in terms of patient-bed-days per day for the geriatric wards, in order to determine appropriate bed capacity levels.

The calculation of the offered load is carried out by solving (6) (and (2), (4), (5)) with an unlimited capacity in Stations 2, 3 and 4 ($N_i \equiv \infty$, $i = 2, 3, 4$). (Note that $b_i(t) \equiv 0$, for $i = 2, 3, 4$, which means that no patients are blocked.) These conditions

yield the following set of DEs for the offered load r_i, $i = 1, \ldots, 4$ (just substitute r_i for x_i in (6)):

$$\begin{aligned} \dot{r}_1(t) &= \lambda(t) + \sum_{i=2}^{4} \beta_i r_i(t) - \theta_1 r_1(t) - \mu_1\big(r_1(t) \wedge N_1\big), \\ \dot{r}_i(t) &= p_{1i}(t) \cdot \mu_1\big(r_1(t) \wedge N_1\big) - \big(\beta_i + \theta_i + \mu_i\big) r_i(t), \qquad i = 2, 3, 4. \end{aligned} \tag{10}$$

2.6.1 Estimating the Optimal Number of Beds based on the Offered load

The estimated overall cost for Stations 2, 3 and 4, based on the offered load over the planning horizon T, is

$$C(N_2, N_3, N_4) = \sum_{i=2}^{4} C(N_i); \tag{11}$$

here $C(N_i)$ is the underage plus overage cost for Station i, given by

$$C(N_i) = \int_0^T \Big[C_{u_i} \cdot \big(r_i(t) - N_i\big)^+ + C_{o_i} \cdot \big(N_i - r_i(t)\big)^+ \Big] \mathrm{d}t, \quad i = 2, 3, 4. \tag{12}$$

The first integrand corresponds to the underage cost, which is calculated by multiplying C_{u_i} with the (proxy for) bed shortage $(r_i(t) - N_i)^+$ and integrating it over the planning horizon. The second integrand, the overage cost, is obtained by multiplying C_{o_i} with the proxy for bed surplus $(N_i - r_i(t))^+$ and integrating it over the planning horizon as well.

Remark 2.2. *Why are these two proxies justified?*
First, under bed shortage (at cost C_{u_i} per bed), we substitute r_i for x_i. Second, under bed surplus (at cost C_{o_i} per bed), we substitute r_i for q_i. Third, since practically $C_{u_i} \gg C_{o_i}$ (see Section 2.7.1), the optimal solution must amplify reducing the number of blocked patients, hence the more significant cost is incurred by bed surplus. Finally, for calculating the latter cost and according to the offered-load definition, $q_i \approx r_i$ when the system is underloaded. And indeed, comparing the solutions according to the fluid model, to the offered-load approximation and to simulation results (Section 2.7.2), shows an excellent fit.

The offered load for each station is a known function of t, that depends solely on input parameters but not on N_2, N_3, N_4. Thus, minimizing (11) is, in fact, a separable problem, which can be solved for each station separately. (When doing so below, we

shall omit the i in (12) for simplicity of notations.)

To minimize $C(N)$, we adopt the approach of Jennings et al. (1997) and treat N as a continuous variable. We let $r_d = \{r_d(t)|\ 0 \leq t \leq T\}$ denote the *decreasing rearrangement* of r on the interval $[0, T]$: r_d on $[0, T]$ is characterized by being the unique decreasing function such that, for all $x \geq 0$, we have

$$\int_0^T 1_{\{r(t)\geq x\}} \mathrm{d}t = \int_0^T 1_{\{r_d(t)\geq x\}} \mathrm{d}t; \tag{13}$$

here $1_{\{r(t)\geq x\}}$ denotes the indicator function for the event $\{r(t) \geq x\}$. Existence and uniqueness of r_d were established in Hardy et al. (1952). The interpretation of Equation (13) is that both $r(t)$ and $r_d(t)$ spend the same amount of time above and under any level x. We can now rewrite $C(N)$ as follows:

$$\begin{aligned}
C(N) &= \int_0^T [C_u \cdot (r(t) - N)^+ + C_o \cdot (N - r(t))^+] \mathrm{d}t \qquad (14)\\
&= \int_N^\infty C_u \int_0^T 1_{\{r(t)\geq x\}} \mathrm{d}t\ \mathrm{d}x + \int_0^N C_o \int_0^T 1_{\{r(t)\leq x\}} \mathrm{d}t\ \mathrm{d}x \\
&= \int_0^\infty C_u \int_0^T 1_{\{r(t)\geq x\}} \mathrm{d}t\ \mathrm{d}x - \int_0^N C_u \int_0^T 1_{\{r(t)\geq x\}} \mathrm{d}t\ \mathrm{d}x + \int_0^N C_o [T - \int_0^T 1_{\{r(t)\geq x\}} \mathrm{d}t]\ \mathrm{d}x \\
&= \int_0^\infty C_u \int_0^T 1_{\{r(t)\geq x\}} \mathrm{d}t\ \mathrm{d}x - \int_0^N (C_u + C_o) \int_0^T 1_{\{r(t)\geq x\}} \mathrm{d}t\ \mathrm{d}x + C_o T N \\
&= \int_0^\infty C_u \int_0^T 1_{\{r_d(t)\geq x\}} \mathrm{d}t\ \mathrm{d}x - \int_0^N (C_u + C_o) \int_0^T 1_{\{r_d(t)\geq x\}} \mathrm{d}t\ \mathrm{d}x + C_o T N,
\end{aligned}$$

where the first equality is achieved by substituting:

$$(r(t) - N)^+ = \int_N^\infty 1_{\{r(t)\geq x\}} \mathrm{d}x, \qquad (N - r(t))^+ = \int_0^N 1_{\{r(t)\leq x\}} \mathrm{d}x, \tag{15}$$

and interchanging the order of integration.

We are now ready for Theorem 2.1, which identifies the optimal number of beds, N^*. The proof of the Theorem is provided in Appendix C. Note that our proof does not require that $r(t)$ and $\lambda(t)$ be continuous or differentiable. (These assumptions were needed in Jennings et al., 1997.)

Theorem 2.1. *The number of beds that minimizes C(N) is given by*

$$N^* = r_d \left(\frac{C_o T}{C_o + C_u} \right). \tag{16}$$

In Appendix D we explain how N^* arose as a candidate for minimizing $C(N)$.

Remark 2.3. *Alternatively, one can obtain the solution by building the cumulative relative frequency function for r and noting the similarity between our problem and the Newsvendor problem (Arrow et al., 1951; Nahmias and Cheng, 2009), for inventory management. In this case, we interpret the frequency as probability. This approach is similar to the reduction to the Newsvendor problem in Harrison and Zeevi (2005). However, our solution in (16) is more natural (more directly related to the time-varying nature of our models and their underlying systems); but, more importantly, this time-varying view naturally enables the solution of two extensions: setup cost per new bed (Section 2.8.1) and periodic reallocation of beds (Section 2.8.2) (such extensions are beyond the scope of the Newsvendor problem extension).*

2.7 Numerical Results

In this section, we apply our model to data in order to validate our solution (Sections 2.7.1 –2.7.2), calculate the imputed costs (Section 2.7.3) and provide structural insights and managerial recommendations (Section 2.7.4).

2.7.1 An Illustrative Example

Our healthcare partners were willing to share with us some of their financial reports and cost data. Rigorous calculations, based on these data (some of which are confidential), yielded the following critical fractiles required for (16). The hospitalization cost in mechanical ventilation wards is the highest among the geriatric wards and, as it turns out, $C_{u_3} = 1.882 C_{o_3}$. In rehabilitation wards the ratio is $C_{u_2} = 2.667 C_{o_2}$, as the hospitalization there is less expensive. Finally, the ratio for skilled nursing care is $C_{u_4} = 4.267 C_{o_4}$, as the hospitalization cost there is the lowest among the geriatric wards.

We used the fluid model developed in Section 2.4, together with our two-year historical data, to forecast the offered load for a subsequent three-year planning horizon,

where the demand for beds (e.g. the arrival rate) increases every year. Then, by using Matlab we numerically constructed the functions r_d for each ward (by sorting the function values of r). The optimal number of beds is the value of these functions at the critical point as in (16). Since the value of N^* is not necessarily an integer, it must be rounded. Rounding up vs. down has minor significance, since the solution here serves as a guide for a large organization that provides healthcare services for an entire district. Therefore, our solution provides insights regarding the difference between the suggested allocation and the current capacity.

The left plot in Figure 4 presents the optimal number of beds (the dashed lines) compared to the offered load (solid lines). The optimal number of beds for each ward was calculated by rounding up the result from Equation (16). The optimal solution implies increasing the current number of beds by 25%, 35% and 33% in rehabilitation, mechanical ventilation and skilled nursing care, respectively. In total, an increase to 577 beds from present 439 beds. This will lead to an overage and underage cost reduction of 51%, 53% and 69%; here, we compared to the cost under the current number of beds for the same arrival forecast. We believe that there are two major reasons for this dramatic cost reduction. The first is the lack of a model in practice, such as the one introduced here: such a model would take blocking and its related costs into account, which would guide planners. The second reason is the difficulties in increasing the present budget towards acquiring new beds. We provide more details and calculate imputed costs in Section 2.7.3.

The right plot in Figure 4 presents the waiting list length to each geriatric ward under the optimal number of beds. Note that the waiting lists were shortened (compared to the current situation presented in Figure 3), by 67%, 74% and 88% in rehabilitation, mechanical ventilation and skilled nursing care, respectively. This shortening occurred even though shortening the waiting lists was not directly included in our objective function. Indeed, we aimed at minimizing overage and underage costs; but since blocking costs are significant, reducing the total cost is achieved by reducing blocking which, in turn, leads to significant shorter waiting lists.

2.7.2 Solution Validation and Cost Comparison

In addition to validating our fluid model against data and stochastic simulation results (see Appendix A), in this section we validate our bed planning solution.

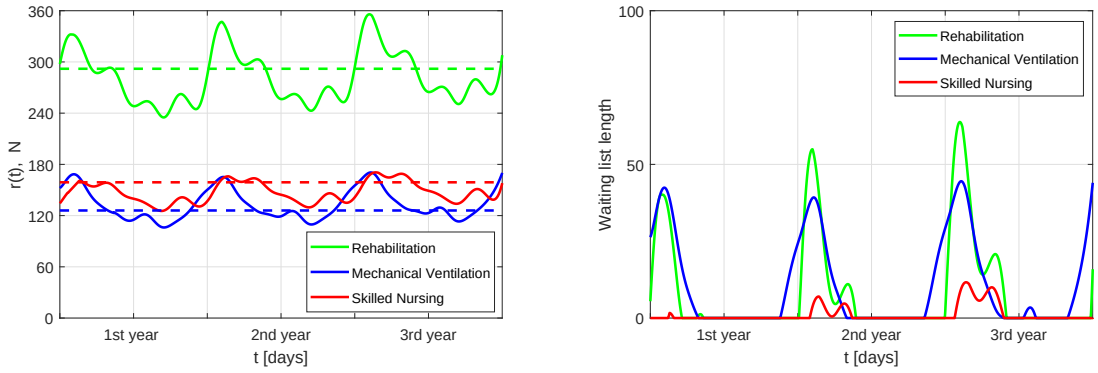


Figure 4: Optimal solution. On the left, the solid lines represent the offered load for each geriatric ward and the dashed lines represent the optimal number of beds. On the right, depicted are the waiting list lengths in hospital, according to the optimal solution; this is relative to the current waiting list lengths presented in Figure 3.

Thus far, two cost functions were presented for estimating the optimal number of geriatric beds. The first, $C^{(0)}(N_2, N_3, N_4)$ in (8), is based on the time-varying number of patients, as derived from the solution of the fluid equations in (7). Since minimizing $C^{(0)}(N_2, N_3, N_4)$ is analytically intractable, we introduced the second cost function, $C(N_2, N_3, N_4)$ in (11), which estimates the total cost based on an offered-load approximation to the time-varying demand for beds.

In order to validate the approximated cost function, we compared the optimal solutions for the two problems with the optimal solution derived from our stochastic simulation model. In the latter, the arrivals, duration times and routing percentages are random variables (see Appendix A). All parameters, including the size of the system, are realistic for the system we analyze.

The solution for $C(N_2, N_3, N_4)$ was calculated by our closed-form expression in (16). The solution for $C^{(0)}(N_2, N_3, N_4)$ was achieved by numerically solving the optimization problem in (8)–(9); this was done by solving the fluid model in (6)–(7) for each capacity combination, calculating the total cost according to (8) and choosing the capacity combination with the minimal cost. Finally, the solution for the stochastic simulation model was achieved by calculating, for each capacity combination, the total underage and overage cost. This was done by using (8) and (9), where instead of q_i and b_i, $i = 2, 3, 4$, we used the corresponding numbers from the simulation results. Then, we chose the combination which minimized the cost. In other words, the solutions according to $C^{(0)}(N_2, N_3, N_4)$ and simulation, was carried out by a three-dimensional

search (over N_2, N_3 and N_4). Table 4 summarizes this comparison by presenting the optimal number of beds and the optimal cost according to each method. In addition, we calculated the differences in percentages between each two methods for each ward separately and then all of them together. The last column in Table 4 presents the maximal difference between the solutions. The maximal difference varied from 1–1.6%, when comparing bed allocations and 1.1–3.4% when comparing total cost. This excellent fit is typical; indeed, we obtained similar differences when comparing the three solutions, under several other scenarios of overage and underage costs.

Ward	$C^{(0)}(N_2, N_3, N_4)$ N^* (Total cost)	$C(N_2, N_3, N_4)$ N^* (Total cost)	Simulation N^* (Total cost)	Maximal difference N^* (Total cost)
Rehabilitation	295 (2,601,667)	292 (2,683,042)	294 (2,633,167)	1.0% (3.0%)
Mechanical Ventilation	128 (1,493,917)	126 (1,547,000)	128 (1,499,167)	1.6% (3.4%)
Skilled Nursing	161 (1,213,333)	159 (1,226,750)	160 (1,215,667)	1.3% (1.1%)
Total Number of beds	584 (5,308,917)	577 (5,456,792)	582 (5,348,000)	1.2% (2.7%)

Table 4: Comparing optimal solutions (number of beds and overage and underage cost per year) – $C^{(0)}(N_2, N_3, N_4)$ vs. $C(N_2, N_3, N_4)$ vs. simulation.

2.7.3 The Imputed Overage and Underage Costs

In addition to the estimation of the C_o/C_u ratio given to us by our healthcare organization, it is of interest to examine C_o and C_u as imputed costs. These imputed costs are based on observed decisions that, in our case, are the number of beds that decision makers allocate to each geriatric ward. To this end, we use the current number of beds in each geriatric ward in order to extract the model's parameters C_o and C_u or, more accurately, the ratio C_o/C_u. (A similar approach was taken by Olivares et al., 2008.) Suppose that the current allocation N is optimal, we define

$$r_d^{-1}(N) \equiv \sup\{t | r_d(t) \geq N\}, \tag{17}$$

as the time during which underage costs were incurred. Let I denote the fraction of time during which underage costs were incurred. Consequently, from Theorem 2.1 we have

$$I = \frac{r_d^{-1}(N)}{T} = \frac{C_o}{C_o + C_u}, \tag{18}$$

We now present our data as a sequence of n days: $(t_i, r(t_i))$ for $i = 1, ..., n$, where

t_i denotes a single time point for day i. Then, we define $\bar{I}$ to be an estimator for the fraction of time during which underage costs were incurred:

$$\bar{I} = \frac{1}{n}\sum_{i=1}^{n} 1_{\{r(t_i)\geq N\}}. \tag{19}$$

We replace $r_d^{-1}(N)/T$ with $\bar{I}$ in (18) to get

$$\bar{I} = \frac{C_o}{C_o + C_u}. \tag{20}$$

According to our data, $\bar{I}_2 = 0.74$ in rehabilitation, $\bar{I}_3 = 0.91$ in skilled nursing care and $\bar{I}_4 = 1$ in mechanical ventilation. Therefore, the imputed costs are $C_{u_2} = 0.35C_o$ (vs. $C_{u_2} = 2.667C_o$ according to the financial reports) in rehabilitation, $C_u = 0.099C_o$ (vs. $C_{u_3} = 1.882C_o$) in skilled nursing care and $C_u = 0$ (vs. $C_{u_4} = 4.267$) in mechanical ventilation. The differences in the imputed costs among the three wards are due to different hospitalization costs, as explained in Section 2.7.1.

There is a big difference between the ratio C_u/C_o according to the financial reports, and according to the imputed costs. This may imply that blocking costs are neglected or underestimated when determining the geriatric bed capacity. Another possible explanation is that although there is a central decision maker that owns both the hospitals and geriatric institutions, decisions are locally optimized.

2.7.4 Managerial Insights for the Optimal Solution

The function r_d in the optimal solution (16) is decreasing in $[0, T]$. As explained already, the ratio $C_o/(C_o + C_u)$ in the optimal solution is the hospitalization cost ratio between a geriatric bed and a hospital bed. As the gap between these two costs widens, more geriatric beds will be needed. Indeed, in Figure 4, the optimal number of beds in skilled nursing care is relatively high compared to the offered load. The reason for this is the relatively low hospitalization cost in this ward. In mechanical ventilation, however, the optimal number of beds is relatively low compared to the offered load, since the hospitalization cost there is higher.

Figure 4 demonstrates long periods of overage, especially in skilled nursing care and rehabilitation. To accommodate for the seasonal demand, we seek a more flexible

solution, such as the possibility to reallocate beds between wards. To this end, we first sum the total offered load for the three wards then, we minimize (12) in order to find the total required number of beds. The optimal solution will then require fewer beds overall (566 beds instead of 577), but will lead to only an additional decrease of 5% in the total cost. The reason for this relatively modest advantage is the similar offered-load patterns among the wards, which implies that more beds are needed in all three wards at the same time. Thus, reallocating beds between wards is less effective in reducing the cost.

Consequently, a more flexible and responsive policy to fluctuations in demand, can be achieved by adding and removing beds throughout the year. Our healthcare partners argue that setting two capacity levels each year, which implies reallocating beds twice a year, is feasible. For example, it is possible to open a specific area/ward when demand is high (usually in the winter), and close this area when demand is low (usually in the summer). The described policy is feasible since most 'bed cost' is related to labor cost and medical supplies; the latter can be purchased seasonally while the former can be changed due to the existing flexibility of staffing levels (e.g. reallocating workers within facilities in the same organization or changing the work load of part-time workers throughout the year). We formally introduce and analyze the periodic reallocation problem in Section 2.8.2

2.8 Extensions

In this section we present two extensions to our model. The first extension, at the strategic level, adds setup costs for allocating new beds. The second extension, at the operational level, allows periodic reallocation of beds.

2.8.1 Including Setup Cost per New Bed

In this section, we analyze a case where there is a fixed setup cost, K, associated with the introduction of each new bed. The setup cost may be associated with recruitment and training of new staff or the purchase of new equipment. We assume that the setup cost may vary with bed types. Let B denote the current bed capacity, then the overall cost for a geriatric ward is

$$C_K(N) = C(N) + K(N - B)^+, \tag{21}$$

where $C(N)$ is the overall cost, analyzed in Section 2.5 and $(N-B)^+$ is the number of new beds. The planning horizon, T, reflects an organizational policy regarding investments and, hence, should be long enough for an investment in new beds to be worthwhile.

Theorem 2.2. *The optimal number of beds that minimizes $C_K(N)$ is given by*

$$N_k^* = \begin{cases} r_d\left(\dfrac{C_oT}{C_o+C_u}\right), & \text{if} \quad r_d\left(\dfrac{C_oT}{C_o+C_u}\right) \leq B \\ r_d\left(\dfrac{C_oT+K}{C_o+C_u}\right), & \text{if} \quad r_d\left(\dfrac{C_oT+K}{C_o+C_u}\right) \geq B \\ B, & \text{otherwise.} \end{cases} \tag{22}$$

We prove Theorem 2.2 in Appendix E.

Note that $r_d(\cdot)$ is defined on the interval $[0,T]$; hence, when $C_uT < K$, then $r_d(\cdot)$ is undefined, since

$$\frac{C_oT+K}{C_o+C_u} > \frac{C_oT+C_uT}{C_o+C_u} = T.$$

In this case, only the first condition of N_K^* is relevant. Therefore, the solution will not include the introduction of new beds. An intuitive explanation is that for a high bed setup cost it may be preferable to pay the underage cost for the entire planning horizon.

Note that the optimal solution depends on the available bed capacity. For a very large B, there is no point introducing new beds and, hence, the optimal solution equals the solution with no setup cost. On the other hand, if the current capacity, B, is very small, then adding new beds is essential for decreasing the total cost. In all other cases, it may be preferable to keep the capacity as is.

2.8.2 Periodic Reallocation of Beds

Managers of geriatric institutions acknowledge that it is feasible to change the number of beds during the year in order to compensate for seasonal variations in demand. Note that changing the number of beds also implies changing staff levels (which are typically proportional to the number of beds) and other related costs. The planning horizon

remains the same, but we divide each year into several periods. We then determine the preferable periods (location and length) and the number of beds required for each period. For example, an optimal reallocation policy would determine a certain capacity during the first three and the last two months of every year in the planning horizon, and possibly a different capacity during the seven other months of every year. To this end, we introduce a reallocation cost, C_r, associated with adding and removing a bed.

Due to feasibility constraints from our partner hospital chain, we allow only two capacity levels throughout the planning horizon. Nevertheless, the methodology we present can be implemented in other settings where more capacity levels are possible. Moreover, due to the nature/shape of the demand, having two capacity levels corresponds to changing capacity levels twice each year.

Let $\mathcal{T} = [0, T]$ denote the planning horizon interval and let $\mathcal{I}$ denote the time interval (location and length) in which there are $N_{\mathcal{I}}$ geriatric beds (in $\mathcal{T} \setminus \mathcal{I}$, there are $N_{\mathcal{T}\setminus\mathcal{I}}$ geriatric beds). Our objective is to find $\mathcal{I}$, $N_{\mathcal{I}}$ and $N_{\mathcal{T}\setminus\mathcal{I}}$ that minimize the total underage and overage costs.

To this end, we split $r(t)$ into two functions: $r_{\mathcal{I}}(t)$ for the capacity level in $\mathcal{I}$ and $r_{\mathcal{T}\setminus\mathcal{I}}(t)$ for the capacity level in $\mathcal{T} \setminus \mathcal{I}$. The functions $r_{\mathcal{I}}(t)$ and $r_{\mathcal{T}\setminus\mathcal{I}}(t)$ are defined on the intervals $[0, |\mathcal{I}|]$ and $[0, |\mathcal{T} \setminus \mathcal{I}|]$, respectively, by concatenating the relevant intervals from $r(t)$ and shifting the functions to $t = 0$. We define the functions $r_{d_{\mathcal{I}}}(t)$ and $r_{d_{\mathcal{T}\setminus\mathcal{I}}}(t)$ to be the decreasing rearrangements of $r_{\mathcal{I}}(t)$ and $r_{\mathcal{T}\setminus\mathcal{I}}(t)$, respectively, exactly as we defined $r_d(t)$ in Section 2.5. The total underage and overage costs are, therefore,

$$\begin{aligned} C(\mathcal{I}, N_{\mathcal{I}}, N_{\mathcal{T}\setminus\mathcal{I}}) &= C(\mathcal{I}, N_{\mathcal{I}}) + C(\mathcal{T} \setminus \mathcal{I}, N_{\mathcal{T}\setminus\mathcal{I}}) + C_r \left| N_{\mathcal{T}\setminus\mathcal{I}} - N_{\mathcal{I}} \right| \\ &= \int_{\mathcal{I}} \left[C_u \big(r(t) - N_{\mathcal{I}}\big)^+ + C_o \big(N_{\mathcal{I}} - r(t)\big)^+ \right] \mathrm{d}t \\ &+ \int_{\mathcal{T}\setminus\mathcal{I}} \left[C_u \big(r(t) - N_{\mathcal{T}\setminus\mathcal{I}}\big)^+ + C_o \big(N_{\mathcal{T}\setminus\mathcal{I}} - r(t)\big)^+ \right] \mathrm{d}t + C_r \left| N_{\mathcal{T}\setminus\mathcal{I}} - N_{\mathcal{I}} \right|, \end{aligned} \tag{23}$$

where $C(\mathcal{I}, N_{\mathcal{I}})$ and $C(\mathcal{T}\setminus\mathcal{I}, N_{\mathcal{T}\setminus\mathcal{I}})$ denote the overage and underage costs for intervals $\mathcal{I}$ and $\mathcal{T} \setminus \mathcal{I}$, respectively.

Theorem 2.3. *The number of beds that minimizes (23), for a fixed $\mathcal{I}$, is*

$$\begin{cases} N^*_{\mathcal{I}} = N^{\mathcal{I}}_{-}, \;\; N^*_{\mathcal{T}\setminus\mathcal{I}} = N^{\mathcal{T}\setminus\mathcal{I}}_{+}, & \text{if} \;\; N^{\mathcal{I}}_{-} \le N^{\mathcal{T}\setminus\mathcal{I}}_{+}, \\ N^*_{\mathcal{I}} = N^{\mathcal{I}}_{+}, \;\; N^*_{\mathcal{T}\setminus\mathcal{I}} = N^{\mathcal{T}\setminus\mathcal{I}}_{-}, & \text{if} \;\; N^{\mathcal{I}}_{+} \ge N^{\mathcal{T}\setminus\mathcal{I}}_{-}, \\ N^*_{\mathcal{I}} = N^*_{\mathcal{T}\setminus\mathcal{I}} = N^*, \text{ as in (16)}, & \text{otherwise.} \end{cases} \tag{24}$$

Here, $N^{\mathcal{A}}_{\pm} = r_{d_{\mathcal{A}}}\left(\frac{C_o|\mathcal{A}|\pm C_r}{C_o+C_u}\right)$, for every interval $\mathcal{A}$.

We prove Theorem 2.3 in Appendix F.

Note that the option in the third line in (24) suggests determining only one capacity level (e.g. it is preferable not to reallocate beds throughout the planning horizon). In particular, since $r_{d_{\mathcal{I}}}(\cdot)$ and $r_{d_{\mathcal{T}\setminus\mathcal{I}}}(\cdot)$ are defined on the intervals $[0, |\mathcal{I}|]$ and $[0, |\mathcal{T}\setminus\mathcal{I}|]$, respectively, when $C_u|\mathcal{I}| > C_r$ or when $C_u|\mathcal{T}\setminus\mathcal{I}| > C_r$, it is preferable to pay the underage cost for the entire period than to pay the reallocation cost, C_r.

2.8.3 A Numerical Example

We now solve the periodic reallocation problem for a three-year planning horizon. Figure 5 depicts the solutions for three cases. The solid lines represent the offered load for each ward, while the dashed lines represent the optimal number of beds. The first case (top left plot) is when no reallocation costs are introduced ($C_r = 0$). This solution yields a 35%, 22% and 31% underage and overage cost reduction, in rehabilitation, mechanical ventilation and skilled nursing care, respectively, compared to the constant allocation. The second case (top right plot) is when reallocation costs are introduced; in this case, the gaps between the two capacity levels narrows. In particular, the optimal allocation in mechanical ventilation is constant, since it is not worthy to invest the reallocation cost (e.g. $C_r > C_u|\mathcal{I}|$ or $C_r > C_u|\mathcal{T}\setminus\mathcal{I}|$). The third case (bottom right plot), presents the optimal periodic reallocation when four reallocation points are allowed and no reallocation costs are introduced. The left bottom plot in Figure 5 presents the waiting list lengths for each ward under the optimal reallocation policy when no reallocation costs are introduced; this is in comparison with the current situation presented in Figure 3 and the constant allocation presented in Figure 4 (right).

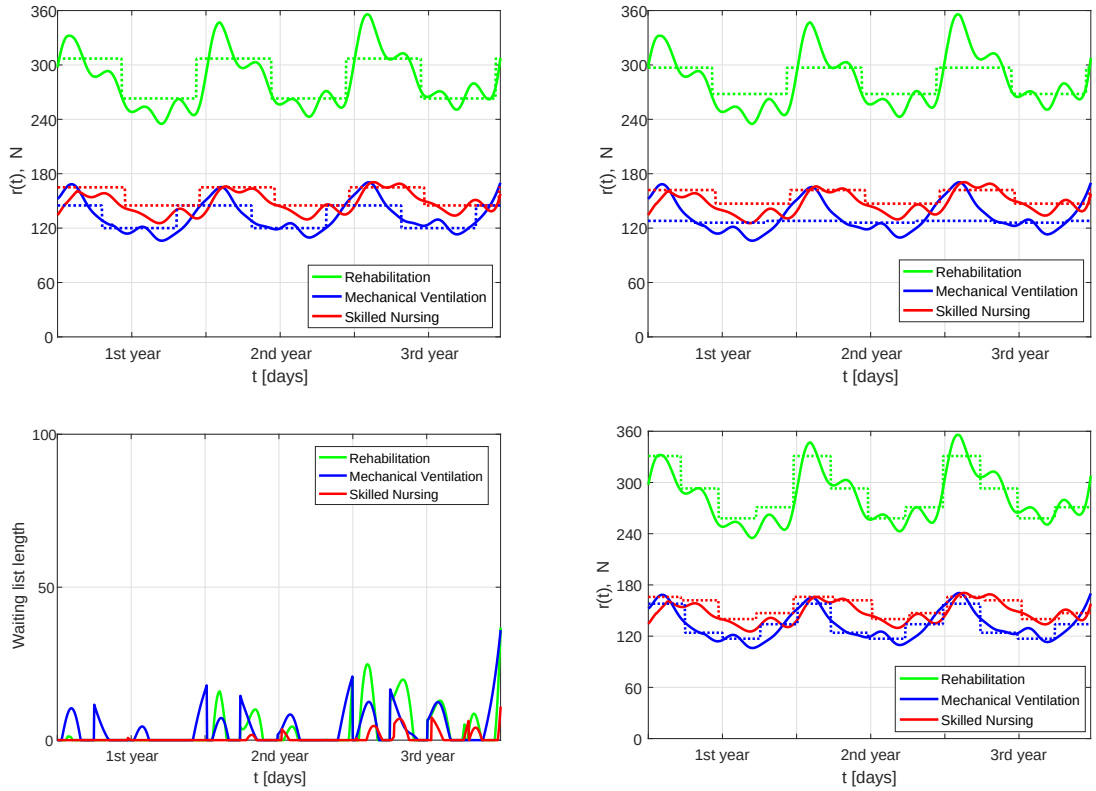


Figure 5: Optimal reallocation of beds when no reallocation costs are introduced (left top plot), when reallocation costs are introduced (right top plot) and when four reallocation points are allowed (bottom right plot). Waiting list length under the optimal reallocation policy when no reallocation costs are introduced (left bottom plot).

2.8.4 Managerial Recommendations on Extensions

The major cost reduction, compared to the current situation for the three wards, is achieved by adopting the proposed policy of a constant number of beds. Periodic allocations allow for extra cost reductions, when compared to the policy with a constant number of beds. Thus, a reasonable policy would be to adopt the constant allocation at a first step and implement the periodic reallocation as a second step. In some cases, when the reallocation cost is higher than the underage period cost, it is preferable to remain with the constant allocation (see the case for mechanical ventilation ward in the right top plot of Figure 5). Another option which can help reduce the load is to divert more geriatric patients in peak periods to home healthcare services or virtual hospitals rather than to geriatric institutions (Ticona and Schulman, 2016). In this case, multidisciplinary home healthcare teams treat the patient at home rather than in hospital. Home care hospitalization was found to be more effective, shorter and

increases patient satisfaction, compared to the same treatment received in hospital (Shepperd et al., 2008; Caplan et al., 2012). Moreover, according to our analysis, even a 10% diversion of patients requiring geriatric hospitalization to home care, will reduce the overage, and underage costs by about 25% on average and will shorten the waiting lists in hospital by 30% on average.

2.9 Future Research

There are multiple directions worthy of future research, two of which will be now described. The first is to modify the structure of the system by adding an intermediate ward (i.e., a step-down unit) for sub-acute geriatrics (Wolstenholme, 1999), between the hospital and the geriatric institutions. Such an intermediate ward would be designated for elderly patients with an expected long stay in the hospital, before continuing on to a geriatric ward. Adding a sub-acute ward can both reduce the workload and bed occupancy in hospitals and improve the patient flow in and out of the hospital.

Another direction is a capacity allocation problem, in which given a predefined budget, the planners must decide where it is most beneficial to add new beds: in hospitals, in intermediate wards or in geriatric wards. The simple version of this question (without intermediate wards), in fact, triggered the present research.

3 Time-varying Tandem Queues under the BAS Mechanism

3.1 Introduction

Achieving a more comprehensive view of the system analyzed in Section 2, can be done by including ED boarded patients, waiting for admission to hospital wards (Figure 1). This analysis should also include finite waiting room before the first station and customer loss when this waiting room is full. This has motivated us to model and analyze time-varying tandem networks with blocking and finite waiting rooms throughout the network – before the first station and between the stations.

The models we focus on (flow lines) have been researched for decades (Avi-Itzhak, 1965; Avi-Itzhak and Levy, 1995; Li and Meerkov, 2009; Meerkov and Yan, 2016); our research takes the analysis to the new territories of time-varying environments and many-server stations.

In particular, 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 (Mandelbaum et al., 1998, 1999; Whitt, 2004, 2006; Pang and Whitt, 2009; Liu and Whitt, 2011a, 2014).

Our basic model (Section 3.4) is a network with two queues in tandem (Figure 6), 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*

(Buzacott and Shanthikumar, 1993; Balsamo et al., 2001); and here, as it turns out (Section 2.4), 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 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) (Filippov, 2013); 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 Section 3.5 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 Section 3.6 we provide operational insights regarding the performance of time-varying tandem queues with finite buffers. 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 on 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.

3.2 Literature Review

Despite the fact that time-varying parameters are common in production (Leachman and Gascon, 1988; Nahmias and Cheng, 2009) and service systems (Green et al., 2007b; Feldman et al., 2008), such as in healthcare (Armony et al., 2015; Cohen et al., 2014;

Yom-Tov and Mandelbaum, 2014), research on time-varying models with blocking is scarce. We now review the three research areas, most relevant to this work.

3.2.1 Flow Lines with Blocking

Previous research on tandem queueing networks with blocking has focused on steady-state analysis for small networks (Grassmann and Drekic, 2000; Akyildiz and von Brand, 1994; Langaris and Conolly, 1984), steady-state approximations for larger networks (Takahashi et al., 1980; Brandwajn and Jow, 1988; Gershwin, 1987; Dallery and Gershwin, 1992; Perros, 1994; Balsamo and de Nitto Personè, 1994; Tolio and Gershwin, 1998; van Vuuren et al., 2005; Osorio and Bierlaire, 2009) and simulation models (Conway et al., 1988; El-Darzi et al., 1998; Katsaliaki et al., 2005; Bretthauer et al., 2011; Millhiser and Burnetas, 2013).

Several papers have analyzed tandem queueing networks with an unlimited waiting room before the first station and a Blocking After Service (BAS) mechanism between the stations. In Avi-Itzhak and Yadin (1965), the steady-state of a 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 Prabhu (1967). The model in Avi-Itzhak and Yadin (1965) was extended in Avi-Itzhak (1965) 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 Avi-Itzhak (1965) 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 Avi-Itzhak and Yadin (1965) was generalized in Avi-Itzhak and Levy (1995) 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 Kelly (1984); the service times for each customer are identical at each station. In Whitt (1985) 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 Conway et al. (1988) to analyze Work in Process (WIP) in serial production lines, with and without buffers in balanced and unbalanced lines. The results of Glynn and Whitt (1991) were extended in Martin (2002) 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 Meerkov and Yan (2016). 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.

3.2.2 Time-Varying Fluid Models

Fluid models were successfully implemented in modeling different types of service systems. These models cover the early applications for post offices (Oliver and Samuel, 1962), claims processing in social security offices (Vandergraft, 1983), call centers (Green et al., 2007b; Afèche et al., 2017) and healthcare systems (Yom-Tov and Mandelbaum, 2014; Cohen et al., 2014; Zychlinski et al., 2018c). Fluid models of service systems were extended to include state-dependent arrival rates, general arrival and service rates (Whitt, 2005, 2006). Time-varying queueing models were analyzed for setting staffing requirements in service systems with unlimited waiting rooms, by using the offered load heuristics (Green et al., 2007b; Whitt, 2007, 2013).

Time-varying heavy traffic fluid limits were developed in Mandelbaum et al. (1998, 1999) 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 Liu and Whitt (2011a) for a single station, and for a network in Liu and Whitt (2011b). These models were extended to general service times in Liu and Whitt (2012a,b, 2014).

Heavy traffic approximations for systems with blocking have focused on stationary loss models (Borisov and Borovkov, 1981; Borovkov, 2012; Srikant and Whitt, 1996). An approximation for the steady-state blocking probability, with service times being dependent and non-exponential, was developed in Li and Whitt (2014). A recent work

in Li et al. (2016) 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.

3.2.3 Queueing Models with Reflection

Queueing models with reflection were analyzed in Harrison (1973) 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 Wenocur (1982) 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 Harrison (1985); Chen and Yao (2013) for a single-station and for a network is extended in Mandelbaum and Pats (1995, 1998) for state-dependent queues. Loss systems for one station with reflection were analyzed in Whitt (2002); Garnett et al. (2002). More recently, Reed et al. (2013) 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.

3.3 Contributions

The main contributions of this section 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 (Filippov, 2013).

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 (e.g. (28)–(29), for $k = 2$). It turns out, however, that a state description in terms of non-utilized servers is more amenable to analysis ((31)–(32)). Indeed, it enables a representation of the network in terms of reflection, which yields useful properties of the network reflection operator (e.g. Lipschitz continuity).

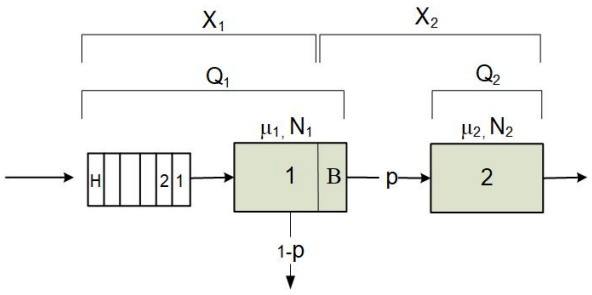


Figure 6: Two tandem stations with a finite waiting room before the first station.

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 (Section 3.6), 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.

3.4 Two Stations in Tandem with Finite Waiting Room

We now develop a fluid model with blocking for two stations in tandem, as illustrated in Figure 6. In Section 3.5, we further extend this model for a network with k stations in tandem and finite internal waiting rooms between the stations.

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

$$\mathbb{E}A(t) = \int_0^t \lambda(u)\,\mathrm{d}u, \quad t \geq 0. \tag{25}$$

   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. 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. 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 (waiting room 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) (Balsamo et al., 2001). 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 Zychlinski et al. (2018c) 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

$$\begin{aligned} X_1(t) = X_1(0) &+ \int_0^t 1_{\{X_1(u-)+(X_2(u-)-N_2)^+ < N_1+H\}} \, \mathrm{d}A(u) \\ &- D_1\left(p\mu_1 \int_0^t [X_1(u) \wedge (N_1 - B(u))] \, \mathrm{d}u\right) \\ &- D_3\left((1-p)\mu_1 \int_0^t [X_1(u) \wedge (N_1 - B(u))] \, \mathrm{d}u\right), \\ X_2(t) = X_2(0) &+ D_1\left(p\mu_1 \int_0^t [X_1(u) \wedge (N_1 - B(u))] \, \mathrm{d}u\right) \\ &- D_2\left(\mu_2 \int_0^t [X_2(u) \wedge N_2] \, \mathrm{d}u\right); \quad t \geq 0. \end{aligned} \tag{26}$$

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 (26) represents the number of arrivals that entered service up to time t. As noted in Mandelbaum and Pats (1998), an inductive construction over time shows that (26) 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.

3.4.1 Representation in Terms of Reflection

First we rewrite (26) by using the fact that

$$\begin{aligned} &\int_0^t 1_{\{X_1(u-)+(X_2(u-)-N_2)^+ < N_1+H\}} \, \mathrm{d}A(u) \\ &\qquad = A(t) - \int_0^t 1_{\{X_1(u-)+(X_2(u-)-N_2)^+ = N_1+H\}} \, \mathrm{d}A(u); \end{aligned} \tag{27}$$

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 (26) and (27):

$$\begin{cases} \begin{bmatrix} X_1(t) \\ X_1(t)+X_2(t) \end{bmatrix} = \begin{bmatrix} Y_1(t)-L(t) \\ Y_2(t)-L(t) \end{bmatrix} \le \begin{bmatrix} N_1+H \\ N_1+N_2+H \end{bmatrix}, \quad t \ge 0, \\ \mathrm{d}L(t) \ge 0, \;\; L(0)=0, \\ \int_0^\infty 1_{\{X_1(t)+(X_2(t)-N_2)^+ < N_1+H\}} \, \mathrm{d}L(t) = 0, \end{cases} \tag{28}$$

where

$$\begin{aligned} Y_1(t) &= X_1(0) + A(t) - D_1\left(p\mu_1 \int_0^t [X_1(u) \wedge (N_1 - B(u))] \, \mathrm{d}u\right) \\ &\quad - D_3\left((1-p)\mu_1 \int_0^t [X_1(u) \wedge (N_1 - B(u))] \, \mathrm{d}u\right), \\ Y_2(t) &= X_1(0) + X_2(0) + A(t) - D_3\left((1-p)\mu_1 \int_0^t [X_1(u) \wedge (N_1 - B(u))] \, \mathrm{d}u\right) \\ &\quad - D_2\left(\mu_2 \int_0^t [X_2(u) \wedge N_2] \, \mathrm{d}u\right), \\ L(t) &= \int_0^t 1_{\{X_1(u-)+(X_2(u-)-N_2)^+ = N_1+H\}} \, \mathrm{d}A(u). \end{aligned} \tag{29}$$

Figure 7 (left) geometrically illustrates the reflection in (28). 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 (28) 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 (29) in the last equation for $L(t)$ of (28), which yields the following:

$$\int_0^\infty 1_{\{X_1(t)+(X_2(t)-N_2)^+ < N_1+H\}} \cdot 1_{\{X_1(t-)+(X_2(t-)-N_2)^+ = N_1+H\}} \, \mathrm{d}A(t) = 0. \tag{30}$$

Now, if (30) 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

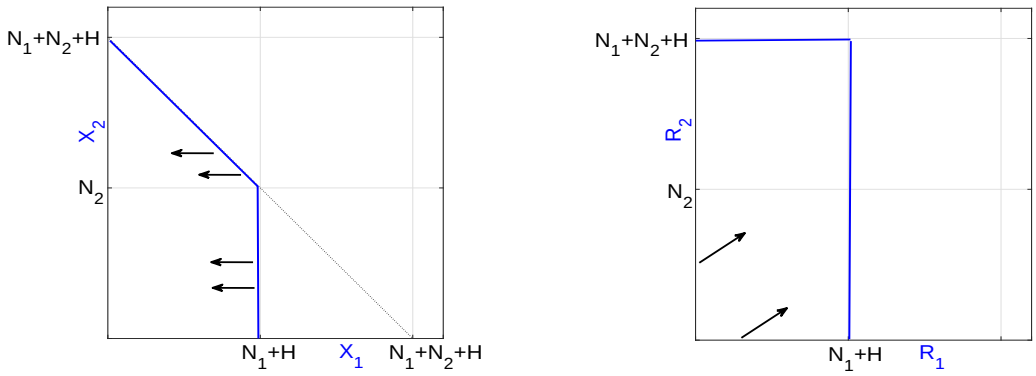


Figure 7: Geometrical representation of the reflection. On the left – in terms of X, and on the right – in terms of R.

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 (28), 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 (28):

$$\begin{cases} \begin{bmatrix} R_1(t) \\ R_2(t) \end{bmatrix} = \begin{bmatrix} \tilde{Y}_1(t) + L(t) \\ \tilde{Y}_2(t) + L(t) \end{bmatrix} \geq 0, \quad t \geq 0, \\ \mathrm{d}L(t) \geq 0, \;\; L(0) = 0, \\ \int_0^\infty 1_{\{R_1(t) \wedge R_2(t) > 0\}} \, \mathrm{d}L(t) = 0, \end{cases} \tag{31}$$

where

$$\tilde{Y}(t) = \begin{bmatrix} \tilde{Y}_1(t) \\ \tilde{Y}_2(t) \end{bmatrix} = \begin{bmatrix} N_1 + H - Y_1(t) \\ N_1 + N_2 + H - Y_2(t) \end{bmatrix}; \tag{32}$$

the last line in (31) is derived from

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

The processes $\tilde{Y}_1$, $\tilde{Y}_2$ and L (see (31)) can be stated in the "language" of R:

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

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

Figure 7 (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 Figure 7.

From (31), 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

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

Note that this solution is applicable even though $\tilde{Y}$ depends on R (see Mandelbaum and Pats, 1995 for details, though recall that they do not cover blocking).

3.4.2 Fluid Approximation

We now develop a fluid limit for our queueing model through the Functional Strong Law of Large Numbers (FSLLN). We begin with (31) 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's representation:

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

where

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

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

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

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

$$\begin{bmatrix} \tilde{y}_1^\eta(\cdot) \\ \tilde{y}_2^\eta(\cdot) \end{bmatrix} = \begin{bmatrix} r_1^\eta(0) - \eta^{-1}A^\eta(\cdot) + \eta^{-1}D_1\left(\eta p\mu_1 \int_0^\cdot [(N_1 + H - r_1^\eta(u)) \wedge (N_1 - b^\eta(u))]\, \mathrm{d}u\right) \\ \quad +\eta^{-1}D_3\left(\eta(1-p)\mu_1 \int_0^\cdot [(N_1 + H - r_1^\eta(u)) \wedge (N_1 - b^\eta(u))]\, \mathrm{d}u\right) \\ r_2^\eta(0) - \eta^{-1}A^\eta(\cdot) + \eta^{-1}D_3\left(\eta(1-p)\mu_1 \int_0^\cdot [(N_1 + H - r_1^\eta(u)) \wedge (N_1 - b^\eta(u))]\, \mathrm{d}u\right) \\ \quad +\eta^{-1}D_2\left(\eta\mu_2 \int_0^\cdot [N_2 \wedge (r_1^\eta(u) - r_2^\eta(u) + N_2)]\, \mathrm{d}u\right) \end{bmatrix}. \tag{33}$$

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

Theorem 3.1. *Suppose that*

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

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

$$\begin{cases} r_1(t) = r_1(0) - \int_0^t [\lambda(u) - \mu_1((N_1 + H - r_1(u)) \wedge (N_1 - b(u)))]\, \mathrm{d}u + l(t) \geq 0, \\ r_2(t) = r_2(0) - \int_0^t [\lambda(u) - (1-p)\mu_1((N_1 + H - r_1(u)) \wedge (N_1 - b(u)))]\, \mathrm{d}u \\ \qquad\qquad + \int_0^t [\mu_2(N_2 \wedge (r_1(u) - r_2(u) + N_2))]\, \mathrm{d}u + l(t) \geq 0, \\ \mathrm{d}l(t) \geq 0,\ \ l(0) = 0, \\ \int_0^\infty 1_{\{r_1(t) \wedge r_2(t) > 0\}}\, \mathrm{d}l(t) = 0; \end{cases} \tag{34}$$

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

Returning to our original formulation (28), (34) can in fact be written in terms of

$x(\cdot)$ for $t \geq 0$ as follows:

$$\begin{cases} x_1(t) = x_1(0) + \int_0^t \left[\lambda(u) - \mu_1(x_1(u) \wedge \left(N_1 - b(u)\right))\right] \mathrm{d}u - l(t) \leq N_1 + H, \\ x_1(t) + x_2(t) = x_1(t) + x_2(0) + \int_0^t \left[p\mu_1(x_1(u) \wedge (N_1 - b(u))) - \mu_2(N_2 \wedge x_2(u))\right] \mathrm{d}u \\ \qquad\qquad \leq N_1 + N_2 + H, \\ \mathrm{d}l(t) \geq 0, \;\; l(0) = 0, \\ \int_0^\infty 1_{\{x_1(t) + (x_2(t) - N_2)^+ < N_1 + H\}} \, \mathrm{d}l(t) = 0. \end{cases} \tag{35}$$

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 (35); see Appendix I for details. As opposed to (35), this solution (36) is given by a set of differential equations with discontinuous RHS but without reflection. Thus, implementing (36) numerically is straightforward via recursion, which would not be the case with (35).

Proposition 3.1. *The fluid limit approximation for X in (26) is given by*

$$\begin{aligned} x_1(t) = x_1(0) &- \mu_1 \int_0^t [x_1(u) \wedge (N_1 - b(u))] \, \mathrm{d}u \\ &+ \int_0^t \left[1_{\{x_1(u) < N_1 + H\}} \cdot 1_{\{x_1(u) + x_2(u) < N_1 + N_2 + H\}} \cdot \lambda(u)\right] \mathrm{d}u \\ &+ \int_0^t \left[1_{\{x_1(u) = N_1 + H\}} \cdot 1_{\{x_1(u) + x_2(u) < N_1 + N_2 + H\}} \cdot [\lambda(u) \wedge l_1^*(u)]\right] \mathrm{d}u \\ &+ \int_0^t \left[1_{\{x_1(u) < N_1 + H\}} \cdot 1_{\{x_1(u) + x_2(u) = N_1 + N_2 + H\}} \cdot [\lambda(u) \wedge l_2^*(u)]\right] \mathrm{d}u \\ &+ \int_0^t \left[1_{\{x_1(u) = N_1 + H\}} \cdot 1_{\{x_1(u) + x_2(u) = N_1 + N_2 + H\}} \cdot [\lambda(u) \wedge l_1^*(u) \wedge l_2^*(u)]\right] \mathrm{d}u, \\ x_2(t) = x_2(0) &+ \int_0^t [p\mu_1(x_1(u) \wedge (N_1 - b(u))) - \mu_2(x_2(u) \wedge N_2)] \, \mathrm{d}u, \end{aligned} \tag{36}$$

where

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

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:

$$q_1(t) = x_1(t) + b(t);$$
$$q_2(t) = x_2 \wedge N_2.$$

Remark 3.1. *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 3.2. *Abandonments from the waiting room can occur when customers have finite patience. This is a prevalent phenomenon in service systems and healthcare, in particular (e.g. customers that abandon the Emergency Department are categorized as Left Without Being Seen (LWBS) (Baker et al., 1991; Arendt et al., 2003). Such abandonments can be added to our model by following Mandelbaum et al. (1999) and Pender (2015). 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 (36); here $[x_1(t) + b(t) - N_1]^+$ represents the number of waiting customers at Station 1 at time t.*

3.4.3 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 (He et al., 2016; Ma and Whitt, 2016). In Liu and Whitt (2012a), the authors introduce an algorithm that is based on the standard equilibrium renewal process (SERP). This algorithm is implemented in Pen-

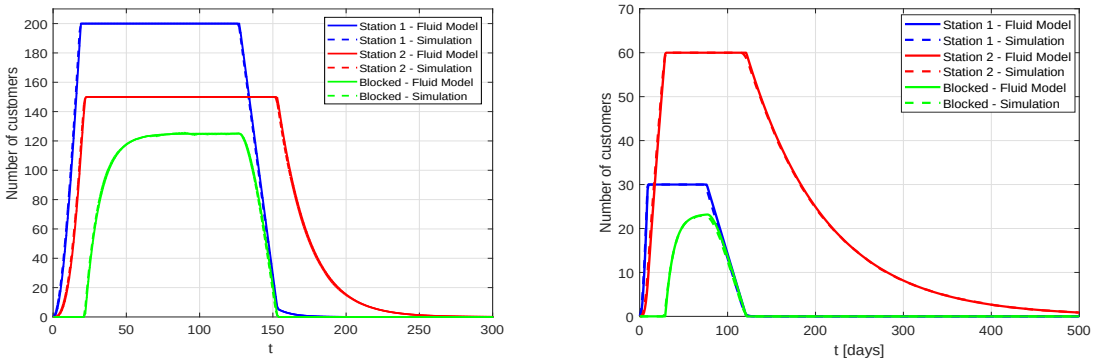


Figure 8: Total number in each station – fluid formulation vs. simulation for two scenarios. The fluid model curves overlap the simulation curves.

der and Ko (2017) to approximate the general inter-arrival times for the phase-type distribution.

The fluid equations in (36) were solved recursively, by discretizing time. Figure 8 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 $\lambda(t) = 2t$, $0 \leq t \leq 120$. 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 $\lambda(t) = t$, $0 \leq t \leq 60$.
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.

3.5 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 Figure 9. The notations remain as before, only with an i subscript, $i = 1, \ldots, k$, indicating Station i. Moreover, we denote the transfer 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

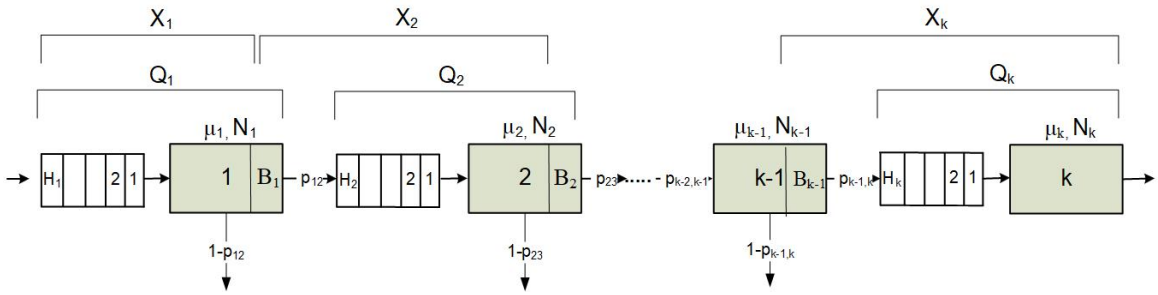


Figure 9: Multiple stations in tandem with finite internal waiting rooms.

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: External arrival process $A = \{A(t), t \geq 0\}$, as was defined in (25), processes $D_i = \{D_i(t), t \geq 0\}$, $i = 1, .., 2k-1$, 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 $Q = \{Q_1(t), Q_2(t), .., Q_k(t), t \geq 0\}$ 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:

$$\begin{aligned} Q_1(t) =& X_1(t) + B_1(t); \\ Q_i(t) =& \big[X_i(t) + B_i(t)\big] \wedge (N_i + H_i), \quad i = 2, \ldots, k-1; \\ Q_k(t) =& X_k(t) \wedge (N_k + H_k), \qquad\qquad t \geq 0. \end{aligned} \tag{37}$$

Here,

$$X_1(t) = X_1(0) + A(t) - D_1\left(p_{12} \cdot \mu_1 \int_0^t \left[X_1(u) \wedge (N_1 - B_1(u))\right] \mathrm{d}u\right) \tag{38}$$

$$
\begin{aligned}
&\quad - D_{k+1}\left((1-p_{12})\cdot\mu_1\int_0^t \left[X_1(u)\wedge\left(N_1-B_1(u)\right)\right]\mathrm{d}u\right)\\
&\quad - \int_0^t 1_{\{X_1(u-)+B_1(u-)=N_1+H_1\}}\,\mathrm{d}A(u),\\
X_i(t) &= X_i(0) + D_{i-1}\left(p_{i-1,i}\cdot\mu_{i-1}\int_0^t \left[X_{i-1}(u)\wedge(N_{i-1}-B_{i-1}(u))\right]\,\mathrm{du}\right)\\
&\quad - D_i\left(p_{i,i+1}\cdot\mu_i\int_0^t \left[X_i(u)\wedge(N_i-B_i(u))\right]\,\mathrm{du}\right)\\
&\quad - D_{k+i}\left((1-p_{i,i+1})\cdot\mu_i\int_0^t \left[X_i(u)\wedge(N_i-B_i(u))\right]\,\mathrm{du}\right), i=2,\ldots,k-1,\\
X_k(t) &= X_k(0) + D_{k-1}\left(p_{k-1,k}\cdot\mu_{k-1}\int_0^t \left[X_{k-1}(u)\wedge(N_{k-1}-B_{k-1}(u))\right]\,\mathrm{du}\right)\\
&\quad - D_k\left(\mu_k\int_0^t \left[X_k(u)\wedge N_k\right]du\right),\\
B_i(t) &= \left[X_{i+1}(t)+B_{i+1}(t)-N_{i+1}-H_{i+1}\right]^+,\ \ i=1,\ldots,k-2,\\
B_{k-1}(t) &= \left[X_k(t)-N_k-H_k\right]^+.
\end{aligned}
$$

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 (38) uniquely determines the processes X and B.

By using similar methods as for the two-station network in Section 3.4, with more cumbersome algebra and notations, we establish that x, the fluid limit for the stochastic queueing family X^η, is given, for $t\geq 0$, by

$$
\begin{aligned}
x_1(t) &= x_1(0) - \mu_1\int_0^t \left[x_1(u)\wedge(N_1-b_1(u))\right]\,\mathrm{du} \qquad (39)\\
&\quad + \sum_{m=0}^{k}\sum_{\substack{A\subset\{1,\ldots,k\}:\\ |A|=m}}\int_0^t\left[\prod_{j\in A} 1_{\left\{\sum_{i=1}^j x_i(u)=\sum_{i=1}^j(N_i+H_i)\right\}}\right.\\
&\quad \left.\times\prod_{j\in\{1,\ldots,k\}\cap\bar{A}} 1_{\left\{\sum_{i=1}^j x_i(u)<\sum_{i=1}^j(N_i+H_i)\right\}}\left[\lambda(u)\wedge\bigwedge_{y\in A} l_y^*(u)\right]\right]\mathrm{du},\\
x_i(t) &= x_i(0) + \int_0^t\left[p_{i-1,i}\cdot\mu_{i-1}\left(x_{i-1}(u)\wedge(N_{i-1}-b_{i-1}(u))\right)\right.\\
&\quad \left. - \mu_i\left(x_i(u)\wedge(N_i-b_i(u))\right)\right]\mathrm{du}, \qquad i=2,\ldots,k-1,\\
x_k(t) &= x_k(0) + \int_0^t\left[p_{k-1,k}\cdot\mu_{k-1}\left(x_{k-1}(u)\wedge(N_{k-1}-b_{k-1}(u))\right) - \mu_k\left(x_k(u)\wedge N_k\right)\right]\mathrm{du},
\end{aligned}
$$

where

$$
\begin{aligned}
l_1^*(u) &= \mu_1 N_1, \\
l_n^*(u) &= \mu_n N_n + \sum_{j=1}^{n-1} (1 - p_{j,j+1}) \mu_j \left(x_j(u) \wedge (N_j - b_j(u)) \right), \quad n = 2, \ldots, k, \\
b_i(t) &= \left[x_{i+1}(t) + b_{i+1}(t) - N_{i+1} - H_{i+1} \right]^+, \quad i = 1, \ldots k-2, \\
b_{k-1}(t) &= \left[x_k(t) - N_k - H_k \right]^+ .
\end{aligned}
$$

The term in the second line of (39) is a generalization of the last 4 terms in the expression for $x_1(t)$ in (36), 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

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

Remark 3.3. *A special case for the model analyzed in Section 3.5 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 Zychlinski et al. (2018c).*

3.6 Numerical Experiments and Operational Insights

In this section, we demonstrate how our models yield operational insights on 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 Sections 3.6.1–3.6.2, 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 ($H = \infty$). Sections 3.6.3–3.6.4 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.

3.6.1 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. stations). This implies that the stations are identical in the fluid model; in Section 3.6.2 we relax this assumption.

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

$$\lambda(t) = \bar{\lambda} + \beta \sin(\gamma t), \quad t \geq 0, \tag{40}$$

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

Figure 10 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 of 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 11 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 (Figure 10) also occurs when stations have ample capacities to eliminate blocking and customer loss. In these cases, system performance

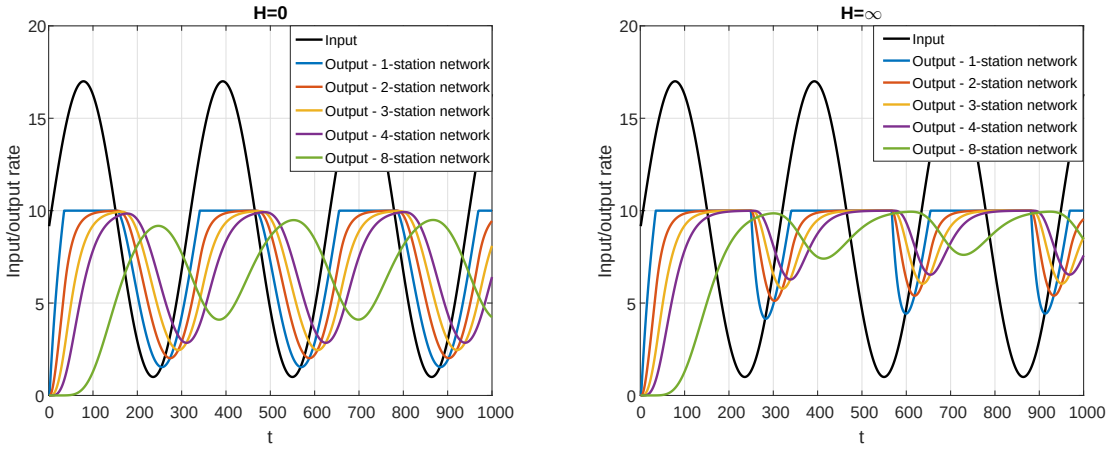


Figure 10: Line length effect on the network output rate with k i.i.d. stations, the sinusoidal arrival rate function in (40) with $\bar{\lambda} = 9$, $\beta = 8$ and $\gamma = 0.02$, $N_i = 200$, $\mu_i = 1/20$ and $q_i(0) = 0$, $\forall i \in \{1, \ldots, k\}$. Five networks of different length are considered.

reaches its upper bound. Here, the output from one station is the input for the next one. In Eick et al. (1993) an analytic expression was developed for the number of customers in the $M_t/G/\infty$ queue, with a sinusoidal arrival rate, as in (40). In particular, the output rate from Station 1 is given by

$$\delta_1(t) = \bar{\lambda} + \beta \left(\frac{\mu^2}{\mu^2 + \gamma^2} \sin(\gamma t) - \frac{\gamma\mu}{\mu^2 + \gamma^2} \cos(\gamma t) \right), \qquad t \geq 0. \tag{41}$$

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 (41) 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, $i = 1, 2, ..,$ i.i.d. stations in tandem, and exponential service times, is given by the following expression:

$$\delta_i(t) = \bar{\lambda} + \beta \left(C_1^{(i)} \sin(\gamma t) - C_2^{(i)} \cos(\gamma t) \right), \qquad t \geq 0, \tag{42}$$

where

$$\begin{aligned} &C_1^{(1)} = A_1, \quad C_2^{(1)} = B_1, \quad A_i = \frac{\mu_i^2}{\mu_i^2 + \gamma^2}, \quad B_i = \frac{\gamma\mu_i}{\mu_i^2 + \gamma^2}, \qquad i = 1, \ldots, k, \\ &C_1^{(i)} = C_1^{(i-1)} A_i - C_2^{(i-1)} B_i, \quad C_2^{(i)} = C_1^{(i-1)} B_i + C_2^{(i-1)} A_i, \qquad i = 2, \ldots, k. \end{aligned} \tag{43}$$

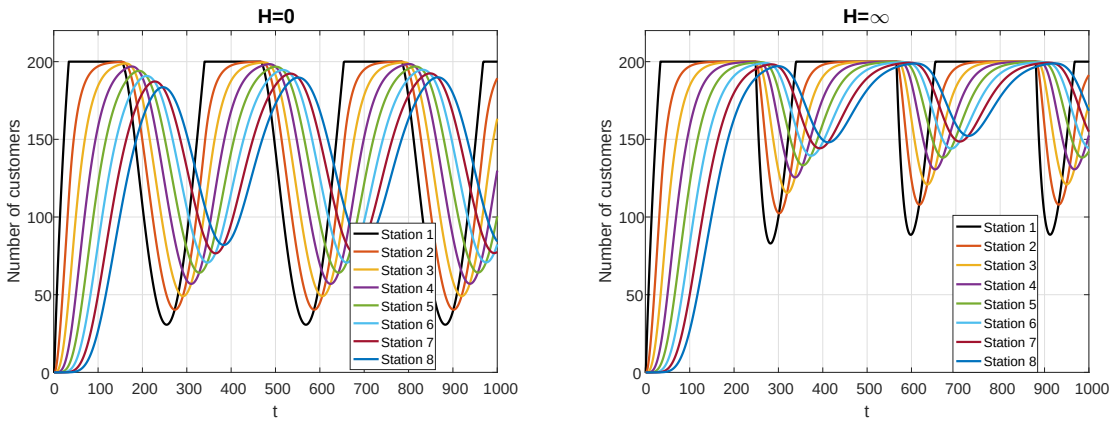


Figure 11: Total number of customers in each station in a network with eight i.i.d. stations and the sinusoidal arrival rate function in (40) with $\bar{\lambda} = 9$, $\beta = 8$ and $\gamma = 0.02$, $N_i = 200$, $\mu_i = 1/20$ and $q_i(0) = 0$, $i = 1, \ldots, 8$.

Figure 12 demonstrates that, in the special case of no blocking and sinusoidal arrival rate, our results are consistent with those derived in Eick et al. (1993). Using (42) and (43), 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 (42) do not exist for stochastic models with blocking, which renders our fluid model essential for analyzing system dynamics.

3.6.2 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 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 13 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 14 presents the total number of blocked customers in each station where the last station is the bottleneck. When $H = \infty$, blocking begins at Station 7 and surges backwards 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,

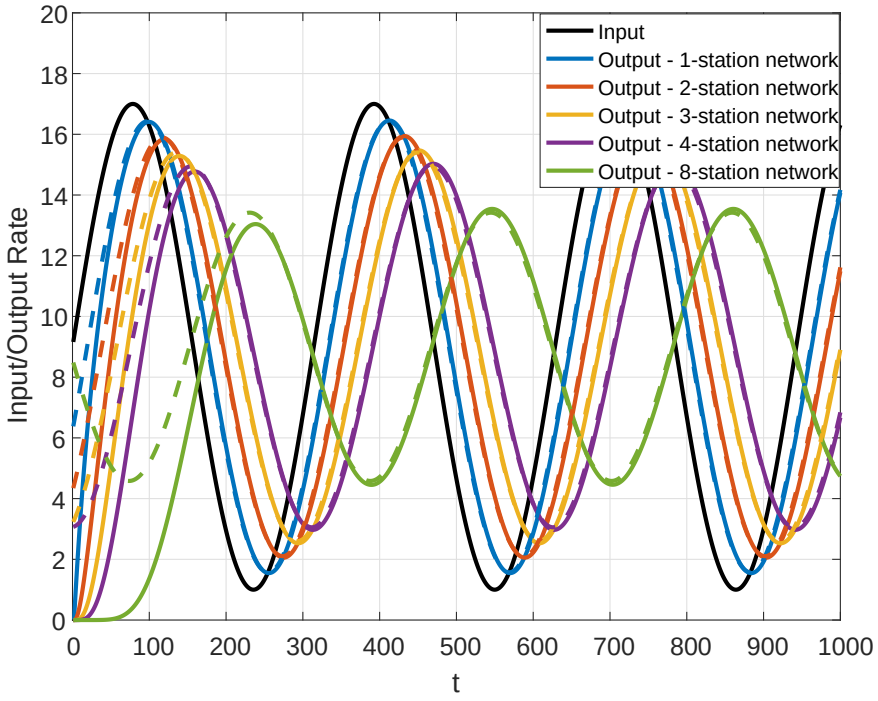


Figure 12: Input and output rates from networks with k i.i.d. stations – fluid model (solid lines) vs. values from (42) (dashed lines). The sinusoidal arrival rate function in (40) with $\bar{\lambda} = 9$, $\beta = 8$ and $\gamma = 0.02$, $N = 200$, $\mu = 1/20$ and $q_i(0) = 0$, $\forall i \in \{1, \ldots, k\}$. Five networks of different length are considered. Once the system reaches steady-state, the curves from the fluid model and the analytic formula overlap.

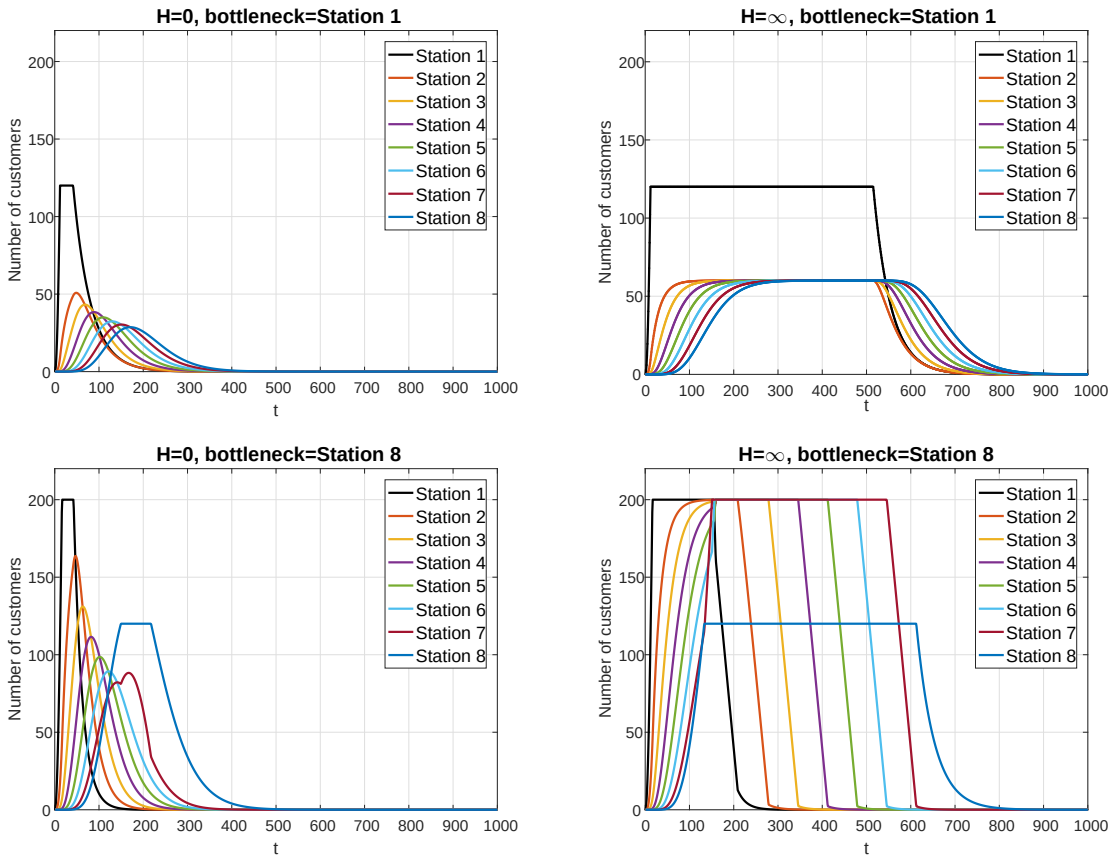


Figure 13: The bottleneck location effect on the total number of customers in each station. For the bottleneck station, j, $N_j = 120$, $\mu_j = 1/40$. For the other stations, $i = 1, \ldots, 8$, $i \neq j$ $N_i = 200$, $\mu_i = 1/20$, $q_m(0) = 0$, $m = 1, 2, \ldots, 8$, and $\lambda(t) = 2t$, $0 \leq t \leq 40$.

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.

3.6.3 Waiting Room Size

We now examine the effect of waiting room size before the first station. Figure 15 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 Figure 15 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.

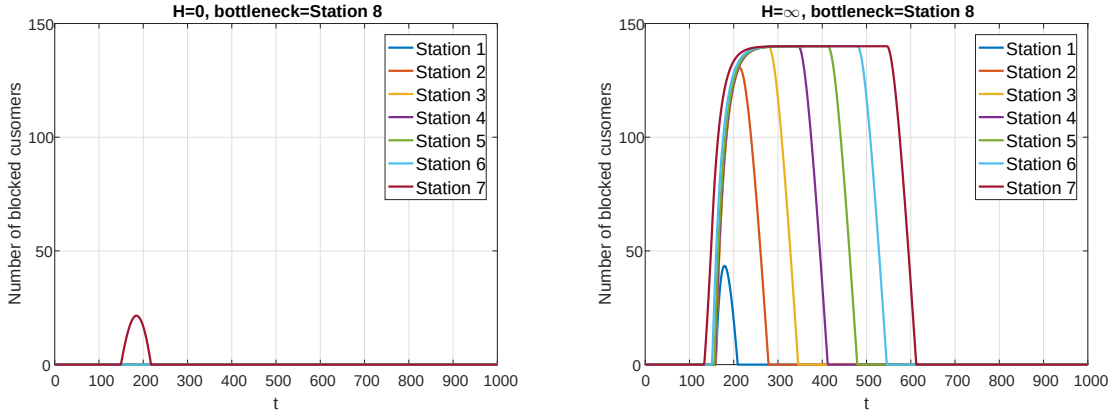


Figure 14: Number of blocked customers in each station when the last station (Station 8) is the bottleneck. $N_i = 200$, $\mu_i = 1/20$, $i = 1, \ldots, 7$, $N_8 = 120$, $\mu_8 = 1/40$. $q_m(0) = 0$, $m = 1, \ldots, 8$, and $\lambda(t) = 2t$, $0 \leq t \leq 40$. On the left, the curves for Stations 1–6 are zero and overlap.

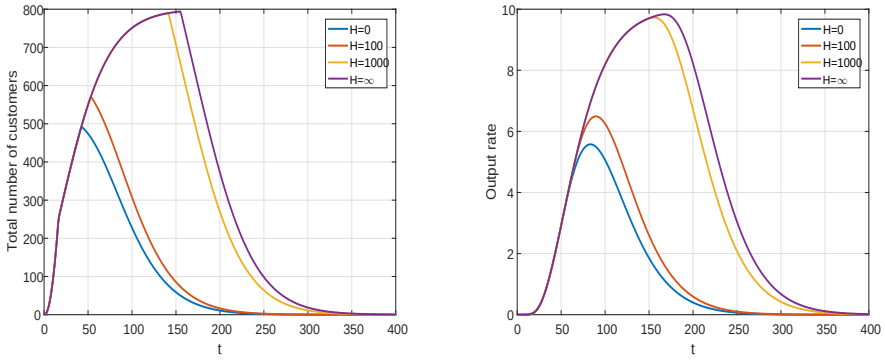


Figure 15: Waiting room size effect on the total number of customers (left plot) and on the output rate (right plot) in a network with four i.i.d. stations, where $N_i = 200$, $\mu_i = 1/20$, $q_i(0) = 0$, $i = 1, 2, 3, 4$ and $\lambda(t) = 2t$, $0 \leq t \leq 40$.

3.6.4 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 16 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 Avi-Itzhak (1965) and Avi-Itzhak and Yadin (1965), 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

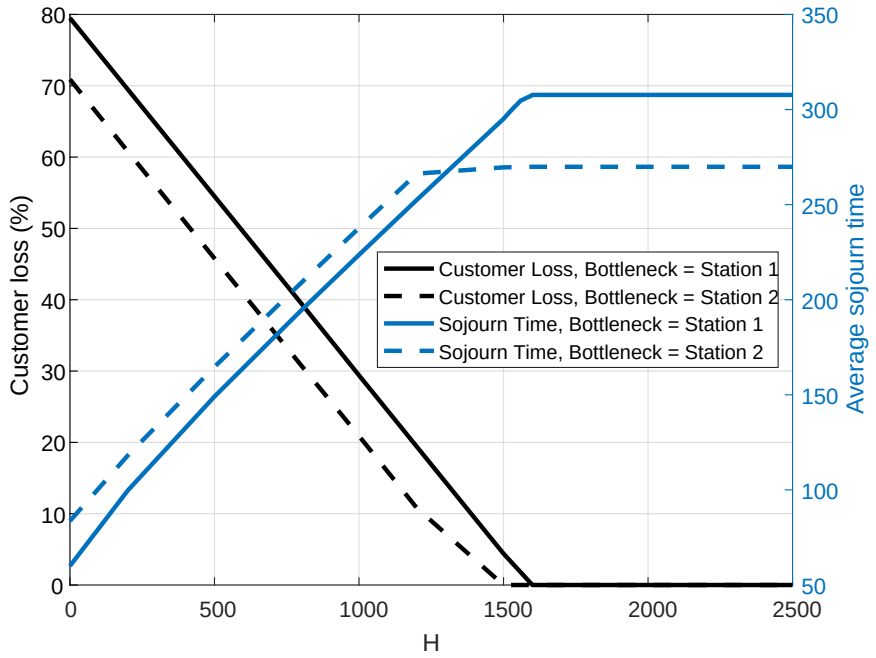


Figure 16: The effects of waiting room size and bottleneck location on sojourn time and customer loss in a tandem network with two stations, where $q_m(0) = 0$, $m = 1, 2$, and $\lambda(t) = 20$, $0 \leq t \leq 100$. In the bottleneck station, j, $N_j = 120$ and $\mu_j = 1/40$; in the other station, i, $N_i = 200$ and $\mu_i = 1/20$.

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 17 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 (Hillier and Boling, 1967; Conway et al., 1988) 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

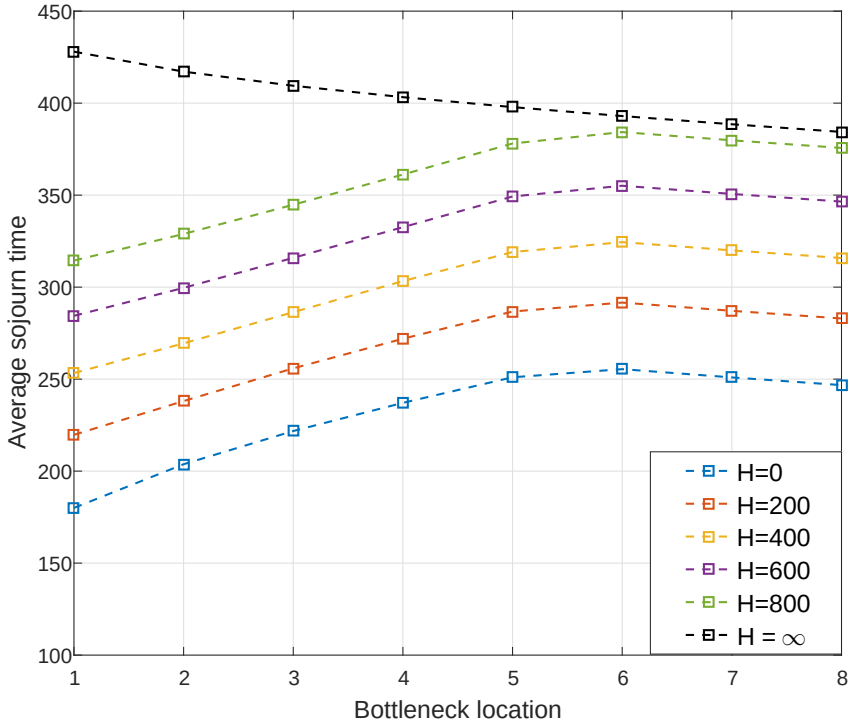


Figure 17: The effects of waiting room size and bottleneck location on the average sojourn time in a tandem network with eight station. Here, $q_m(0) = 0$, $m = 1, \ldots, 8$, and $\lambda(t) = 20$, $0 \leq t \leq 100$. In the bottleneck station, j, $N_j = 120$ and $\mu_j = 1/40$; in all other stations, $i = 1, 2, \ldots, 8$, $i \neq j$, $N_i = 200$ and $\mu_i = 1/20$.

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 18 presents each of these two components. The average waiting time (right plot) decreases as 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 19 presents the blocking time in each station and overall when $H = 0$. The 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.

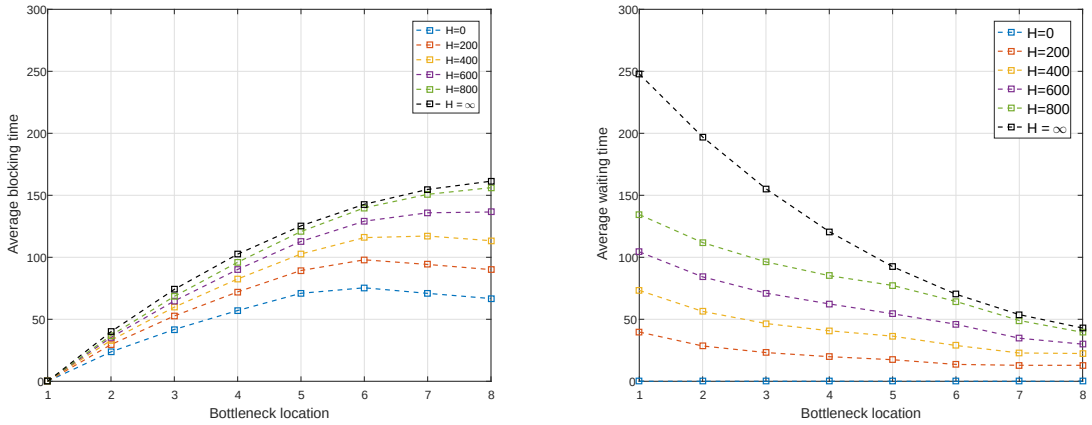


Figure 18: The effects of waiting room size and bottleneck location on the average blocking time (left plot) and the average waiting time (right plot). The summation of the waiting time, blocking time and service time yields the sojourn times presented in Figure 17.

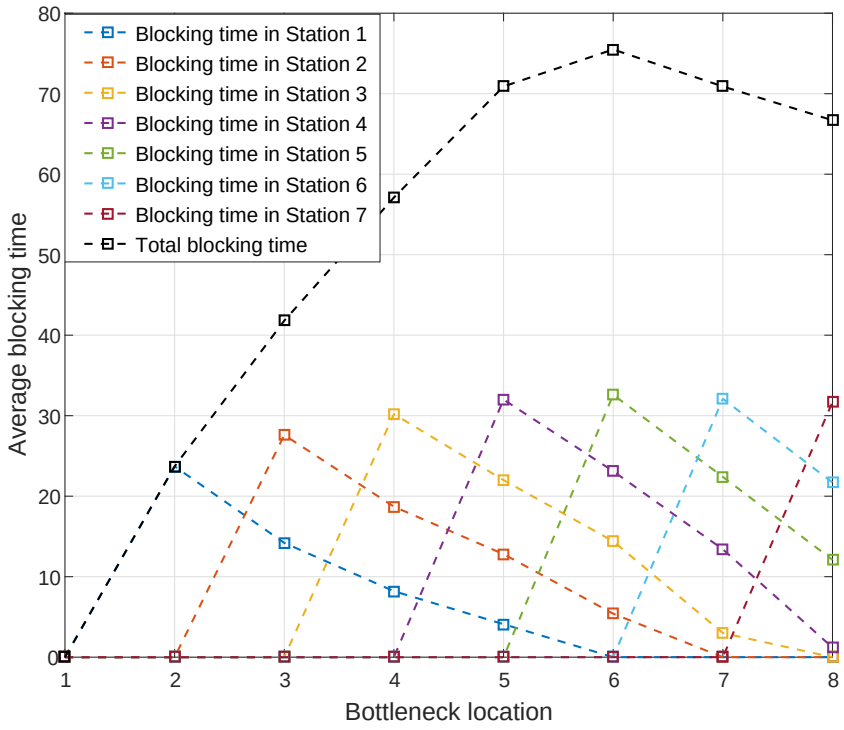


Figure 19: Average blocking time in each station and overall when $H = 0$.

4 Time-varying Tandem Queues under the BBS Mechanism

4.1 Introduction

Thus far, in Sections 2 and 3, we analyzed the Blocking After Service (BAS) mechanism. In this section, we analyze the Blocking Before Service (BBS) mechanism, which is also referred to as communication blocking or two-stage blocking (Perros, 1994; Balsamo et al., 2001). Under BBS, a service cannot begin at Station i, if there is no available capacity (storage or service) at Station $i+1$.

4.1.1 Motivation and Examples

Clearly, the BBS mechanism is prevalent in telecommunication networks (Suri and Diehl, 1984; Frein and Dallery, 1989; Seo et al., 2008). However, BBS is not uncommon in production lines; for example, in the chemical and pharmaceutical industries (Dogan-Sahiner and Altiok, 1998). In these production lines, 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 De Bruin et al. (2007). 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 (Daganzo et al., 2012). Such implementations could support/evaluate the practice of releasing cars to highways during rush hours (Bickel et al., 2003) or estimate travel times by navigation software (autonomous vehicles).

4.1.2 Results

In Section 4.4 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 (Mandelbaum et al., 1998, 1999; Whitt, 2004, 2006; Pang and Whitt, 2009; Liu and Whitt, 2011a, 2014). 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 discontinuous right-hand side (RHS) (Filippov, 2013; Zychlinski et al., 2018b), 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 4.5), 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 in each station and network throughput. In Section 4.5.2, we conclude the paper with an example of designing transfer protocols from surgery to recovery rooms in hospitals.

4.2 Literature Review

The most common types of blocking mechanisms for tandem flow lines are BAS and BBS (Altiok (1982); Perros (1994); Balsamo et al. (2001)). 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 (Desel and Silva, 1998). Thus, under this mechanism, a job can enter Station i, but cannot begin service until there is available capacity (buffer space or 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 (storage or service) 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 (Balsamo et al., 2001).

In Avi-Itzhak and Yadin (1965), 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 Avi-Itzhak (1965) and to the BBS mechanism in Avi-Itzhak and Levy (1995). Under the analyzed BBS, a job can enter a station only if the next k stations are available. In Avi-Itzhak and Halfin (1993), a k-station single-server network, with no intermediate buffers and an unlimited buffer before the first station, was analyzes under BAS and BBS. Note that the methodology we develop can, with slight modification (see Remark 4.2), 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 (Gershwin, 1987; Brandwajn and Jow, 1988; Dallery and Frein, 1993; van Vuuren et al., 2005; Osorio and Bierlaire, 2009). Several papers develop algorithms for approximating the steady-state throughput of closed single-server cyclic queueing networks with finite buffers (under both BBS and BAS in Onvural and Perros (1989) and under BBS in Suri and Diehl (1984) and Frein and Dallery (1989)).

4.3 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.

4.4 The Model

4.4.1 Notations and Assumptions

We model a network with k stations in tandem, as illustrated in Figure 20. This FCFS

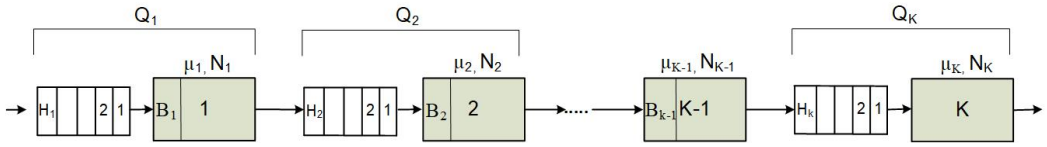


Figure 20: A network with k stations in tandem under the BBS mechanism.

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 (25).

2. "Basic" nominal service processes $D_i = \{D_i(t), t \geq 0\}$, $i = 1, 2, \ldots, k$, where $D_i(t)$ are 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 $Q_i(0)$, $i = 1, 2, \ldots, k$.

4.4.2 The Stochastic Model

Service at Station i can begin 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 Figure 20, 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 number of jobs at each station, is characterized by the following equations:

$$
\begin{aligned}
Q_1(t) &= Q_1(0) + A(t) - \int_0^t 1_{\{Q_1(u-)=H_1+N_1\}} \mathrm{d}A(u) \\
&\quad - D_1\left(\mu_1 \int_0^t [Q_1(u) \wedge N_1 \wedge (H_2 + N_2 - Q_2(u))]\mathrm{d}u\right), \\
Q_i(t) &= Q_i(0) + D_{i-1}\left(\mu_{i-1} \int_0^t [Q_{i-1}(u) \wedge N_{i-1} \wedge (H_i + N_i - Q_i(u))]\mathrm{d}u\right) \\
&\quad - D_i\left(\mu_i \int_0^t [Q_i(u) \wedge N_i \wedge (H_{i+1} + N_{i+1} - Q_{i+1}(u))]\mathrm{d}u\right), \quad i = 2, \ldots, k-1; \\
Q_k(t) &= Q_k(0) + D_{k-1}\left(\mu_{k-1} \int_0^t [Q_{k-1}(u) \wedge N_{k-1} \wedge (H_k + N_k - Q_k(u))]\mathrm{d}u\right) \\
&\quad - D_k\left(\mu_k \int_0^t [Q_k(u) \wedge N_k]\mathrm{d}u\right); \quad t \geq 0.
\end{aligned}
\tag{44}
$$

The integral in the first line of (44) represents the number of jobs that were forced to leave the system up until time t, since 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 (44) 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 (44), as follows:

$$
\begin{cases}
\begin{bmatrix} Q_1(t) \\ Q_2(t) \\ \vdots \\ Q_k(t) \end{bmatrix} = \begin{bmatrix} Y_1(t) - L(t) \\ Y_2(t) \\ \vdots \\ Y_k(t) \end{bmatrix} \leq \begin{bmatrix} H_1 + N_1 \\ H_2 + N_2 \\ \vdots \\ H_k + N_k \end{bmatrix}, \quad t \geq 0, \\
\mathrm{d}L(t) \geq 0, \;\; L(0) = 0, \\
\int_0^\infty 1_{\{Q_1(u-) < H_1+N_1\}} \mathrm{d}L(u) = 0,
\end{cases}
\tag{45}
$$

where

$$
\begin{aligned}
Y_1(t) &= Q_1(0) + A(t) - D_1\left(\mu_1 \int_0^t [Q_1(u) \wedge N_1 \wedge (H_2 + N_2 - Q_2(u))]\mathrm{d}u\right), \\
Y_i(t) &= Q_i(t), \quad i = 2, \ldots, k, \\
L(t) &= \int_0^t 1_{\{Q_1(u-) = H_1 + N_1\}} \mathrm{d}A(u).
\end{aligned} \tag{46}
$$

The last equation of (46) 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 Section 3.4.1 for details).

We simplify (45), so that the reflection will occur at zero, by letting

$$
R_i(t) = N_i + H_i - Q_i(t), \quad i = 1, \ldots, k, \quad t \geq 0, \tag{47}
$$

which gives rise to the following equivalent to (45):

$$
\begin{cases}
\begin{bmatrix} R_1(t) \\ R_2(t) \\ \vdots \\ R_k(t) \end{bmatrix} = \begin{bmatrix} \tilde{Y}_1(t) + L(t) \\ \tilde{Y}_2(t) \\ \vdots \\ \tilde{Y}_k(t) \end{bmatrix} \geq \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad t \geq 0, \\
\mathrm{d}L(t) \geq 0, \quad L(0) = 0, \\
\int_0^\infty 1_{\{R_1(t) > 0\}} \mathrm{d}L(t) = 0,
\end{cases} \tag{48}
$$

where $\tilde{Y}_i = H_i + N_i - Y_i$. From (48), 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 Mandelbaum and Pats (1995); Zychlinski et al. (2018b) for details).

4.4.3 Fluid Approximation

We now develop a fluid limit for our queueing model through the Functional Strong Law of Large Numbers (FSLLN). We begin with (48) and scale up the arrival rate and the size of the system (servers and waiting room) 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:

$$\begin{cases} R_1^\eta(t) = \tilde{Y}_1^\eta(t) + L^\eta(t), \\ R_i^\eta(t) = \tilde{Y}_i^\eta(t), \quad i = 2, \dots k, \quad t \geq 0, \end{cases} \tag{49}$$

where

$$\begin{aligned}
\tilde{Y_1}^\eta(\cdot) =& R_1^\eta(0) - A^\eta(\cdot) + D_1 \left(\mu_1 \int_0^\cdot [(\eta H_1 + \eta N_1 - R_1^\eta(u)) \wedge \eta N_1 \wedge R_2^\eta] \mathrm{d}u \right) \\
\tilde{Y_i}^\eta(\cdot) =& R_i^\eta(0) - D_{i-1} \left(\mu_{i-1} \int_0^\cdot \left[\left(\eta H_{i-1} + \eta N_{i-1} - R_{i-1}^\eta(u) \right) \wedge \eta N_{i-1} \wedge R_i^\eta \right] \mathrm{d}u \right) \\
& + D_i \left(\mu_i \int_0^t \left[(\eta H_i + \eta N_i - R_i^\eta) \wedge \eta N_i \wedge R_{i+1}^\eta(u) \right] \mathrm{d}u \right), \quad i = 2, \dots k-1; \\
\tilde{Y_k}^\eta(\cdot) =& R_k^\eta(0) - D_{k-1} \left(\mu_{k-1} \int_0^\cdot \left[\left(\eta H_{k-1} + \eta N_{k-1} - R_{k-1}^\eta(u) \right) \wedge \eta N_{k-1} \wedge R_k^\eta \right] \mathrm{d}u \right) \\
& + D_k \left(\mu_i \int_0^t [(\eta H_k + \eta N_k - R_k^\eta) \wedge \eta N_k] \mathrm{d}u \right); \\
L^\eta(\cdot) =& \int_0^\cdot 1_{\left\{ R_1^\eta(u-) = 0 \right\}} \mathrm{d}A^\eta(u).
\end{aligned}$$

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 $y^\eta = \{y^\eta(t), t \geq 0\}$, 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 Zychlinski et al. (2018b), Theorem 1, yields the following asymptotic behavior of r^η. Suppose that

$$\left\{ \eta^{-1} A^\eta(t), \, t \geq 0 \right\} \to \left\{ \int_0^t \lambda(u) du, \, t \geq 0 \right\}, \quad u.o.c. \ as \ \eta \to \infty, \tag{50}$$

as well as

$$\lim_{\eta \to \infty} r^\eta(0) = r(0), \quad a.s., \tag{51}$$

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:

$$\begin{cases} r_1(t) = r_1(0) - \int_0^t \left[\lambda(u) - \mu_1\left((H_1 + N_1 - r_1(u)) \wedge N_1 \wedge r_2(u)\right)\right] \mathrm{d}u + \, l(t) \geq 0, \\ r_i(t) = r_i(0) - \int_0^t \Big[\mu_{i-1}\left((H_{i-1} + N_{i-1} - r_{i-1}(u)) \wedge N_i \wedge r_i(u)\right) \\ \qquad\qquad -\mu_i\left((H_i + N_i - r_i(u)) \wedge N_i \wedge r_{i+1}(u)\right)\Big] \mathrm{d}u \geq 0, \;\; i = 2, \ldots, k-1; \\ r_k(t) = r_k(0) - \int_0^t \left[\mu_{k-1}\left((H_{k-1} + N_{k-1} - r_{k-1}(u)) \wedge N_{k-1} \wedge r_k(u)\right)\right] \\ \qquad\qquad -\mu_k\left((H_k + N_k - r_k(u)) \wedge N_k\right) \mathrm{d}u \geq 0, \\ \mathrm{d}l(t) \geq 0, \;\; l(0) = 0, \\ \int_0^\infty 1_{\{r_1(t) > 0\}} \mathrm{d}l(t) = 0; \end{cases} \tag{52}$$

The following proposition provides an equivalent representation to (52) in terms of our original formulation (i.e. $q(\cdot)$); see Appendix L for details. Implementing the solution in (53) numerically is straightforward since it is given by a set of differential equations with discontinuous RHS but, notable, without reflection.

Proposition 4.1. *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:*

$$\begin{aligned} q_1(t) &= q_1(0) - \mu_1 \int_0^t \left[q_1(u) \wedge N_1 \wedge (H_2 + N_2 - q_2(u))\right] \mathrm{d}u + \int_0^t \Big[1_{\{q_1(u) < H_1 + N_1\}} \cdot \lambda(u) \\ &\qquad + 1_{\{q_1(u) = H_1 + N_1\}} \cdot \left[\lambda(u) \wedge \mu_1 \left[N_1 \wedge (H_2 + N_2 - q_2(u))\right]\right] \mathrm{d}u, \\ q_i(t) &= q_i(0) + \mu_{i-1} \int_0^t \left[q_{i-1}(u) \wedge N_{i-1} \wedge (H_i + N_i - q_i(u))\right] \mathrm{d}u \\ &\qquad - \mu_i \int_0^t \left[q_i(u) \wedge N_i \wedge (H_{i+1} + N_{i+1} - q_{i+1}(u))\right] \mathrm{d}u, \quad i = 2, \ldots, k-1; \\ q_k(t) &= q_k(0) + \mu_{k-1} \int_0^t \left[q_{k-1}(u) \wedge N_{k-1} \wedge (H_k + N_k - q_k(u))\right] \mathrm{d}u \\ &\qquad - \mu_k \int_0^t \left[q_k(u) \wedge N_k\right] \mathrm{d}u. \end{aligned} \tag{53}$$

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

The function q will be referred to as the *fluid limit* associated with the queueing family Q^η, $\eta > 0$.

Remark 4.1. *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 $\theta\left[N_i - q_i(t) \wedge (H_{i+1} + N_{i+1} - q_{i+1}(t))\right]^+$; the abandonment rate of waiting jobs from Station i, $i = 1, \ldots, k$, at time t would be $\theta\left[q_i(t) - N_i\right]^+$. The mathematical analysis of models with abandonments does not differ from the one without.*

Remark 4.2. *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 (53) with $\wedge \bigwedge_{j=i}^{k} (H_j + N_j - q_j(u))$.*

4.4.4 Numerical Examples

To demonstrate that our proposed fluid model accurately describe the flow of jobs in the networks, we compared it to the average behavior of a stochastic simulation model constructed in SimEvents/MATLAB. 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. Let the arrival rate function be the sinusoidal function in (40). Solving the fluid equations in (53) was done by recursion and time discretization. Figure 21 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 stochastic system it approximates.

4.5 Network Performance

In this section we focus on steady-state performance, in particular network throughput under BBS and BAS (Section 4.5.1). 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

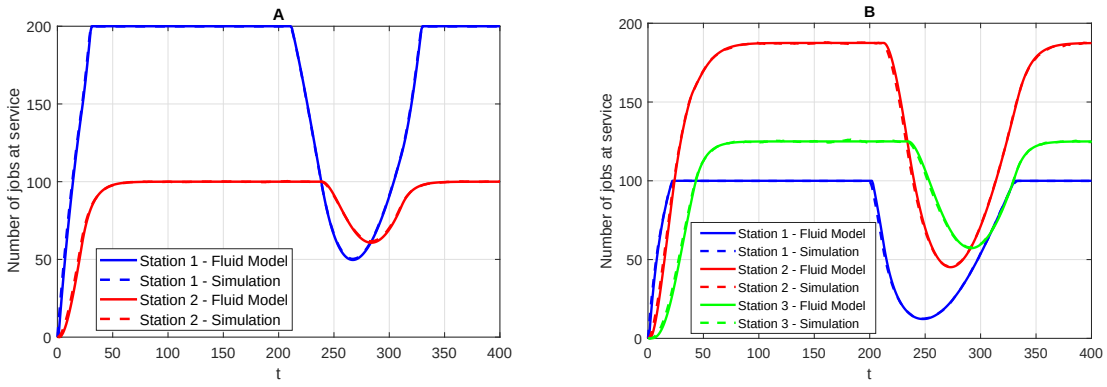


Figure 21: Total number of jobs at service - fluid model vs. simulation results, the sinusoidal arrival rate function in (40) with $\bar{\lambda} = 9$, $\beta = 8$ and $\gamma = 0.02$, $q_i(0) = 0$. In Plot A, $\mu_1 = \mu_2 = 1/20$, $H_1 = H_2 = 50$, $N_1 = 200$, $N_2 = 150$; in Plot B, $\mu_1 = 1/10$, $\mu_2 = 1/20$, $\mu_3 = 1/20$, $H_1 = H_2 = H_3 = 50$, $N_1 = 100$, $N_2 = 200$ and $N_3 = 200$.

i, respectively; thus,

$$s_i = \bar{q}_i \wedge N_i \wedge (H_{i+1} + N_{i+1} - \bar{q}_{i+1}), \quad i = 1, \ldots, k-1, \tag{54}$$
$$s_k = \bar{q}_k \wedge N_k.$$

For calculating steady-state performance, we start with (53), 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

$$\mu_1 s_1 = \lambda \cdot 1_{\{\bar{q}_1 < H_1 + N_1\}} + [\lambda \wedge \mu_1 (N_1 \wedge (H_2 + N_2 - \bar{q}_2))] \cdot 1_{\{\bar{q}_1 = H_1 + N_1\}}, \tag{55}$$
$$\mu_{i-1} s_{i-1} = \mu_i s_i, \qquad i = 2, \ldots, k.$$

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

Theorem 4.1. *Let δ denote the network throughput in the fluid model. Then*

$$\delta = \mu_i s_i = \lambda \wedge \bigwedge_{j=1}^{k} \mu_j N_j \wedge \bigwedge_{j=2}^{k} \frac{H_j + N_j}{1/\mu_{j-1} + 1/\mu_j}, \quad i = 1, \ldots, k. \tag{56}$$

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

$$\bar{q}_1 = H_1 + N_1; \tag{57}$$

$$\bar{q}_j = H_j + N_j - \delta/\mu_{i-1}, \qquad j = 2, \ldots, i;$$
$$\bar{q}_j = \delta/\mu_j, \qquad\qquad\qquad j = i+1, \ldots, k;$$

here

$$i = \min \left\{ \arg\min \bigwedge_{j=1}^{k} \mu_j N_j,\ \arg\min \bigwedge_{j=2}^{k} \frac{H_j + N_j}{1/\mu_{j-1} + 1/\mu_j} \right\}. \tag{58}$$

The interpretation of (56) 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 Dai and Vande Vate (2000), who defined a virtual workload condition for 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.

4.5.1 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) (Balsamo et al., 2001). 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 (buffer/server) at Station $i+1$, the job is blocked at Station i while occupying a server there. Figure 22 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 Figure 20. This change seems small but it is not: as shown momentarily, it can significantly affect network performances (see Figure 23).

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 (56) will be referred to as

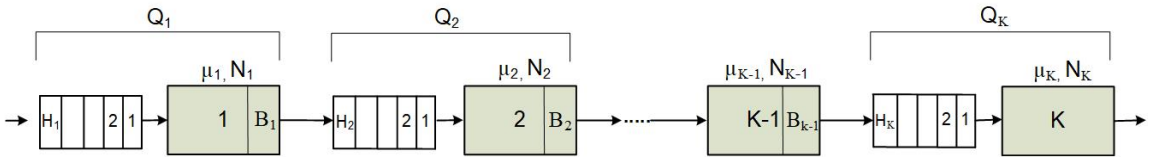


Figure 22: A network with k stations in tandem under the BAS mechanism.

δ^{BBS}); s_i^x, $i = 1, \ldots, k$, denote 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 Equation (15) in Zychlinski et al. (2018b), with $\lambda(t) \equiv \lambda$, $\forall t \geq 0$), yields the following BAS throughput:

$$\delta^{\text{BAS}} = \mu_i s_i^{\text{BAS}} = \lambda \wedge \bigwedge_{j=1}^{k} \mu_j N_j, \qquad i = 1, \ldots, k. \tag{59}$$

Remark 4.3. *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 (56), since they influence the bottleneck processing capacity.*

Remark 4.4. *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 (56)) does not determine the throughput, and we get that $\delta^{BBS} = \delta^{BAS}$.

Figure 23 presents the total number of jobs in service at each station under the two mechanisms. 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.

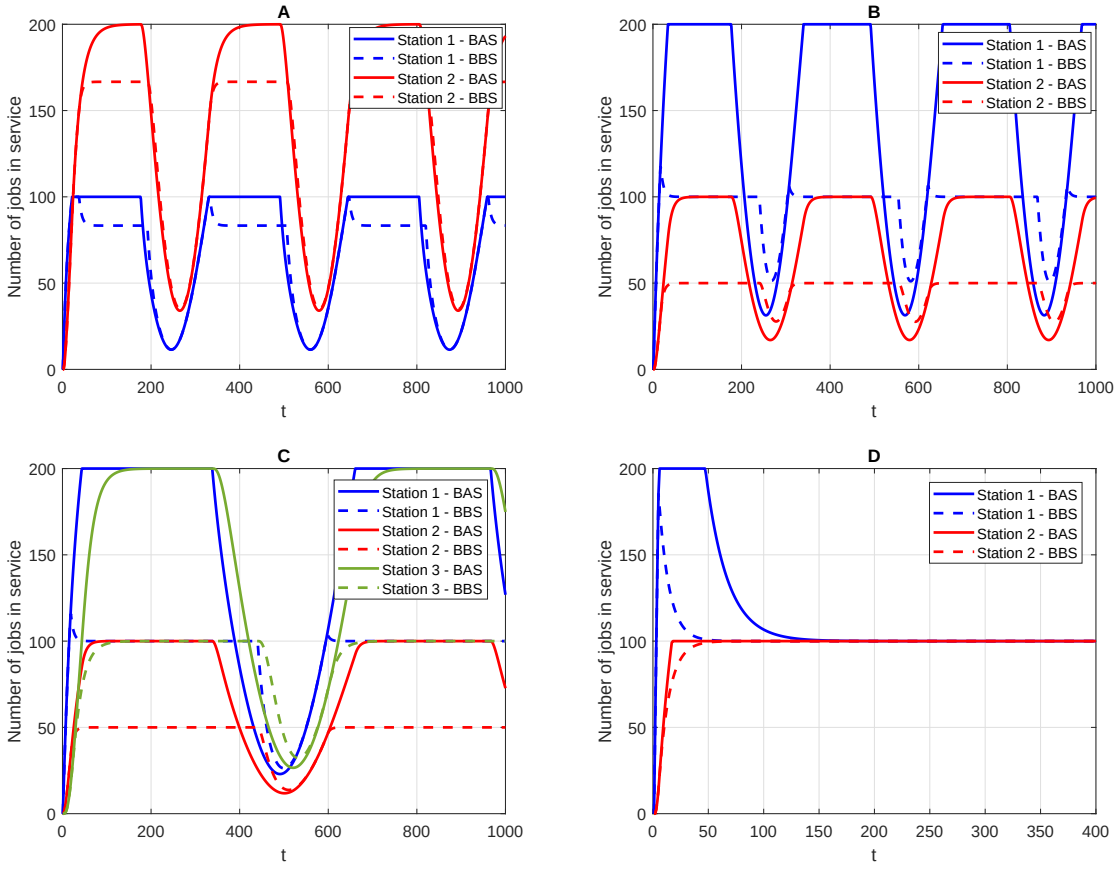


Figure 23: Total number of jobs in service at each station - BBS vs. BAS with $q(0) = 0$. In Plot A, the sinusoidal arrival rate function in (40) with $\bar{\lambda} = 9$, $\beta = 8$ and $\gamma = 0.02$, $N_1 = 100$, $N_2 = 200$, $H_1 = H_2 = 50$, $\mu_1 = 1/10$, $\mu_2 = 1/20$. In Plot B, the station 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 $\mu_1 = \mu_2 = 1/20$.

Combining (56) and (59) yields the following:

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

thus, $\delta^{\mathrm{BBS}} \leq \delta^{\mathrm{BAS}}$. 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 Figure 23, 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

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

4.5.2 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 (56) and (59); the latter is calculated by first calculating the number of patients in the system (Theorem 4.1) 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 ten patients per hour and average sojourn time of two hours. Now, suppose that recovery takes on average two hours (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 hours, under BBS it reaches 5 hours. Under this setting, BAS is superior according to both performance measurements.

5 Summary and Future research Directions

This thesis is grounded on modeling, developing and analyzing time-varying fluid networks with blocking. Beyond having an intrinsic value of their own, these mathe-

matical models are also strong limits of corresponding stochastic systems, which yield operational insights on performance of the latter. Our models are motivated by three applications: The first is patient flow analysis between hospitals and geriatric institutions, in order to improve their joint operation (Section 2); the second application includes analysis of time-varying tandem flow lines with blocking, customer loss and reflection (Section 3); the third application includes analysis of time-varying tandem flow lines under the BBS mechanism, which arises in telecommunication networks, production lines and healthcare systems (Section 4). These three applications are related through their essential characteristics: Time-variability and blocking.

Future research can include practical and theoretical directions. One possible direction is to exploit new data-driven and mathematical tools together with game-theory analysis, to investigate and improve patient flow between the community, hospitals and geriatric institutions. "Clalit", the largest Israeli Health Maintenance Organization (HMO), has recently provided us with patient flow data, at the level of individual patients, between Emergency Departments, hospital wards and geriatric institutions. Such individual patient flow data is usually confidential and very hard to attain. The willingness of "Clalit" to share its data with us is significant and highlights the importance it assigns to this issue. Analyzing these data will open up new opportunities and directions for research in both exploratory data analysis (EDA) and queueing science. The work we envision has the potential to reveal important features that cannot be explained by existing models. The proposed EDA will enable conducting an integrative analysis, for example, relating transfer delays to readmission rates, treatment durations and patient clinical condition. Addressing these issues will most likely require developing new queueing models and theory, jointly with supporting statistical analysis.

Another research direction will include several stakeholders such as the government, HMOs and private or corporation hospitals. In order to capture the balance of forces among these stakeholders, the analysis should accommodate all of them. Combining these factors will require conducting a game theoretic view, in which each stakeholder makes bed allocation decisions for the hospitals and institutions it operates. The mode of analysis we envision is in the spirit of Zhang et al. (2016), who use game theoretic analysis among hospitals to asses incentives by the United States Medicare and Medicaid policy for reducing readmissions.

Yet another possible direction is to extend the development of our time-varying many-server fluid models to fork-join networks with blocking (Dallery et al., 1994, 1997). This direction would require specific definitions of new blocking mechanisms and priority protocols. For example, suppose that all servers at Station X are busy, and there are blocked customers at Station Y and Z awaiting a server at X. When an X-server becomes available, who among the waiting customers will get it?

Appendices

A Fluid Model Validation

To validate our model we used the following patient flow data:

1. Two years of patient flow data from a district that includes four hospitals and three geriatric institutions (three rehabilitation wards, two mechanical ventilation wards and three skilled nursing wards).

2. Two years of waiting lists for geriatric wards, including individual waiting times from our Partner Hospital.

Based on the patient flow data, model parameters were first estimated, then inspected and validated by expert doctors. The parameter values used for the validation are: $\mu_1 = 1/4.85$, $\mu_2 = 1/30$, $\mu_3 = 1/160$, $\mu_4 = 1/45$, $\beta_2 = 1/250$, $\beta_3 = 1/1000$, $\beta_4 = 1/1000$, $\theta_1 = 1/125$, $\theta_2 = 1/2500$, $\theta_3 = 1/1000$, $\theta_4 = 1/1000$, $N_1 = 600$, $N_2 = 226$, $N_3 = 93$, $N_4 = 120$ (we used day as a time unit). For example, Station 1 contains 600 beds; the average treatment duration there is 4.85 days and the average time to death is 125 days.

Estimating the rates of mortality and readmission were done using the MLE (Maximum Likelihood Estimator), that is prevalent for estimating censored data, such as patience and retries in service systems (see Zohar et al., 2002 for details). Here, we adjust the estimator for the case where patients die while being in treatment, rather than just while waiting in queue. To this end, instead of the actual waiting time, we consider the actual treatment time.

The time-varying arrival rates and routing probabilities were also derived from the data. The average monthly arrival rate was 3,632 patients per month (with a minimum of 3,559 and maximum 3,774), and the average routing probabilities to each geriatric ward were 9% for rehabilitation wards, 0.8% for mechanical ventilation and 2.4% for skilled nursing care.

Using these parameters, we numerically (via Matlab) solved (7), which resulted in the number of patients in each ward at any time ($q_i(t)$ for $i = 1, 2, 3, 4$) and the number of blocked patients waiting for each ward ($b_i(t)$ for $i = 2, 3, 4$). Figure 3 shows the length of the waiting lists for each ward, using a daily resolution during one calendar year, according to the data and the fluid model. The very good fit implies that the

fluid model is appropriate for modeling the system considered here. The three geriatric wards work at full capacity throughout the year; there are always blocked patients in the hospital and any vacant geriatric bed is immediately filled.

In addition to comparing the fluid model with real data, we validated its accuracy against a discrete event simulation of a stochastic system, which we developed for this purpose in SimEvents/MATLAB. We conducted experiments for several scenarios; in each one, we considered three levels of the scaling parameter η. In our simulation model, the patients arrive according to a non-homogeneous Poisson process that was used to represent a process with a general, time-dependent arrivals, as prevalent in hospitals (Bekker and de Bruin, 2010; Yom-Tov and Mandelbaum, 2014; Shi et al., 2015; Armony et al., 2015). The treatment rates were randomly generated from exponential, Phase-type (as a mixture of two exponentials) and Lognormal distributions, which are typical for describing lengths of stay in hospitals and geriatric wards (McClean and Millard, 1993; Marazzi et al., 1998; Xie et al., 2005; McClean and Millard, 2006; Faddy et al., 2009; Armony et al., 2015). The expectations of these three distributions were equal when compared in a specific scenario. For each scenario and η we used 300 replications, each for 1000 days, and calculated the Root Mean Square Error (RMSE) using the following formula:

$$RMSE = \sqrt{\frac{\int_{t=0}^{T} \sum_{i=2}^{4} \left[q_i^{sim}(t) - q_i^{fluid}(t)\right]^2 \mathrm{d}t}{T}};$$

here $q_i^{sim}(t)$ is the total number of patients in Station i at time t according to the simulation results and $q_i^{fluid}(t)$ is the number according to the fluid model. The results are summarized in Tables 5 and 6. An example for Scenario 1 with $\eta = 10$ is illustrated in Figure 24. As expected, fluid models become more accurate as the scaling parameter η becomes larger. In general, the best results were achieved for the Exponential distributions. However, the model is quite accurate even for the Phase-type and Lognormal distributions. In all cases, the fluid model accurately forecasts, within a 95% confidence interval, the stochastic behavior of the corresponding simulation. The percentage of error, relative to system capacity, varied from 0.6% to 2.4%. However, for the size of systems in which we are interested (Scenarios 1–18), the percentage of error was less than 1%.

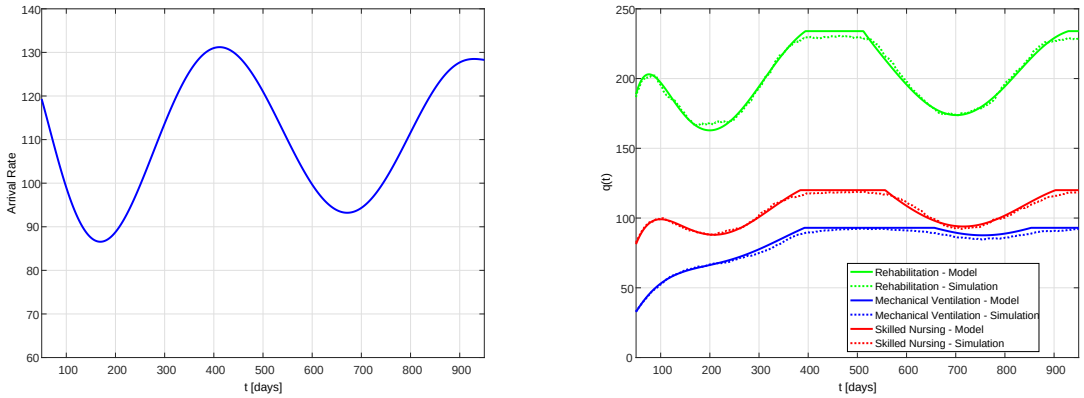


Figure 24: Scenario 1 in Table 5. On the right: Total number of patients in each geriatric ward - fluid model vs. simulation. On the left: The arrival rate $\lambda(t)$.

No.	N_1, N_2, N_3, N_4	$\mu_1, \mu_2, \mu_3, \mu_4$	p_{12}, p_{13}, p_{14}	distribution	$\lambda(t)$
1	600, 234, 93, 120	1/4.85, 1/30, 1/160, 1/45	0.09, 0.008, 0.024	Exponential	polyno.
2	600, 234, 93, 120	1/4.85, 1/30, 1/160, 1/45	0.09, 0.008, 0.024	Phase-Type	polyno.
3	600, 234, 93, 120	1/4.85, 1/30, 1/160, 1/45	0.09, 0.008, 0.024	Lognormal	polyno.
4	600, 234, 93, 120	1/4.85, 1/30, 1/160, 1/45	0.09, 0.008, 0.024	Exponential	polyno./10
5	600, 234, 93, 120	1/4.85, 1/30, 1/160, 1/45	0.09, 0.008, 0.024	Phase-Type	polyno./10
6	600, 234, 93, 120	1/4.85, 1/30, 1/160, 1/45	0.09, 0.008, 0.024	Lognormal	polyno./10
7	600, 234, 93, 120	1/4.85, 1/30, 1/160, 1/45	0.09, 0.008, 0.024	Exponential	polyno.·10
8	600, 234, 93, 120	1/4.85, 1/30, 1/160, 1/45	0.09, 0.008, 0.024	Phase-Type	polyno.·10
9	600, 234, 93, 120	1/4.85, 1/30, 1/160, 1/45	0.09, 0.008, 0.024	Lognormal	polyno.·10
10	600, 200, 200, 200	1/5, 1/30, 1/30, 1/30	0.25, 0.25, 0.25	Exponential	polyno.
11	600, 200, 200, 200	1/5, 1/30, 1/30, 1/30	0.25, 0.25, 0.25	Phase-Type	polyno.
12	600, 200, 200, 200	1/5, 1/30, 1/30, 1/30	0.25, 0.25, 0.25	Lognormal	polyno.
13	600, 200, 200, 200	1/5, 1/30, 1/30, 1/30	0.25, 0.25, 0.25	Exponential	polyno.·10
14	600, 200, 200, 200	1/5, 1/30, 1/30, 1/30	0.25, 0.25, 0.25	Phase-Type	polyno.·10
15	600, 200, 200, 200	1/5, 1/30, 1/30, 1/30	0.25, 0.25, 0.25	Lognormal	polyno.·10
16	600, 200, 100, 100	1/5, 1/15, 1/15, 1/15	0.25, 0.25, 0.25	Exponential	polyno.
17	600, 200, 100, 100	1/5, 1/15, 1/15, 1/15	0.25, 0.25, 0.25	Phase-Type	polyno.
18	600, 200, 100, 100	1/5, 1/15, 1/15, 1/15	0.25, 0.25, 0.25	Lognormal	polyno.
19	60, 20, 20, 20	1/5, 1/30, 1/30, 1/30	0.09, 0.008, 0.024	Exponential	polyno./10
20	60, 20, 20, 20	1/5, 1/30, 1/30, 1/30	0.09, 0.008, 0.024	Phase-Type	polyno./10
21	60, 20, 20, 20	1/5, 1/30, 1/30, 1/30	0.09, 0.008, 0.024	Lognormal	polyno./10

Table 5: Parameters of scenarios. The polynomial arrival rate is $\lambda(t) = C_1t^7 + C_2t^6 + C_3t^5 + C_4t^4 + C_5t^3 + C_6t^2 + C_7t + C_8$ where $C_1 = 5.8656 \cdot 10^{-17}, C_2 = -2.1573 \cdot 10^{-13}, C_3 = 3.0756 \cdot 10^{-10}, C_4 = -2.1132 \cdot 10^{-7}, C_5 = 6.9813 \cdot 10^{-5}, C_6 = -0.0091, C_7 = 0.0718, C_8 = 130.8259$.

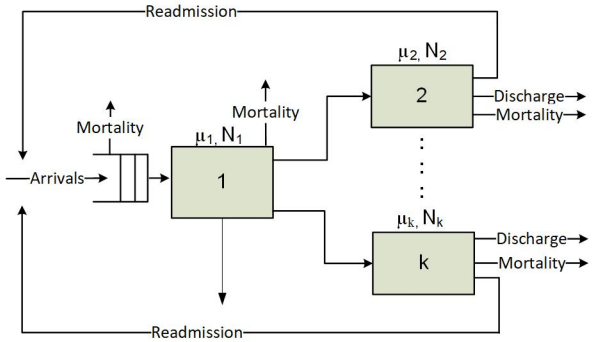
No.	$\eta = 1$	$\eta = 10$	$\eta = 100$		No.	$\eta = 1$	$\eta = 10$	$\eta = 100$
1	8.07	2.42	0.89		12	11.4	5.12	1.14
2	8.92	3.52	1.24		13	7.42	2.13	0.78
3	11.68	5.21	1.32		14	7.74	3.41	0.96
4	9.87	2.78	0.97		15	10.98	4.64	1.01
5	10.76	3.85	1.45		16	8.01	2.23	0.72
6	12.52	5.69	1.38		17	8.59	3.41	0.91
7	7.67	2.28	0.82		18	11.23	4.76	0.98
8	8.32	3.44	1.05		19	2.35	1.95	0.58
9	11.21	5.09	1.14		20	2.76	2.28	1.24
10	8.03	2.28	0.82		21	2.91	2.43	1.32
11	8.65	3.5	1.05		Avg	8.53	3.56	1.04

Table 6: Total number in each station - fluid model vs. Simulation - RMSE results

B Fluid Model for Blocking: Convergence of the Stochastic Model

We now develop a fluid model with blocking, mortality and readmissions for a network with k stations, as illustrated in Figure 25. Our system is characterized by the following

Figure 25: A k-station network



(deterministic) parameters:

1. Arrival rate to Station 1 is $\lambda(t)$, $t \geq 0$;
2. Service rate $\mu_i > 0$, $i = 1, \ldots, k$;
3. Mortality rate $\theta_i > 0$, $i = 1, \ldots, k$;
4. readmission rate $\beta_i > 0$, $i = 2, \ldots, k$, from Station i back to Station 1;
5. Number of servers (beds) N_i, $i = 1, \ldots, k$;
6. Transfer probability $p_{ij}(t)$ from Station i to Station j;

5. Unlimited waiting room before Station 1;

6. No waiting room before Stations $i = 2, \ldots, k$.

The stochastic model is created from the following stochastic building blocks A, S_i, $i = 1, \ldots, (2k-1)$, M_i, $i = 1, \ldots, k$ and R_i, $i = 2, \ldots, k$, which are assumed to be independent, as well as $X_i(0)$, $i = 1, \ldots, k$:

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. The arrival rate $\lambda(t)$, $t \geq 0$ is related to A via

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

   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 (constant arrival rate 1).

2. "Basic" nominal service processes $S_i = \{S_i(t), t \geq 0\}$, $i = 1, .., (2k-1)$, where $S_i(t)$ are standard Poisson processes.

3. "Basic" nominal mortality processes $M_i = \{M_i(t), t \geq 0\}$, $i = 1, .., k$, where $M_i(t)$ are standard Poisson processes.

4. "Basic" nominal readmission processes $R_i = \{r_i(t), t \geq 0\}$, $i = 2, .., k$, where $r_i(t)$ are standard Poisson processes.

5. Initial number of customers in each state $X_i(0)$, $i = 1, \ldots, k$.

The above building blocks will yield the following k 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.

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 1, require service at Station i, but have not yet completed their service at Station i at time t.

We assume that the blocking mechanism is blocking-*after*-service (BAS) (Balsamo et al., 2001). Thus, if upon service completion at Station 1, the destination station is saturated, the customer will be forced to stay in Station 1, while occupying a server there until the destination station becomes available. The latter means that when a server completes service, the blocked customer immediately transfers and starts service.

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

$$\begin{aligned} Q_1(t) =& X_1(t) + \sum_{i=2}^{k} \left(X_i(t) - N_i\right)^+; \\ Q_j(t) =& X_j(t) \wedge N_j, \qquad j = 2, \ldots, k; \end{aligned}$$

here

$$\begin{aligned} X_1(t) = X_1(0) + A(t) + \sum_{m=2}^{k} R_m \left(\beta_m \int_0^t (X_m(u) \wedge N_m)\mathrm{d}m\right) - M_1 \left(\theta_1 \int_0^t X_m(u)\mathrm{d}u\right) \\ - \sum_{m=2}^{k} S_m \left(\mu_1 \int_0^t p_{1m}(u) \left[\cdot X_1(u) \wedge \left(N_1 - \sum_{i=2}^{k} \left(X_i(u) - N_i\right)^+\right)\right] \mathrm{d}u\right) \\ - S_1 \left(\mu_1 \int_0^t \left(1 - \sum_{i=2}^{k} p_{1i}(u)\right) \left[X_1(u) \wedge \left(N_1 - \sum_{i=2}^{k} \left(X_i(u) - N_i\right)^+\right)\right] \mathrm{d}u\right), \end{aligned} \tag{60}$$

$$\begin{aligned} X_j(t) = X_j(0) + S_1 \left(\mu_1 \int_0^t p_{1j}(t) \left[X_1(u) \wedge \left(N_1 - \sum_{i=2}^{k} \left(X_i(u) - N_i\right)^+\right)\right] \mathrm{d}u\right) \\ - R_j \left(\beta_j \int_0^t (X_j(u) \wedge N_j)\mathrm{d}m\right) - M_j \left(\theta_j \int_0^t X_j(u)\mathrm{d}u\right) \\ - S_{k-1+j} \left(\mu_j \int_0^t (X_j(u) \wedge N_j)du\right), \qquad j = 2, \ldots, k. \end{aligned} \tag{61}$$

An inductive construction over time shows that (60) uniquely determines the process X.

Note that $\left(X_i(t) - N_i\right)^+$, $i = 2, \ldots, k$, is the number of blocked customers waiting for an available server in Station i.

B.1 Fluid Approximation - FSLLN

We now develop a fluid limit for our queueing model through a Functional Strong Law of Large Numbers (FSLLN). We begin with (60) and scale up the arrival rate and the number of servers by $\eta > 0$, $\eta \to \infty$. This η will serve as an index of a corresponding queueing process X^η:

$$\begin{aligned}
X_1^\eta(t) = X_1^\eta(0) + A^\eta(t) + \sum_{m=2}^{k} R_m\left(\beta_m \int_0^t (X_m^\eta(u) \wedge \eta N_m)\mathrm{d}m\right) - M_1\left(\theta_1 \int_0^t X_m^\eta(u)\mathrm{d}u\right) \\
- \sum_{m=2}^{k} S_m\left(\mu_1 \int_0^t p_{1m}(u)\left[X_1^\eta(u) \wedge \left(\eta N_1 - \sum_{i=2}^{k}\left(X_i^\eta(t) - \eta N_i\right)^+\right)\right]\mathrm{du}\right) \\
- S_1\left(\mu_1 \int_0^t \left(1 - \sum_{i=2}^{k} p_{1i}(u)\right)\left[X_1^\eta(u) \wedge \left(\eta N_1 - \sum_{i=2}^{k}\left(X_i^\eta(t) - \eta N_i\right)^+\right)\right]\mathrm{du}\right), \\
X_j^\eta(t) = X_j^\eta(0) + S_1\left(\mu_1 \int_0^t p_{1j}(u)\left[X_1^\eta(u) \wedge \left(\eta N_1 - \sum_{i=2}^{k}\left(X_i^\eta(t) - \eta N_i\right)^+\right)\right]\mathrm{du}\right) \\
- R_j\left(\beta_j \int_0^t (X_j^\eta(u) \wedge \eta N_j)\mathrm{d}m\right) - M_j\left(\theta_j \int_0^t X_j^\eta(u)\mathrm{d}u\right) \\
- S_{k-1+j}\left(\mu_j \int_0^t (X_j^\eta(u) \wedge \eta N_j)du\right), \qquad j = 2, \ldots, k.
\end{aligned}$$

Suppose that A^η, $\eta > 0$, the family of arrival processes satisfies the following FSLLN:

$$\lim_{\eta\to\infty} \frac{1}{\eta} A^\eta(t) = \int_0^t \lambda(u)\mathrm{d}u; \tag{62}$$

here the convergence is uniformly on compact sets of $t \geq 0$ (u.o.c.). For example, in the non-homogenuous Poisson process

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

Other examples can be found in Liu and Whitt (2011a, 2012a, 2014).
Assumption (62) is all that is required in order to apply Theorem 2.2 in Mandelbaum et al. (1998) and get

$$\lim_{\eta\to\infty} \frac{1}{\eta} X_i^\eta(t) = x_i(t), \quad \text{u.o.c.}, \quad i = 1, \ldots, k,$$

where x_i, $i = 1, 2, .., k$, are referred to as the *fluid limit* associated with the queueing

family X_i^η, $i = 1, \ldots, k$. The functions x_i constitute the unique solution of the following ODE:

$$x_1(t) = x_1(0) + \int_0^t \Big[\lambda(u) + \sum_{i=2}^{4} \beta_i\big(x_i(t) \wedge N_i\big) - \mu_1\Big(x_1(u) \wedge \big(N_1 - \sum_{i=2}^{k}(x_i(u) - N_i)^+\big)\Big) - \theta_1 x_1(t)\Big]\,\mathrm{d}u,$$

$$x_j(t) = x_j(0) + \int_0^t \Big[p_{1j}(u) \cdot \mu_1\Big(x_1(u) \wedge \big(N_1 - \sum_{i=2}^{k}(x_i(u) - N_i)^+\big)\Big) - (\mu_j + \beta_j)\big(x_j(u) \wedge N_j\big) - \theta_j x_j(t)\Big]\,\mathrm{d}u, \qquad j = 2, \ldots, k.$$

We now introduce the functions q_i, $i = 1, \ldots, k$, as the *fluid limit* associated with the queueing family Q^η; these functions are given by

$$q_1(t) = x_1(t) + \sum_{i=2}^{k} \big(x_i(t) - N_i\big)^+,$$
$$q_j(t) = x_j(t) \wedge N_j, \qquad j = 2, \ldots, k.$$

C Proof of Theorem 2.1

The function $C(N)$ in (14) equals

$$C(N) = \mathit{constant} - (C_o + C_u) \int_0^N \big[f(x) - Z\big]\mathrm{d}x, \tag{63}$$

where

$$f(x) = \int_0^T 1_{\{r_d(t) \geq x\}}\mathrm{d}t \quad \text{and} \quad Z = \frac{C_o T}{C_o + C_u}. \tag{64}$$

Therefore, it suffices to prove that the function $F(N)$, given by

$$F(N) = \int_0^N [f(x) - Z]\mathrm{d}x, \tag{65}$$

is maximized by N^* in (16).

Note that $f(x)$ is non-negative and non-increasing in x, where $f(0) = T$ and $\lim_{x\to\infty} f(x) = 0$. In addition, $Z \in [0, T]$, hence $f(x)$ crosses level Z. The function $F(N)$, for N starting from 0, is first an integral of a non-negative integrand, hence is increasing in N. Then, after the first N for which $f(N) = Z$, it is decreasing. This proves that $F(N)$

is maximized (globally) at point N, where $f(N) = Z$.

We conclude the proof by showing that N^* in (16) satisfies $f(N^*) = Z$. Substituting N^* into (64) gives

$$f(N^*) = \int_0^T 1_{\{r_d(t) \geq r_d(Z)\}} \mathrm{d}t = \int_0^T 1_{\{t \leq Z\}}(t) \mathrm{d}t = Z,$$

since r_d is a decreasing function. Therefore, $N^* = r_d(Z)$, as in (16).

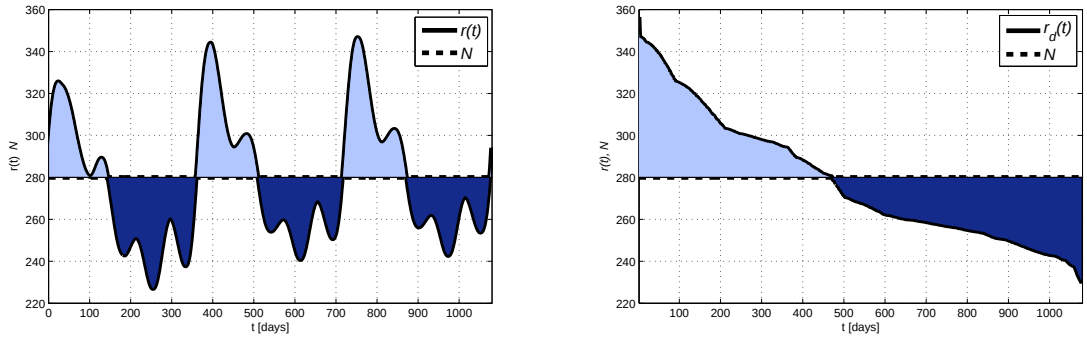
Remark C.1. *When r_d is continuous and strictly decreasing, $f(x)$ is in fact its inverse r_d^{-1}.*

D Choosing the Candidate Solution

We now describe the method that motivates N^*, as in (16), to be a natural candidate for maximizing $C(N)$ in (14). This method requires additional assumptions about $r(t)$, $r_d(t)$ and λ. Theorem 2.1, though, does not make these assumptions and is, therefore, more general.

Figure 26 shows an illustration of the overage and underage periods for a specific number of beds ($N = 280$): on the left, according to $r(t)$ and on the right according to $r_d(t)$. The bright areas mark underage periods, where the offered load is higher than the number of beds. The dark areas mark overage periods. The areas of each color are equal in the two figures.

Figure 26: An illustration of the overage and underage periods according to $r(t)$ and $r_d(t)$



We assume that $r_d(t)$ is an invertible function and define t^* to be the intersection point between $r_d(t)$ and N such that $r_d(t^*) = N$; then $t^* = r_d^{-1}(N)$. We can rewrite $C(N)$

to get

$$C(N) = C_u \int_0^{r_d^{-1}(N)} [r_d(t) - N]\mathrm{d}t + C_o \int_{r_d^{-1}(N)}^{T} [N - r_d(t)]\mathrm{d}t. \tag{66}$$

Now assume that $r_d^{-1}(N)$ is a continuous differential function and differentiate Equation (66) according to Leibniz's differentiation rule:

$$\dot{C}(N) = C_o(T - r_d^{-1}(N)) - C_u r_d^{-1}(N) = -(C_o + C_u)r_d^{-1}(N) + C_o T.$$

Since $C(N)$ approaches ∞ as N approaches ∞ and achieves a high positive value for $N = 0$, we minimize $C(N)$ by equating the derivative to 0. This gives rise to

$$r_d^{-1}(N) = \frac{C_o T}{C_o + C_u}.$$

Applying r_d to both sides yields the optimal N^* in Equation (16).
Since C_o and C_u are non-negative numbers and $r_d^{-1}(N)$ is decreasing in N, $\ddot{C}(N)$ is monotonically non-decreasing, and therefore, $C(N)$ is convex and N^* in Equation (16) minimizes $C(N)$.

E Proof of Theorem 2.2

In our proof, we use the following proposition, which is proved in Appendix G:

Proposition E.1. *$C(N)$ in (14) is a convex function.*

We solve problem (21) for the case where $N \leq B$, and for the case where $N \geq B$. Then, we choose the solution which minimizes the overall cost. The option for $N = B$ is included in both cases since their solutions are identical.
Step 1: Find N_k^1, the optimal number of beds if *no* new beds are added, by solving $C_K(N)$ for $N \leq B$.
Since $C(N)$ is a convex function, if the optimal solution for the unconstrained problem is in the allowed range (i.e., $N^* \leq B$), then this will be the solution for the constraint

problem as well. If not, the solution will be at the edge of the range. Formally:

$$N_k^1 = \begin{cases} r_d\left(\dfrac{C_oT}{C_o+C_u}\right), & r_d\left(\dfrac{C_oT}{C_o+C_u}\right) \leq B \\ B, & \textbf{otherwise}. \end{cases}$$

Step 2: Find N_k^2, the optimal number of beds, where $(N-B)$ new beds are added, by solving $C_K(N)$ for $N \geq B$, as follows:

$$\begin{aligned} \underset{N}{\text{minimize}} \quad & C(N) + K(N-B) \\ \text{subject to} \quad & -N + B \leq 0. \end{aligned} \tag{67}$$

Since the objective function remains convex, we solve the unconstrained problem and check whether the solution is in the allowed range. For this, we use the following statement:

The optimal solution, which minimizes the unconstrained problem

$$C_K^{(u)}(N) = C(N) + K(N-B), \tag{68}$$

is given by

$$N_K^{(u)*} = r_d\left(\frac{C_oT+K}{C_o+C_u}\right). \tag{69}$$

This is because the function $C_K^{(u)}(N)$ in (68) can be written in the same structure as in (63) for

$$C = \frac{C_oT+K}{C_o+C_u}. \tag{70}$$

In order to justify the introduction of new beds, we must have $K \leq TC_u$, and therefore, $0 \leq C \leq T$. Since $0 \leq f(x) \leq T$, $f(x)$ crosses C and the proof in Theorem 2.1 holds. The optimal solution for (68) is $N_K^{(u)*} = r_d(C)$, as in (69).

The solution for (67) is, therefore,

$$N_k^2 = \begin{cases} r_d\left(\dfrac{C_oT+K}{C_o+C_u}\right), & r_d\left(\dfrac{C_oT+K}{C_o+C_u}\right) \geq B \\ B, & \textit{otherwise}. \end{cases}$$

Step 3: Combining the results of Steps 1 and 2, yields the solution in Equation (22).

F Proof of Theorem 2.3

We begin by considering the two cases for (23). Each case yields two separable problems, as follows:

1. When $N_{\mathcal{I}} < N_{\mathcal{T}\setminus\mathcal{I}}$, the two problems are to minimize

$$(1)\quad C(N_{\mathcal{I}}) - C_r \cdot N_{\mathcal{I}} = \int_{\mathcal{I}} \Big[C_u\big(r(t) - N_{\mathcal{I}}\big)^+ + C_o\big(N_{\mathcal{I}} - r(t)\big)^+\Big]\,\mathrm{d}t - C_r \cdot N_{\mathcal{I}},$$

$$(2)\quad C(N_{\mathcal{T}\setminus\mathcal{I}}) + C_r \cdot N_{\mathcal{T}\setminus\mathcal{I}} = \int_{\mathcal{T}\setminus\mathcal{I}} \Big[C_u\big(r(t) - N_{\mathcal{T}\setminus\mathcal{I}}\big)^+ + C_o\big(N_{\mathcal{T}\setminus\mathcal{I}} - r(t)\big)^+\Big]\,\mathrm{d}t + C_r \cdot N_{\mathcal{T}\setminus\mathcal{I}}.$$

2. When $N_{\mathcal{I}} > N_{\mathcal{T}\setminus\mathcal{I}}$

$$(1)\quad C(N_{\mathcal{I}}) + C_r \cdot N_{\mathcal{I}} = \int_{\mathcal{I}} \Big[C_u\big(r(t) - N_{\mathcal{I}}\big)^+ + C_o\big(N_{\mathcal{I}} - r(t)\big)^+\Big]\,\mathrm{d}t + C_r \cdot N_{\mathcal{I}},$$

$$(2)\quad C(N_{\mathcal{T}\setminus\mathcal{I}}) - C_r \cdot N_{\mathcal{T}\setminus\mathcal{I}} = \int_{\mathcal{T}\setminus\mathcal{I}} \Big[C_u\big(r(t) - N_{\mathcal{T}\setminus\mathcal{I}}\big)^+ + C_o\big(N_{\mathcal{T}\setminus\mathcal{I}} - r_{\mathcal{T}\setminus\mathcal{I}}(t)\big)^+\Big]\,\mathrm{d}t - C_r \cdot N_{\mathcal{T}\setminus\mathcal{I}}.$$

Since $r_{\mathcal{I}}(t)$ and $r_{\mathcal{T}\setminus\mathcal{I}}(t)$ are non-negative and measurable on the intervals $\mathcal{I}$ and $\mathcal{T}\setminus\mathcal{I}$, respectively (see Hardy et al., 1952), implementing the results from Theorems 2.1 and 2.2 yields the following:

1. When $N_-^{\mathcal{I}} < N_+^{\mathcal{T}\setminus\mathcal{I}}$, then $N_{\mathcal{I}}^* = N_-^{\mathcal{I}}$ and $N_{\mathcal{T}\setminus\mathcal{I}}^* = N_+^{\mathcal{T}\setminus\mathcal{I}}$.
2. When $N_+^{\mathcal{I}} > N_-^{\mathcal{T}\setminus\mathcal{I}}$, then $N_{\mathcal{I}}^* = N_+^{\mathcal{I}}$ and $N_{\mathcal{T}\setminus\mathcal{I}}^* = N_-^{\mathcal{T}\setminus\mathcal{I}}$.

The two cases are mutually exclusive, since $N_-^{\mathcal{I}} \geq N_+^{\mathcal{I}}$ and $N_-^{\mathcal{T}\setminus\mathcal{I}} \geq N_+^{\mathcal{T}\setminus\mathcal{I}}$. When neither of the two conditions prevail, it is preferable to not reallocate beds throughout the planning horizon. Combining these options yields the solution in (24).

G Proof of Proposition E.1

It is sufficient to prove that $F(N)$ in (65) is a concave function. According to *Sierpinski's Theorem* (see Donoghue, 1969), a midpoint concave function that is continuous is, in fact, concave. Since the function $F(N)$ is an integral of N, and therefore, continuous, it is sufficient to prove that it is midpoint concave. Without loss of generality,

it suffices to prove midpoint concavity by proving that for every $N \geq 0$,

$$F(N/2) \geq \frac{F(N)}{2}.$$

In other words, we need to prove that

$$2\int_0^{N/2}[f(x)-C]\mathrm{d}x \geq \int_0^{N}[f(x)-C]\mathrm{d}x,$$

which is equivalent to proving that

$$2\int_0^{N/2} f(x)\mathrm{d}x \geq \int_0^{N} f(x)\mathrm{d}x.$$

Since f is a non-increasing non-negative function, we must have

$$2\int_0^{N/2} f(x)\mathrm{d}x \geq \int_0^{N/2} f(x)\mathrm{d}x + \int_{N/2}^{N} f(x)\mathrm{d}x = \int_0^{N} f(x)\mathrm{d}x,$$

which completes the proof.

H Proof of Theorem 3.1

Let T be an arbitrary positive constant. Using the Lipschitz property (Appendix J) and subtracting the equation for r in (34) from the equation for r^η in (33) yields that

$$\begin{aligned}
\|r_1^\eta - r_1\|_T \vee \|r_2^\eta - r_2\|_T \leq G\Bigg[& |r_1^\eta(0)-r_1(0)| + \left\| \int_0^{\cdot} \lambda(u)\,\mathrm{d}u - \eta^{-1}A^\eta(\cdot)\right\|_T \\
& + \Bigg\| \eta^{-1}D_1\left(\eta p\mu_1 \int_0^{\cdot}\left[\left(N_1+H-r_1^\eta(u)\right)\wedge\left(N_1-b^\eta(u)\right)\right]\mathrm{d}u\right) \\
& - p\mu_1\int_0^{\cdot}\left[(N_1+H-r_1^\eta(u))\wedge(N_1-b^\eta(u))\right]\mathrm{d}u\Bigg\|_T \\
& + \left\| \eta^{-1}D_3\left(\eta(1-p)\mu_1\int_0^{\cdot}\left[(N_1+H-r_1^\eta(u))\wedge(N_1-b^\eta(u))\right]\mathrm{d}u\right)\right. \\
& \left. - (1-p)\mu_1\int_0^{\cdot}\left[\left(N_1+H-r_1^\eta(u)\right)\wedge(N_1-b^\eta(u))\right]\mathrm{d}u\right\|_T \\
& + \left\| \mu_1\int_0^{\cdot}\left[\left(N_1+H-r_1^\eta(u)\right)\wedge(N_1-b^\eta(u)) - (N_1+H-r_1(u))\wedge\left(N_1-b(u)\right)\right]\mathrm{d}u\right\|_T\Bigg] \vee
\end{aligned} \tag{71}$$

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

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 K.1 in Appendix K. By the FSLLN for Poisson processes,

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

Note that the functions $p\mu_1 \int_0^t \left[(N_1 + H - r_1^\eta(u)) \wedge (N_1 - b^\eta(u)) \right] \mathrm{d}u$ 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 K.1, implies that the third, fourth, eighth and ninth terms in (71) converge to 0.

We get that

$$
\begin{aligned}
& \|r_1^\eta - r_1\|_T \vee \|r_2^\eta - r_2\|_T \le \qquad (72) \\
& \left[\epsilon_1^\eta(T) + G\mu_1 \left\| \int_0^\cdot \left[(N_1 + H - r_1^\eta(u)) \wedge (N_1 - b^\eta(u)) - (N_1 + H - r_1(u)) \wedge (N_1 - b(u)) \right] \mathrm{d}u \right\|_T \right] \vee \\
& \Bigg[\epsilon_2^\eta(T) + G(1-p)\mu_1 \left\| \int_0^\cdot \left[(N_1 + H - r_1^\eta(u)) \wedge (N_1 - b^\eta(u)) - (N_1 + H - r_1(u)) \wedge (N_1 - b(u)) \right] \mathrm{d}u \right\|_T \\
& \quad + G\mu_2 \left\| \int_0^\cdot \left[N_2 \wedge (r_1^\eta(u) - r_2^\eta(u) + N_2) \right] - \left[N_2 \wedge (r_1(u) - r_2(u) + N_2) \right] \mathrm{d}u \right\|_T \Bigg]
\end{aligned}
$$

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

where $\epsilon_1^\eta(T)$ bounds the sum of the first four terms on the right-hand side of (71), 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 (72) is obtained by using the inequalities $|a \wedge b - a \wedge c| \leq |b - c|$ and $|a \wedge b - c \wedge d| \leq |a - c| + |b - d|$ for any a, b, c and d. The third equality in (72) is because $0 \leq p \leq 1$.

We now use

$$
\begin{aligned}
\int_0^T \|b^\eta - b\|_u\ \mathrm{d}u &= \int_0^T \left\|(r_1^\eta - r_2^\eta)^+ - (r_1 - r_2)^+\right\|_u\ \mathrm{d}u \\
&= \int_0^T \|r_1^\eta - r_1^\eta \wedge r_2^\eta - r_1 + r_1 \wedge r_2\|_u\ \mathrm{d}u \\
&\leq \int_0^T \left[\, \|r_1^\eta - r_1\|_u + \|r_1^\eta \wedge r_2^\eta - r_1 \wedge r_2\|_u \right] \mathrm{d}u \\
&\leq \int_0^T \left[2\,\|r_1^\eta - r_1\|_u + \|r_2^\eta - r_2\|_u \right] \mathrm{d}u \\
&= 2\int_0^T \|r_1^\eta - r_1\|_u\ \mathrm{d}u + \int_0^T \|r_2^\eta - r_2\|_u\ \mathrm{d}u.
\end{aligned} \tag{73}
$$

From (72) and (73), we get that

$$
\begin{aligned}
&\|r_1^\eta - r_1\|_T \vee \|r_2^\eta - r_2\|_T \\
&\leq [\epsilon_1^\eta(T) \vee \epsilon_2^\eta(T)] + G\,(3\mu_1 + \mu_2) \int_0^T \|r_1^\eta - r_1\|_u\ \mathrm{d}u + G\,(\mu_1 \vee \mu_2) \int_0^T \|r_2^\eta - r_2\|_u\ \mathrm{d}u \\
&\leq [\epsilon_1^\eta(T) \vee \epsilon_2^\eta(T)] + 2G\,(3\mu_1 \vee \mu_2) \left[\int_0^T \|r_1^\eta - r_1\|_u\ \mathrm{d}u + \int_0^T \|r_2^\eta - r_2\|_u\ \mathrm{d}u \right]
\end{aligned} \tag{74}
$$

$$\leq [\epsilon_1^\eta(T) \vee \epsilon_2^\eta(T)] + 4G\left(3\mu_1 \vee \mu_2\right)\left[\int_0^T \|r_1^\eta - r_1\|_u \, \mathrm{d}u \vee \int_0^T \|r_2^\eta - r_2\|_u \, \mathrm{d}u\right]$$
$$\leq [\epsilon_1^\eta(T) \vee \epsilon_2^\eta(T)] + 4G\left(3\mu_1 \vee \mu_2\right)\left[\int_0^T \|r_1^\eta - r_1\|_u \vee \|r_2^\eta - r_2\|_u \, \mathrm{d}u\right].$$

The first equality in (74) is obtained by using the inequality $(a+b)\vee(c+d) \leq a\vee c+b\vee d$, for any a, b, c and d. Applying Gronwall's inequality (Ethier and Kurtz, 2009) to (74) completes the proof for both the existence and uniqueness of r.

I Proof of Proposition 3.1

We begin by proving that the solution for (35) satisfies, for $t \geq 0$,

$$\begin{aligned} l(t) = &\int_0^t 1_{\{x_1(u) \geq N_1+H\}} \cdot 1_{\{x_1(u)+x_2(u)<N_1+N_2+H\}} \left[\lambda(u) - l_1(u)\right]^+ \mathrm{d}u \\ &+ \int_0^t 1_{\{x_1(u)< N_1+H\}} \cdot 1_{\{x_1(u)+x_2(u)\geq N_1+N_2+H\}} \left[\lambda(u) - l_2(u)\right]^+ \mathrm{d}u \\ &+ \int_0^t 1_{\left\{x_1(u)\geq N_1+H\right\}} \cdot 1_{\left\{x_1(u)+x_2(u)\geq N_1+N_2+H\right\}} \left[\lambda(u) - l_1(u) \wedge l_2(u)\right]^+ \mathrm{d}u, \end{aligned} \tag{75}$$

where

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

In order to prove this, we substitute (75) in (35) and show that the properties in (35) prevail. We begin by substituting (75) in the first line of (35). 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,

$$
\begin{aligned}
x_1(t) = x_1(0) + \int_0^t \Big[& 1_{\{x_1(u) < N_1 + H\}} \cdot 1_{\{x_1(u) + x_2(u) < N_1 + N_2 + H\}} \cdot \lambda(u) \\
& - \mu_1 \left[x_1(u) \wedge (N_1 - b(u)) \right] \Big] \, \mathrm{d}u \\
& + \int_0^t \left[1_{\{x_1(u) \geq N_1 + H\}} \cdot 1_{\{x_1(u) + x_2(u) < N_1 + N_2 + H\}} \cdot (\lambda(u) \wedge l_1(u)) \right] \, \mathrm{d}u \\
& + \int_0^t \left[1_{\{x_1(u) < N_1 + H\}} \cdot 1_{\{x_1(u) + x_2(u) \geq N_1 + N_2 + H\}} \cdot (\lambda(u) \wedge l_2(u)) \right] \, \mathrm{d}u \\
& + \int_0^t \left[1_{\{x_1(u) \geq N_1 + H\}} \cdot 1_{\{x_1(u) + x_2(u) \geq N_1 + N_2 + H\}} \cdot (\lambda(u) \wedge l_1(u) \wedge l_2(u)) \right] \, \mathrm{d}u; \\
x_2(t) = x_2(0) + \int_0^t & \left[p\mu_1 \left[x_1(u) \wedge (N_1 - b(u)) \right] - \mu_2 \left(x_2(u) \wedge N_2 \right) \right] \, \mathrm{d}u.
\end{aligned} \tag{76}
$$

Clearly, the properties in the third and fourth lines in (35) prevail. It is left to verify that the first and second conditions prevail. This is done by the following proposition.

Proposition I.1. *The functions $x_1(\cdot)$ and $x_1(\cdot) + x_2(\cdot)$ as in (76) 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 (76), 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 (76), we get that

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

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 (76), we get that

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

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 J.1), we have established that x, the fluid limit for the stochastic queueing family X^η in (26), is given by (76).

The following two remarks explain why (76) is equivalent to (36):

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 I.1, the indicators in (75) 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, $l_2(u) = l_2^*(u)$.

J 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:

$$\begin{aligned}
\psi(m)(t) &= \sup_{0 \leq s \leq t} \left(-\Big(m_1(s) \wedge m_2(s)\Big)\right)^+; \\
\phi(m)(t) &= m(t) + \psi(m)(t) \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad t \geq 0.
\end{aligned}$$

Proposition J.1. *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 (Harrison, 1985, 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

$$\frac{1}{2}(r_i^*(t) - r_i(t))^2 = \int_0^t (r_i^* - r_i)\,\mathrm{d}l^* + \int_0^t (r_i - r_i^*)\,\mathrm{d}l. \tag{77}$$

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 (77) are non-positive. Since the left side ≥ 0, both sides must be zero, thus $r_1^* = r_1$, $r_2^* = r_2$ and $l^* = l$.

Proposition J.2. *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

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

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

$$\begin{aligned}
\|\psi(m) - \psi(m')\|_T &= \Big\| \sup_{0\leq s\leq \cdot} \Big(-\big(m_1(s) \wedge m_2(s)\big)\Big)^+ - \sup_{0\leq s\leq t} \Big(-\big(m_1'(s) \wedge m_2'(s)\big)\Big)^+ \Big\|_T \\
&\leq \Big\| \sup_{0\leq s\leq \cdot} \big|\big(m_1(s) \wedge m_2(s)\big) - \big(m_1'(s) \wedge m_2'(s)\big)\big| \Big\|_T \qquad (78) \\
&= \big\|\big(m_1 \wedge m_2\big) - \big(m_1' \wedge m_2'\big)\big\|_T \leq 2\Big[\,\|m_1 - m_1'\|_T \vee \|m_2 - m_2'\|_T\,\Big].
\end{aligned}$$

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,

$$\begin{aligned} m_1(t) \wedge m_2(t) &\leq m_1'(t) \wedge m_2'(t) + \|m_1 - m_1'\|_T + \|m_2 - m_2'\|_T\,, \\ m_1(t) \wedge m_2(t) &\geq m_1'(t) \wedge m_2'(t) - \|m_1 - m_1'\|_T - \|m_2 - m_2'\|_T\,, \end{aligned}$$

and

$$|m_1(t) \wedge m_2(t) - m_1'(t) \wedge m_2'(t)| \leq \|m_1 - m_1'\|_T + \|m_2 - m_2'\|_T\,,$$

which yields

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

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

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

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

We begin with the left-hand side:

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

where the last inequality is derived from (78).

K Lemma K.1

Lemma K.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 $|g_\eta(t)| \leq C_T$, 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.

L Proof of Proposition 4.1

From (52), we return to our original formulation in terms of $q(\cdot)$ for $t \geq 0$, as follows:

$$\begin{cases} q_1(t) = q_1(0) + \int_0^t \left[\lambda(u) - \mu_1\left(q_1(u) \wedge N_1 \wedge (H_2 + N_2 - q_2(u))\right)\right] \mathrm{d}u - l(t) \leq H_1 + N_1, \\ q_i(t) = q_i(0) + \int_0^t \Big[\mu_{i-1}\left(q_{i-1}(u) \wedge N_{i-1} \wedge (H_i + N_i - q_i(u))\right) \\ \qquad -\mu_i\left(q_i(u) \wedge N_i \wedge (H_{i+1} + N_{i+1} - q_{i+1}(u))\right)\Big] \mathrm{d}u \leq H_i + N_i, \;\; i = 2, \ldots, k-1; \\ q_k(t) = q_k(0) + \int_0^t \Big[\mu_{k-1}\left(q_{k-1}(u) \wedge N_{k-1} \wedge (H_k + N_k - q_k(u))\right) \\ \qquad -\mu_i\left(q_k(u) \wedge N_k\right)\Big] \mathrm{d}u \leq H_k + N_k, \\ \mathrm{d}l(t) \geq 0, \;\; l(0) = 0, \\ \int_0^\infty 1_{\{q_1(u-) < H_1 + N_1\}} \mathrm{d}l(t) = 0; \end{cases} \tag{79}$$

Now, we prove that the solution for (79) satisfies

$$l(t) = \int_0^t 1_{\{q_1(u) \geq H_1 + N_1\}} \left[\lambda(u) - l_1(u)\right]^+ \mathrm{d}u, \quad t \geq 0, \tag{80}$$

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 (80) in the equation of $q_1(t)$ in (79) and show that the properties in (79) prevail:

$$\begin{aligned} q_1(t) = q_1(0) &+ \int_0^t \left[\lambda(u) - \mu_1\left(q_1(u) \wedge N_1 \wedge (H_2 + N_2 - q_2(u))\right)\right] \mathrm{d}u \\ &- \int_0^t 1_{\{q_1(u) \geq H_1 + N_1\}} \left[\lambda(u) - \lambda(u) \wedge l_1(u)\right] \mathrm{d}u \end{aligned} \tag{81}$$

$$= q_1(0) + \int_0^t \left[1_{\{q_1(u) < H_1+N_1\}} \cdot \lambda(u) - \mu_1 \left(q_1(u) \wedge N_1 \wedge (H_2 + N_2 - q_2(u))\right)\right] \mathrm{d}u$$
$$+ \int_0^t \left[1_{\{q_1(u) \geq H_1+N_1\}} \cdot (\lambda(u) \wedge l_1(u))\right] \mathrm{d}u$$

Clearly, the properties in the last two lines in (79) prevail. It is left to verify that the first k conditions prevail. This is done by the following proposition.

Proposition L.1. *The functions $q_i(\cdot)$, $i = 1, \ldots, k$, as in (81) are bounded by $H_i + N_i$, respectively.*

Proof: First we prove that the function $q_1(\cdot)$, as in (81), 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 (81), we get that

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

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 (79), we get that

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

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 Zychlinski et al. (2018b)), we have established that q, the fluid limit for the stochastic queueing family Q^η in (44), is given by (53). Note that after proving that $q_1(\cdot) \leq H_1 + N_1$ in Proposition L.1, the indicators in (80) can accommodate only the case when $q_1(\cdot) = H_1 + N_1$.

M Proof of Theorem 4.1

Due to the uniqueness of q (Proposition 4.1), it suffices to show that δ and $\bar{q}_i$, $i = 1, \ldots, k$, in Equations (56)–(58) satisfy the model equations in (53). In particular, it suffices to show that the steady-state equations in (55) are satisfied. Since the second equation in (55) 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 (55) yields the following:

$$\lambda = \lambda \cdot 1_{\{\lambda < \mu_1(H_1+N_1)\}} + [\lambda \wedge \mu_1 (N_1 \wedge (H_2 + N_2 - \lambda/\mu_2))] \cdot 1_{\{\lambda = \mu_1(H_1+N_1)\}}. \tag{82}$$

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 (56) we know that $\lambda \leq \mu_1 N_1$. Therefore, the second indicator in (82) 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 (56): when $\delta = \lambda$, we get that $\lambda = \mu_1 N_1 \leq (H_2 + N_2)/(1/\mu_1 + 1/\mu_2)$, which is equivalent to $N_1 \leq H_2 + N_2 - \mu_1 N_1/\mu_1$. Therefore, (82) is satisfied. It is easy to show that the second line in (55) is also satisfied by $\bar{q}_j = \lambda/\mu_j$, $j = 1, \ldots, k$.

Now, when $\delta < \lambda$, from (55) we get that $\bar{q}_1 = H_1 + N_1$ (the first indicator in the first line is zero), and we get that

$$\delta = \lambda \wedge \mu_1 (N_1 \wedge (H_2 + N_2 - \bar{q}_2)) = \mu_1 (N_1 \wedge (H_2 + N_2 - \bar{q}_2)). \tag{83}$$

If Station 1 is the first bottleneck ($i = 1$, in (58)) then, from (54) and (56), we get that $\delta = \mu_1 N_1 \leq \mu_1(H_2 + N_2 - \mu_1 N_1/\mu_2)$; therefore, (83) 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 (54) 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 Equation (83).

For completing the proof for $\bar{q}_i$, $i = 3, \ldots, k$, in (57), 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, $3 \leq i \leq k$, is the first bottleneck. From (54) 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 contra-

dicts 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, $2 \leq i \leq k-1$, is the first bottleneck. From (54) and (55), we get that $\delta = \mu_{i+1}\left[\bar{q}_{i+1} \wedge N_{i+1} \wedge (H_{i+2} + N_{i+2} - \bar{q}_{i+2})\right]$. 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 $\delta \leq \mu_{i+1}N_{i+1}$, as well as $\delta \leq (H_{i+2}+N_{i+2})/(1/\mu_{i+1}+1/\mu_{i+2})$, which is equivalent to $\delta \leq \mu_{i+1}(H_{i+2}+N_{i+2}-\delta/\mu_{i+2})$. Hence, (55) is satisfied. We iteratively continue this argument up until Station k.

References

Afèche, P., Araghi, M., and Baron, O. (2017). Customer acquisition, retention, and queueing-related service quality: Optimal advertising, staffing, and priorities for a call center. *Manufacturing and Service Operations Management*, 19(4):674–691. 14, 37

Akcali, E., Coˆté, M., and Lin, C. (2006). A network flow approach to optimizing hospital bed capacity decisions. *Health Care Management Science*, 9(4):391–404. 13

Akyildiz, I. and von Brand, H. (1994). Exact solutions for networks of queues with blocking-after-service. *Theoretical Computer Science*, 125(1):111–130. 36

Altiok, T. (1982). Approximate analysis of exponential tandem queues with blocking. *European Journal of Operational Research*, 11(4):390–398. 63

Arendt, K., Sadosty, A., Weaver, A., Brent, C., and Boie, E. (2003). The left-without-being-seen patients: what would keep them from leaving? *Annals of Emergency Medicine*, 42(3):317–IN2. 48

Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y., Tseytlin, Y., and Yom-Tov, G. (2015). On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems*, 5(1):146–194. 11, 35, 79

Arrow, K., Harris, T., and Marschak, J. (1951). Optimal inventory policy. *Econometrica: Journal of the Econometric Society*, pages 250–272. 23

Ata, B., Killaly, B., Olsen, T., and Parker, R. (2013). On hospice operations under medicare reimbursement policies. *Management Science*, 59(5):1027–1044. 13, 14

Avi-Itzhak, B. (1965). A sequence of service stations with arbitrary input and regular service times. *Management Science*, 11(5):565–571. 34, 36, 58, 64

Avi-Itzhak, B. and Halfin, S. (1993). Servers in tandem with communication and manufacturing blocking. *Journal of Applied Probability*, pages 429–437. 64

Avi-Itzhak, B. and Levy, H. (1995). A sequence of servers with arbitrary input and regular service times revisited: in memory of Micha Yadin. *Management Science*, 41(6):1039–1047. 34, 36, 64

Avi-Itzhak, B. and Yadin, M. (1965). A sequence of two servers with no intermediate queue. *Management Science*, 11(5):553–564. 36, 58, 64

Baker, D., Stevens, C., and Brook, R. (1991). Patients who leave a public hospital emergency department without being seen by a physician: causes and consequences. *JAMA*, 266(8):1085–1090. 48

Balsamo, S. and de Nitto Personè, V. (1994). A survey of product form queueing networks with blocking and their equivalences. *Annals of Operations research*, 48(1):31–61. 36

Balsamo, S., de Nitto Personé, V., and Onvural, R. (2001). *Analysis of Queueing Networks with Blocking*. Springer. 11, 16, 35, 40, 62, 63, 64, 72, 83

Bassamboo, A., Harrison, J., and Zeevi, A. (2006). Design and control of a large call center: Asymptotic analysis of an lp-based method. *Operations Research*, 54(3):419–435. 14

BBC News (2016). Hospital bed-blocking costs NHS England £900m a year. http://www.bbc.com/news/health-35481849. 4

Bekker, R. and de Bruin, A. (2010). Time-dependent analysis for refused admissions in clinical wards. *Annals of Operations Research*, 178(1):45–65. 12, 79

Bickel, P., Chen, C., Kwon, J., Rice, J., Varaiya, P., and van Zwet, E. (2003). Traffic flow on a freeway network. In *Nonlinear Estimation and Classification*, pages 63–81. Springer. 62

Borisov, I. and Borovkov, A. (1981). Asymptotic behavior of the number of free servers for systems with refusals. *Theory of Probability & Its Applications*, 25(3):439–453. 37

Borovkov, A. (2012). *Stochastic Processes in Queueing Theory*. Springer Science & Business Media. 37

Brandwajn, A. and Jow, Y. (1988). An approximation method for tandem queues with blocking. *Operations Research*, 36(1):73–83. 36, 64

Bretthauer, K., Heese, H., Pun, H., and Coe, E. (2011). Blocking in healthcare operations: A new heuristic and an application. *Production and Operations Management*, 20(3):375–391. 12, 36

Buzacott, J. and Shanthikumar, J. (1993). *Stochastic Models of Manufacturing Systems*. Prentice Hall Englewood Cliffs, NJ. 35

Caplan, G., Sulaiman, N., Mangin, D., Ricauda, N., Wilson, A., and Barclay, L. (2012). A meta-analysis of "hospital in the home". *The Medical Journal of Australia*, 197(9):512–519. 33

Cheah, J. and Smith, J. (1994). Generalized M/G/C/C state dependent queueing models and pedestrian traffic flows. *Queueing Systems*, 15(1-4):365–386. 12

Chen, H. and Yao, D. (2013). *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*. Springer Science & Business Media. 38

Cochran, J. and Bharti, A. (2006). Stochastic bed balancing of an obstetrics hospital. *Health Care Management Science*, 9(1):31–45. 4

Cohen, I., Mandelbaum, A., and Zychlinski, N. (2014). Minimizing mortality in a mass casualty event: fluid networks in support of modeling and staffing. *IIE Transactions*, 46(7):728–741. 13, 16, 35, 37

Conway, R., Maxwell, W., McClain, J., and Thomas, L. (1988). The role of work-in-process inventory in serial production lines. *Operations Research*, 36(2):229–241. 36, 37, 59

Daganzo, C., Gayah, V., and Gonzales, E. (2012). The potential of parsimonious models for understanding large scale transportation systems and answering big picture questions. *EURO Journal on Transportation and Logistics*, 1(1-2):47–65. 62

Dai, J. and Vande Vate, J. (2000). The stability of two-station multitype fluid networks. *Operations Research*, 48(5):721–744. 72

Dallery, Y. and Frein, Y. (1993). On decomposition methods for tandem queueing networks with blocking. *Operations Research*, 41(2):386–399. 64

Dallery, Y. and Gershwin, S. (1992). Manufacturing flow line systems: A review of models and analytical results. *Queueing Systems*, 12(1-2):3–94. 36

Dallery, Y., Liu, Z., and Towsley, D. (1994). Equivalence, reversibility, symmetry and concavity properties in fork-join queuing networks with blocking. *Journal of the ACM (JACM)*, 41(5):903–942. 77

Dallery, Y., Liu, Z., and Towsley, D. (1997). Properties of fork/join queueing networks with blocking under various operating mechanisms. *IEEE Transactions on Robotics and Automation*, 13(4):503–518. 77

De Bruin, A., Van Rossum, A., Visser, M., and Koole, G. (2007). Modeling the emergency cardiac in-patient flow: an application of queuing theory. *Health Care Management Science*, 10(2):125–137. 62

De Vries, T. and Beekman, R. (1998). Applying simple dynamic modelling for decision support in planning regional health care. *European Journal of Operational Research*, 105(2):277–284. 13

Desai, M., Penn, M., Brailsford, S., and Chipulu, M. (2008). Modelling of Hampshire adult services – gearing up for future demands. *Health Care Management Science*, 11(2):167–176. 11

Desel, J. and Silva, M. (1998). *Application and Theory of Petri Nets 1998: 19th International Conference, ICATPN'98, Lisbon, Portugal, June 22–26, 1998 Proceedings*. Springer. 63

Dogan-Sahiner, E. and Altiok, T. (1998). Blocking policies in pharmaceutical transfer lines. *Annals of Operations Research*, 79:323–347. 62

Donoghue, W. (1969). *Distributions and Fourier Transforms*. Academic Press. 89

Eick, S., Massey, W., and Whitt, W. (1993). $M_t/G/\infty$ queues with sinusoidal arrival rates. *Management Science*, 39(2):241–252. 54, 55

El-Darzi, E., Vasilakis, C., Chaussalet, T., and Millard, P. (1998). A simulation modelling approach to evaluating length of stay, occupancy, emptiness and bed blocking in a hospital geriatric department. *Health Care Management Science*, 1(2):143–149. 4, 11, 36

Ethier, S. and Kurtz, T. (2009). *Markov Processes: Characterization and Convergence*. John Wiley & Sons. 93

Faddy, M., Graves, N., and Pettitt, A. (2009). Modeling length of stay in hospital and other right skewed data: Comparison of phase-type, gamma and log-normal distributions. *Value in Health*, 12(2):309–314. 79

Faddy, M. and McClean, S. (2005). Markov chain modelling for geriatric patient care. *Methods of Information in Medicine-Methodik der Information in der Medizin*, 44(3):369–373. 10

Feldman, Z., Mandelbaum, A., Massey, W., and Whitt, W. (2008). Staffing of time-varying queues to achieve time-stable performance. *Management Science*, 54(2):324–338. 35

Filippov, A. (2013). *Differential Equations with Discontinuous Righthand Sides: Control Systems*. Springer Science & Business Media. 35, 38, 63

Frein, Y. and Dallery, Y. (1989). Analysis of cyclic queueing networks with finite buffers and blocking before service. *Performance Evaluation*, 10(3):197–210. 62, 64

Garnett, O., Mandelbaum, A., and Reiman, M. (2002). Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, 4(3):208–227. 12, 38

Gershwin, S. (1987). An efficient decomposition method for the approximate evaluation of tandem queues with finite storage space and blocking. *Operations Research*, 35(2):291–305. 12, 36, 64

Glynn, P. and Whitt, W. (1991). Departures from many queues in series. *Annals of Applied Probability*, 1(4):546–572. 37

Grassmann, W. and Drekic, S. (2000). An analytical solution for a tandem queue with blocking. *Queueing Systems*, 36(1-3):221–235. 36

Gray, L., Broe, G., Duckett, S., Gibson, D., Travers, C., and McDonnell, G. (2006). Developing a policy simulator at the acute-aged care interface. *Australian Health Review*, 30(4):450–457. 11

Green, L. (2004). Capacity planning and management in hospitals. In *Operations Research and Health Care*, pages 15–41. Springer. 13

Green, L., Kolesar, P., and Whitt, W. (2007a). Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management*, 16(1):13–39. 13

Green, L., Kolesar, P., and Whitt, W. (2007b). Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management*, 16(1):13–39. 35, 37

Hardy, G., Littlewood, J., and Pólya, G. (1952). *Inequalities*. Cambridge University Press. 22, 89

Hare, W., Alimadad, A., Dodd, H., Ferguson, R., and Rutherford, A. (2009). A deterministic model of home and community care client counts in British Columbia. *Health Care Management Science*, 12(1):80–98. 13

Harrison, G. and Millard, P. (1991). Balancing acute and long-term care: the mathematics of throughput in departments of geriatric medicine. *Methods of Information in Medicine*, 30(3):221. 10

Harrison, J. (1973). Assembly-like queues. *Journal of Applied Probability*, 10(02):354–367. 38

Harrison, J. (1985). *Brownian Motion and Stochastic Flow Systems*. Wiley New York. 38, 96

Harrison, J. and Zeevi, A. (2005). A method for staffing large call centers based on stochastic fluid models. *Manufacturing & Service Operations Management*, 7(1):20–36. 14, 23

He, B., Liu, Y., and Whitt, W. (2016). Staffing a service system with non-Poissin non-stationary arrivals. *Probability in the Engineering and Informational Sciences*, 30(4):593–621. 48

Hillier, F. and Boling, R. (1967). Finite queues in series with exponential or erlang service times—a numerical approach. *Operations Research*, 15(2):286–303. 12, 59

Jennings, O., Massey, W., and McCalla, C. (1997). Optimal profit for leased lines services. In *Proceedings of the 15th International Teletraffic Congress-ITC*, volume 15, pages 803–814. 14, 20, 22

Kao, E. and Tung, G. (1981). Bed allocation in a public health care delivery system. *Management Science*, 27(5):507–520. 14

Katsaliaki, K., Brailsford, S., Browning, D., and Knight, P. (2005). Mapping care pathways for the elderly. *Journal of Health Organization and Management*, 19(1):57–72. 11, 36

Kelly, F. (1984). Blocking, reordering, and the throughput of a series of servers. *Stochastic Processes and Their Applications*, 17(2):327–336. 36

Kerbache, L. and MacGregor Smith, J. (1987). The generalized expansion method for open finite queueing networks. *European Journal of Operational Research*, 32(3):448–461. 12

Kerbache, L. and MacGregor Smith, J. (1988). Asymptotic behavior of the expansion method for open finite queueing networks. *Computers & Operations Research*, 15(2):157–169. 12

Koizumi, N., Kuno, E., and Smith, T. (2005). Modeling patient flows using a queuing network with blocking. *Health Care Management Science*, 8(1):49–60. 4, 12

Langaris, C. and Conolly, B. (1984). On the waiting time of a two-stage queueing system with blocking. *Journal of Applied Probability*, 21(03):628–638. 36

Leachman, R. and Gascon, A. (1988). A heuristic scheduling policy for multi-item, single-machine production systems with time-varying, stochastic demands. *Management Science*, 34(3):377–390. 35

Li, A. and Whitt, W. (2014). Approximate blocking probabilities in loss models with independence and distribution assumptions relaxed. *Performance Evaluation*, 80:82–101. 37

Li, A., Whitt, W., and Zhao, J. (2015). Staffing to stabilize blocking in loss models with time-varying arrival rates. *Probability in the Engineering and Informational Sciences*, pages 1–27. 13

Li, A., Whitt, W., and Zhao, J. (2016). Staffing to stabilize blocking in loss models with time-varying arrival rates. *Probability in the Engineering and Informational Sciences*, 30(02):185–211. 38

Li, J. and Meerkov, S. (2009). *Production Systems Engineering*. Springer Science & Business Media. 34

Liu, Y. and Whitt, W. (2011a). Large-time asymptotics for the $G_t/M_t/s_t + GI_t$ many-server fluid queue with abandonment. *Queueing Systems*, 67(2):145–182. 34, 37, 63, 84

Liu, Y. and Whitt, W. (2011b). A network of time-varying many-server fluid queues with customer abandonment. *Operations Research*, 59(4):835–846. 13, 37

Liu, Y. and Whitt, W. (2012a). The $G_t/GI/s_t + GI$ many-server fluid queue. *Queueing Systems*, 71(4):405–444. 37, 48, 84

Liu, Y. and Whitt, W. (2012b). A many-server fluid limit for the $G_t/GI/s_t + GI$ queueing model experiencing periods of overloading. *Operations Research Letters*, 40(5):307–312. 37

Liu, Y. and Whitt, W. (2014). Many-server heavy-traffic limit for queues with time-varying parameters. *Annals of Applied Probability*, 24(1):378–421. 34, 37, 63, 84

Ma, N. and Whitt, W. (2016). Efficient simulation of non-Poisson non-stationary point processes to study queueing approximations. *Statistics & Probability Letters*, 109:202–207. 48

Mandelbaum, A., Massey, W., and Reiman, M. (1998). Strong approximations for Markovian service networks. *Queueing Systems*, 30(1-2):149–201. 13, 34, 37, 63, 84

Mandelbaum, A., Massey, W., Reiman, M., and Rider, B. (1999). Time varying multiserver queues with abandonment and retrials. In *Proceedings of the 16th International Teletraffic Conference*, volume 4, pages 4–7. 13, 16, 34, 37, 48, 63

Mandelbaum, A. and Pats, G. (1995). State-dependent queues: approximations and applications. *Stochastic Networks*, 71:239–282. 38, 45, 67

Mandelbaum, A. and Pats, G. (1998). State-dependent stochastic networks. part i. approximations and applications with continuous diffusion limits. *Annals of Applied Probability*, 8(2):569–646. 38, 41

Marazzi, A., Paccaud, F., Ruffieux, C., and Beguin, C. (1998). Fitting the distributions of length of stay by parametric models. *Medical Care*, 36(6):915–927. 79

Martin, J. (2002). Large tandem queueing networks with blocking. *Queueing Systems*, 41(1-2):45–72. 37

McCalla, C. and Whitt, W. (2002). A time-dependent queueing-network model to describe the life-cycle dynamics of private-line telecommunication services. *Telecommunication Systems*, 19(1):9–38. 13

McClean, S. and Millard, P. (1993). Patterns of length of stay after admission in geriatric medicine: an event history approach. *The Statistician*, pages 263–274. 79

McClean, S. and Millard, P. (2006). Where to treat the older patient? Can Markov models help us better understand the relationship between hospital and community care? *Journal of the Operational Research Society*, 58(2):255–261. 10, 79

Meerkov, S. and Yan, C.-B. (2016). Production lead time in serial lines: Evaluation, analysis, and control. *IEEE Transactions on Automation Science and Engineering*, 13(2):663–675. 34, 37

Millhiser, W. and Burnetas, A. (2013). Optimal admission control in series production systems with blocking. *IIE Transactions*, 45(10):1035–1047. 36

Nahmias, S. and Cheng, Y. (2009). *Production and Operations Analysis*. McGraw-Hill New York. 23, 35

Namdaran, F., Burnet, C., and Munroe, S. (1992). Bed blocking in Edinburgh hospitals. *Health Bulletin*, 50(3):223–227. 4

NHS England - Bed Availability and Occupancy Data (2015). https://www.england.nhs.uk/statistics/statistical-work-areas/bed-availability-and-occupancy/bed-data-overnight/. 4

OECD iLibrary - Health at a Glance (2013). http://www.oecd-ilibrary.org/sites/health_glance-2013-en/04/03/index.html?itemId=/content/chapter/health_glance-2013-34-en. 4

Olivares, M., Terwiesch, C., and Cassorla, L. (2008). Structural estimation of the newsvendor model: an application to reserving operating room time. *Management Science*, 54(1):41–55. 26

Oliver, R. and Samuel, A. (1962). Reducing letter delays in post offices. *Operations Research*, 10(6):839–892. 37

Onvural, R. and Perros, H. (1989). Approximate throughput analysis of cyclic queueing networks with finite buffers. *IEEE Transactions on Software Engineering*, 15(6):800–808. 64

Osorio, C. and Bierlaire, M. (2009). An analytic finite capacity queueing network model capturing the propagation of congestion and blocking. *European Journal of Operational Research*, 196(3):996–1007. 4, 12, 36, 64

Pang, G. and Whitt, W. (2009). Heavy-traffic limits for many-server queues with service interruptions. *Queueing Systems*, 61(2):167–202. 34, 63

Pender, J. (2015). Nonstationary loss queues via cumulant moment approximations. *Probability in the Engineering and Informational Sciences*, 29(1):27–49. 48

Pender, J. and Ko, Y. (2017). Approximations for the queue length distributions of time-varying many-server queues. *INFORMS Journal on Computing*, 29(4):688–704. 48

Perros, H. (1994). *Queueing Networks with Blocking*. Oxford University Press, Inc. 11, 36, 62, 63

Porteus, E. (2002). *Foundations of Stochastic Inventory Theory*. Stanford University Press. 19

Prabhu, N. (1967). Transient behaviour of a tandem queue. *Management Science*, 13(9):631–639. 36

Reed, J., Ward, A., and Zhan, D. (2013). On the generalized drift Skorokhod problem in one dimension. *Journal of Applied Probability*, 50(1):16–28. 38

Rohleder, T., Cooke, D., Rogers, P., and Egginton, J. (2013). Coordinating health services: An operations management perspective. In *Handbook of Healthcare Operations Management*, pages 421–445. Springer. 11

Rubin, S. and Davies, G. (1975). Bed blocking by elderly patients in general-hospital wards. *Age and Ageing*, 4(3):142–147. 4

Seo, D.-W., Lee, H.-C., and Ko, S.-S. (2008). Stationary waiting times in m-node tandem queues with communication blocking. *Management Science and Financial Engineering*, 14(1):23–34. 62

Shepperd, S., Doll, H., Angus, R., Clarke, M., Iliffe, S., Kalra, L., Ricauda, N., and Wilson, A. (2008). Admission avoidance hospital at home. 33

Shi, P., Chou, M., Dai, J., Ding, D., and Sim, J. (2015). Models and insights for hospital inpatient operations: Time-dependent ED boarding time. *Management Science*, 62(1):1–28. 4, 11, 79

Srikant, R. and Whitt, W. (1996). Simulation run lengths to estimate blocking probabilities. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 6(1):7–52. 37

Suri, R. and Diehl, G. (1984). A new 'building block' for performance evaluation of queueing networks with finite buffers. In *ACM SIGMETRICS Performance Evaluation Review*, volume 12, pages 134–142. ACM. 62, 64

Takahashi, Y., Miyahara, H., and Hasegawa, T. (1980). An approximation method for open restricted queueing networks. *Operations Research*, 28(3-part-i):594–602. 12, 36

Taylor, G., McClean, S., and Millard, P. (1997). Continuous-time Markov models for geriatric patient behaviour. *Applied Stochastic Models and Data Analysis*, 13(3-4):315–323. 10

Taylor, G., McClean, S., and Millard, P. (2000). Stochastic models of geriatric patient bed occupancy behaviour. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 163(1):39–48. 10

Ticona, L. and Schulman, K. (2016). Extreme home makeover—the role of intensive home health care. *New England Journal of Medicine*, 375(18):1707–1709. 32

Tolio, T. and Gershwin, S. (1998). Throughput estimation in cyclic queueing networks with blocking. *Annals of Operations Research*, 79:207–229. 36

Travers, C., McDonnell, G., Broe, G., Anderson, P., Karmel, R., Duckett, S., and Gray, L. (2008). The acute-aged care interface: Exploring the dynamics of 'bed blocking'. *Australasian Journal on Ageing*, 27(3):116–120. 4, 11

United Nations Population Fund (2014). http://www.unfpa.org/ageing. 4

van Vuuren, M., Adan, I., and Resing-Sassen, S. (2005). Performance analysis of multi-server tandem queues with finite buffers and blocking. *OR Spectrum*, 27(2-3):315–338. 36, 64

Vandergraft, J. (1983). A fluid flow model of networks of queues. *Management Science*, 29(10):1198–1208. 37

Wenocur, M. (1982). A production network model and its diffusion approximation. Technical report, DTIC Document. 38

Whitt, W. (1985). The best order for queues in series. *Management Science*, 31(4):475–487. 36

Whitt, W. (2002). *Stochastic-Process Limits: an Introduction to Stochastic-Process Limits and their Application to Queues*. Springer Science & Business Media. 12, 38

Whitt, W. (2004). Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science*, 50(10):1449–1461. 34, 63

Whitt, W. (2005). Two fluid approximations for multi-server queues with abandonments. *Operations Research Letters*, 33(4):363–372. 37

Whitt, W. (2006). Fluid models for multiserver queues with abandonments. *Operations research*, 54(1):37–54. 34, 37, 63

Whitt, W. (2007). What you should know about queueing models to set staffing requirements in service systems. *Naval Research Logistics (NRL)*, 54(5):476–484. 13, 20, 37

Whitt, W. (2013). OM Forum—Offered load analysis for staffing. *Manufacturing & Service Operations Management*, 15(2):166–169. 13, 37

Wolstenholme, E. (1999). A patient flow perspective of UK health services: exploring the case for new "intermediate care" initiatives. *System Dynamics Review*, 15(3):253–271. 10, 33

World Health Organization (2014). http://www.who.int/kobe_centre/ageing/en/. 4

Xie, H., Chaussalet, T., and Millard, P. (2005). A continuous time Markov model for the length of stay of elderly people in institutional long-term care. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(1):51–61. 10, 79

Yom-Tov, G. and Mandelbaum, A. (2014). Erlang-r: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management*, 16(2):283–299. 13, 36, 37, 79

Zhang, D., Gurvich, I., Van Mieghem, J., Park, E., Young, R., and Williams, M. (2016). Hospital readmissions reduction program: An economic and operational analysis. *Management Science*, 62(11):3351–3371. 76

Zhang, Y., Puterman, M., Nelson, M., and Atkins, D. (2012). A simulation optimization approach to long-term care capacity planning. *Operations Research*, 60(2):249–261. 13

Zohar, E., Mandelbaum, A., and Shimkin, N. (2002). Adaptive behavior of impatient customers in tele-queues: Theory and empirical support. *Management Science*, 48(4):566–583. 78

Zychlinski, N., Mandelbaum, A., and Momčilović, P. (2018a). Time-varying many-server finite-queues in tandem: Comparing blocking mechanisms via fluid models. *Under revision in Operation Research Letters*. 6

Zychlinski, N., Mandelbaum, A., and Momčilović, P. (2018b). Time-varying tandem queues with blocking: Modeling, analysis and operational insights via fluid models with reflection. *Queueing Systems*, 89(1-2):15–47. 6, 63, 67, 68, 73, 99

Zychlinski, N., Mandelbaum, A., Momčilović, P., and Cohen, I. (2018c). Bed blocking in hospitals due to scarce capacity in geriatric institutions – cost minimization via fluid models. *Under revision*. 6, 37, 40, 52

רשתות נוזלים משתנות בזמן עם חסימות:

מודלים התומכים בניתוח זרימת חולים בבתי חולים

נועה ז'יכלינסקי

רשתות נוזלים משתנות בזמן עם חסימות:

מודלים התומכים בניתוח זרימת חולים בבתי חולים

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת התואר דוקטור לפילוסופיה

נועה ז'יכלינסקי

הוגש לסנט הטכניון - מכון טכנולוגי לישראל

תמוז תשע"ח, יוני 2018

המחקר נעשה בהנחיית פרופסור אבישי מנדלבאום וד"ר יצחק כהן בפקולטה להנדסת תעשייה וניהול, הטכניון – מכון טכנולוגי לישראל

אני מודה לטכניון ולמשרד המדע, החלל והטכנולוגיה על התמיכה הכספית הנדיבה בהשתלמותי

פרסומים:

1. Zychlinski, N., Mandelbaum, A., Momcilovic, P., and Cohen, I. (2018). Bed blocking in hospitals due to scarce capacity in geriatric institutions – cost minimization via fluid models. Under revision in Manufacturing and Service Operations Management (MSOM).

2. Zychlinski, N., Mandelbaum, A., and Momcilovic, P. (2018). Time-varying tandem queues with blocking: Modeling, analysis and operational insights via fluid models with reflection. Queueing Systems, 89(1), 15-47.

3. Zychlinski, N., Mandelbaum, A., and Momcilovic, P. (2018). Time-varying many-server queues in tandem: Comparing blocking mechanisms via Fluid. Under revision in Operations Research Letters.

תקציר

הבסיס למחקר זה הוא הצורך להתמודד עם בעיית ה- bed blocking (חסימת מיטות) בבתי חולים. בעיה זו מתרחשת כאשר חולים קשישים מסיימים טיפול בבית חולים, אך נאלצים להשאר מאושפזים שם עד אשר תתפנה עבורם מיטה פנויה במוסד גריאטרי מתאים. בעיית חסימת המיטות מהווה אתגר תפעולי, עקב השלכותיה הכלכליות והבריאותיות: חולים המעוכבים בבתי חולים אינם מקבלים את הטיפול המתאים ביותר עבורם (שיקום, למשל) ובנוסף, משום שהם תופסים מיטה במחלקה, הם חוסמים/מונעים העברה ואשפוז של חולים חדשים מחדר המיון. מעבר לכך, בעיית חסימת המיטות כרוכה בעלויות גבוהות, משום שעלות אשפוז בבית חולים גבוהה באופן משמעותי מעלות אשפוז במוסד גריאטרי.

אספקת שירותים רפואיים באיכות גבוהה לאוכלוסיית הקשישים מהווה אתגר מרכזי במדינות מפותחות. אתגר זה הולך ומעצים, לאור העובדה שמספר הקשישים בגילאי 65 ומעלה, המהווים היום 10% מהאוכלוסיה, צפוי להכפיל את עצמו בתוך שני עשורים. בנוסף, משום שחלק מהחולים הקשישים נאלצים להתאשפז לעיתים תכופות בבתי חולים, העומס בהם צפוי לגדול. כבר בשנים האחרונות מספר מדינות OECD מדווחות על ממוצע של מעל 90% בתפוסת מיטות במחלקות האשפוז בבתי החולים. ממוצעים שנתיים אלה אינם חושפים את המציאות היומיומית בבתי החולים בתקופות העמוסות (בחורף).

החלק הראשון בעבודה זו (פרק 2) מתמקד בניתוח בעיית חסימת המיטות, וזאת במטרה לשפר את תפעולם המשותף של בתי החולים והמוסדות הגריאטריים. לשם כך, אנו מתמקדים בזרימת חולים ברשת הכוללת מחלקות אשפוז בבתי חולים ומוסדות גריאטריים הכוללים שלוש מחלקות מרכזיות: שיקום גריאטרי, הנשמה ממושכת וסיעודי מורכב. בסיום הטיפול בבית החולים, חלק מהחולים הקשישים אינם יכולים להשתחרר לביתם עקב מצבם הרפואי והם מופנים, על פי מצבם, לאשפוז המשך באחת מהמחלקות הגריאטריות. משכי האשפוז הממוצעים הם כחודש במחלקת שיקום גריאטרי, שישה שבועות בסיעודי מורכב וחמישה וחצי חודשים בהנשמה ממושכת. במהלך האשפוז בכל אחת מהמחלקות, חולים עלולים למות או, עקב הדרדרות במצבם, להיות מוחזרים לבית החולים לאשפוז חוזר.

העומס הגבוה במערכת וזמני ההמתנה הארוכים למחלקות הגריאטריות עודדו אותנו לנתח את המערכת ולחפש פתרונות תפעוליים לשיפורה. לשם כך, פיתחנו מודל נוזלים מתמטי אשר כולל חסימות, תמותה וחזרות לאשפוז – כל אלו הם מאפיינים מרכזיים בסביבה בה אנו מתמקדים. ההשוואה בין מודל הנוזלים, נתונים של שנתיים מרשת בתי חולים ותוצאות סימולציה, מראה שהמודל שאנו מציעים הוא מדויק ושימושי. בנוסף, אנו מוכיחים שמודל הנוזלים מהווה גבול למערכת הסטוכסטית המקבילה, בה הגעות, קצבי טיפול, תמותה וחזרות לאשפוז הם משתנים מקריים. מודל הנוזלים שאנו מציעים, ובמיוחד העומס המוצע (offered-load) הנובע ממנו, מתבררים כשימושיים ונוחים ליישום בתהליך קבלת החלטות

הקשורות בתכנון והקצאת מיטות. אנו משתמשים במודל הנוזלים ובנתונים שניתחנו על מנת לבצע תחזית של ה- offered-load, אשר לוקחת בחשבון את הגידול באוכלוסייה במהלך אופק התכנון. אחת התוצאות שקיבלנו היא נוסחה סגורה לכמות המיטות הגריאטריות הנדרשות, על מנת למזער עלויות עודף וחוסר. אנו מדגימים כי הפתרון המוצע יכול להוריד משמעותית את עלויות התפעול, וכן את אורכן של רשימות ההמתנה, ביחס למצב הנוכחי.

למודל ההקצאה הבסיסי אנחנו מציעים שתי הרחבות. הראשונה, היא מודל הכולל עלות קבועה הכרוכה בהקמה של מיטות/מחלקות חדשות. ההרחבה השניה, כוללת מודל המאפשר הקצאה תקופתית של מיטות במהלך השנה. לשם כך, אנו משתמשים בעלות הקצאה מחדש עבור כל מיטה. בעיית ההקצאה התקופתית מאפשר לקבוע את אורך התקופות השונות בכל שנה וכן, את כמות המיטות הנחוצה בכל תקופה. הפתרון התקופתי מתאים טוב יותר לעומס המוצע המשתנה בזמן ועל כן, מאפשר הורדה נוספת בעלויות התפעול.

התרומות המרכזיות של פרק זה הן:

1.) מידול – אנו מפתחים ומנתחים מודל אנליטי הכולל מחלקות אשפוז גריטארי ארוך טווח וכן את מחלקות בתי חולים המזינות אותן. ניתוח משולב של רשת זו זה הוא הכרחי על מנת למדל את אפקט החסימה (להבדיל ממחקרים קודמים שהתמקדו בניתוח של תחנה אחת) ואת העלויות הכרוכות בחסימת המיטות.
2.) מתודולוגיה – המחקר שלנו תורם לספרות המקצועית בנושא רשתות תורים עם חסימות. המודל שאני מציעים מתאר את החסימות ללא שימוש ב- reflection והוא ניתן ליישום גם ברשתות אחרות. אנו משתמשים במודל כי להסיק פתרונות אנליטיים ותובנות תפעוליות לגבי מזעור עלויות בבעיות הקצאה של מיטות. גישת הפתרון שאנו מציעים כוללת ניתוח מערכות משתנות בזמן, בעלות קיבולות סופיות, תמותה ואשפוזים חוזרים – כל אלו מאפיינים מרכזיים במערכות בריאות.
3.) פרקטיקה – מחקר זה כולל פיתוח אסטרטגיות חדשות לבעיות הקצאה. אנו מציעים נוסחה סגורה לפתרון בעיית הקצאה משתנה בזמן, המתאימה לביקוש העונתי. כמו כן, אנו מציעים מודל אנליטי שמביא בחשבון גם עלויות קבועות של הוספת מיטות חדשות. גישת הפתרון שאנו מציעים מאפשרת לסייע למקבלי החלטות במערכת הבריאות בנוגע לתכנון הקצאה של מיטות.

על מנת לנתח את המערכת בצורה מקיפה יותר, חשוב לקחת בחשבון גם את "מעוכבי האשפוז"–אותם חולים הממתינים בחדר המיון למיטה פנויה באחת המחלקות. ניתוח זה צריך לכלול, בנוסף, גם חדרי המתנה עם קיבולת סופית ואובדן של לקוחות הנאלצים לעזוב כאשר המערכת מלאה. לשם כך, בפרק 3

אנו ממדלים ומנתחים רשתות תורים טוריות המשתנות בזמן, עם חסימות וחדרי המתנה בעלי קיבולת סופית, הן לפני התחנה הראשונה והן בין התחנות. מודלים אלו כוללים את המאפיינים המהותיים של המודל אותו ניתחנו בפרק 2: השתנות בזמן וחסימות. אך להבדיל מהמודל הראשון, מודלים אלו לוקחים בחשבון גם אובדן לקוחות, אשר מתרחש כשחדר ההמתנה הראשון מלא. מידול זה מחייב ניתוח של reflection ועל כן, הן המודל הסטוכסטי במקרה זה והן הוכחת ההתכנסות למודל הנוזלים הם מורכבים יותר. מודל הנוזלים עבור קבוצת הרשתות שאנו מנתחים כולל סט של משוואות דיפרנציאליות לא רציפות (Differential Equations with a discontinues right-hand-side). משוואות אלו ניתנות לפתרון בקלות באופן נומרי. פרק 3 מסתיים בתובנות תפעוליות לגבי רשתות תורים טוריות במובנן הרחב, מעבר לרשתות בתי חולים. התובנות אלו כוללות את ההשפעה של מאפייני הרשת על מדדיה התפעוליים (תפוקה, זמני שהייה המתנה וחסימה ומספר לקוחות בכל תחנה בכל זמן).

התרומות המרכזיות של פרק זה הן:

1.) מידול – אנו מנתחים מודל משתנה בזמן של k תחנות מרובות שרתים בטור, הכולל חדרי המתנה בעלי קיבולת סופית לפני התחנה הראשונה ובין התחנות. מודלים אלו כוללים גם חדרי המתנה בעלי קיבולת אינסופית וגם רשתות לא חדרי המתנה כלל. מקרה פרטי של המודלים כולל מערכת $G_t/M/N/(N+H)$ עבור תחנה אחת. לכל הרשתות האלו אנו מפתחים מודל נוזלים מאוחד המאופיין על ידי סט משוואות דיפרנציאליות לא רציפות.
2.) ניתוח המודל הסטוכסטי – המודל הסטוכסטי הראשון שאנו מציגים למשפחת הרשתות שאנו מנתחים מתבסס על תפוסת התחנות (occupancy). מתברר, שהצגת המודל המתבססת על דווקא על השרתים שאינם מנוצלים (non-utilized) נוחה יותר לניתוח. הצגה זו מאפשרת תיאור של הרשת באמצעות reflection, ממנה ניתן להסיק תכונות שימושיות של אופרטור הרשת של ה-reflection (רציפות ליפשיץ).
3.) ניתוח מודל הנוזלים – באמצעות החוק הפונקציונלי של המספרים הגדולים (Functional Strong Law of Large Numbers), אנו מפתחים את גבול הנוזלים למערכת הסטוכסטית, הכולל reflection. באמצעות שימוש בתכונות אופרטור ה- reflection, אנו פותרים את מודל הנוזלים ומבטאים אותו באמצעות סט משוואות דיפרנציאליות ללא reflection. ייצוג זה הוא אפקטיבי, גמיש ומדויק ועל כן, נוח ליישום עבור מגוון של רשתות.
4.) תובנות תפעוליות – המודלים שאנו מציעים מאפשרים הסקת תובנות תפעוליות על רשתות טוריות משתנות בזמן עם חדרי המתנה סופיים. באמצעות ניסויים נומריים, אנו מנתחים את ההשפעה של אורך הקו (מספר התחנות ברשת), מיקומו של צוואר הבקבוק, גודל חדר ההמתנה הראשון והאינטראקציה ביניהם, על ביצועי הרשת ומדדיה התפעוליים.

בעוד פרקים 2 ו-3 עוסקים במנגנון חסימות מסוג "חסימה לאחר שירות" (Blocking After Service - BAS), פרק 4 עוסק ברשתות תורים משתנות בזמן הפועלות על פי מנגנון מסוג "חסימות לפני שירות" (Blocking Before Service - BBS), אשר נפוצות במערכות תקשורת, ייצור ואף במערכות בריאות. במנגנון זה, שירות מתחיל בתחנה רק אם יש מקום פנוי לאותו לקוח גם בתחנה הבאה ברשת. אנו מתחילים בפיתוח המודל הסטוכסטי עבור רשתות תורים טוריות, משתנות בזמן ומרובות שרתים הכוללות חדרי המתנה בעלי קיבולת סופית לפני התחנה הראשונה ובין התחנות. בשלב הבא, אנו מפתחים את גבול הנוזלים המתאים למודל הסטוכסטי. פיתוח זה כולל reflection, שנובע מקיבולתו הסופית של חדר ההמתנה הראשון. אנו מספקים מספר דוגמאות המדגימות את דיוקו ויעילותו של מודל הנוזלים בתיאור המערכת הסטוכסטית אותה הוא מקרב.

לבסוף, אנו מנתחים את המודלים במצב יציב ומקבלים נוסחה סגורה לתפוקת הרשת ולקצב אובדן הלקוחות. תפוקת הרשת היא למעשה המינימום בין קצב ההגעה, קיבולת העיבוד של צוואר הבקבוק וקיבולת העיבוד של צוואר הבקבוק "הוירטואלי", אשר נובע ממנגנון ה- BBS ועל כן, כולל למעשה שתי תחנות עוקבות. סיום הפרק כולל השוואה אנליטית בין שני מנגנוני החסימה (BAS ו- BBS) והסקת תובנות תפעוליות/תכנוניות לגביהם, כולל התנאים בהם בשני המנגנונים יתקבלו אותם מדדים תפעוליים.

התרומות המרכזיות של פרק זה הן:

1.) מידול – מחקר זה מעשיר מודלים קיימים בכך שהוא מוסיף השתנות בזמן ברת חיזוי, מערכות מרובות שרתים וחדרי המתנה בעלי קיבולת סופית, הפועלים על פי מנגנון BBS.
2.) יישום - המודלים שאנו מציעים הם קלים ליישום, מדויקים ואפקטיביים ביחס למערכות הסטוכסטיות שאותן הם מקרבים.
3.) פרקטיקה - אנו מספקים השוואה אנליטית בין מנגנוני חסימה שונים. השוואה זו מובילה לתובנות תפעוליות ומאפשרת לקבוע, בהתאם לפרמטרים של הרשת, תחת אילו תנאים כדאי להשתמש בכל מנגנון.