Models for Capacity Allocation in Anticipation of Time-Varying Demand

Evghenii Furman

A dissertation submitted to the Faculty of Graduate Studies
in partial fulfilment of the requirements
for the degree of

Doctor of Philosophy

Graduate Program in
Business Administration: Operations Management & Information Systems
York University
Toronto, Ontario, Canada

October, 2020

Abstract

In this dissertation, we propose and investigate several stationary capacity allocation methods that anticipate time-varying demand. We apply these techniques to three practical settings in customer acquisition and retention, cloud computing, and healthcare.

In the first part of this dissertation, we model the trade-off between customer acquisition and retention as a multi-class queueing network with returning customers, time-dependent arrivals, and abandonment. Based on its fluid approximation, we propose an approach to determine optimal stationary staffing levels by partitioning the time-limiting solution of the dynamical system. We test our method by applying it to two real-world applications, i.e., advertising campaigns and a clinical setting, and demonstrate its superiority when comparing to other state-of-the-art approaches.

In the second part, we analyze a cloud computing system where a provider wants to determine the optimal number of servers and retrial interval for incoming jobs when all servers are busy. Servers in this setting represent components of a computer network and customers are jobs attempting to access the cloud computing infrastructure. By modeling the system as a fluid queue and using a calculus-of-variations approach, we derive the optimal amount of service capacity and retrial interval in anticipation of time-varying dynamics. We conduct a case study using data collected from a real cloud service provider and show that significant savings can be realized.

Finally, we estimate the demand for personal protective equipment (PPE) in the general internal medicine (GIM) department of a hospital during the COVID-19 pandemic. We derive closed-form estimates of demand for multiple types of PPE using a queueing framework with generally distributed service times that models medical interactions with heterogeneous patients whose hospital admissions are time-varying. We parametrize our predictive model using a data set containing patients' clinical and operational records over a period of 9 years. We find that gloves and surgical masks represent approximately 90% of predicted PPE usage. We also find that while demand for gloves is driven entirely by patient-practitioner interactions, 86% of the predicted demand for surgical masks can be attributed to the requirement that medical practitioners will need to wear them when not interacting with patients.

To my loving parents and grandparents

Acknowledgements

First and foremost I would like to thank my advisor Prof. Adam Diamant. His guidance, and support transformed me from a Ph.D. freshman into a researcher walking on an academic journey with confidence. His meticulous editing and indispensable advice have raised my bars for rigor, dedication and professionalism higher than they could ever be.

I appreciate all the support of Prof. David Johnston and thank Prof. Alexey Kuznetsov for humbling my achievements by his mere presence and helping me understand that strength of body and mind are very much aligned.

I would like to thank Prof. Richard Irving and Prof. Manus Rungtusanatham for supporting my academic and non-academic endeavours. I am grateful to Paula Gowdie Rose for her kind treatment of OMIS Ph.D. students. I also thank all part-time and full-time faculties and staff members who contributed to this dissertation one way or another.

In particular, I am grateful to our best research manager Joanne Pereira for always being there to help, to support with an advice or to share full-hearted laughs.

I express my gratitude to Prof. Hila Cohen whose bagels and hummus recipe have created life lasting memories. The cake you brought for my comprehensive exam was truly precious. Your empathetic nature and kindness are exemplary.

My special appreciation goes to Gina. Your *bright* personality, banana cakes, cookies and Aroma time have made this dissertation so much more fun. Thank you for your life-lasting friendship and a bag of hilarious Greek jokes!

I am particularly grateful to my dear friends Anton, Alan, Jasmine, Ortac, Gozde, Ting for being there for me when needed. Thank you for your insight, patience and all the fun we had over these years.

I thank all members of my family who supported me on this difficult path. Specifically, my brother for inspiring me to begin my academic career. I thank my sister for her support and everlasting trust in my success. I thank my grandmother who always believed in me but unfortunately did not live to see this moment. This dissertation is written in her memory. Finally, I thank my parents whose unconditional love made this dissertation possible.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

A common feature in many operations management applications such as call centers, healthcare, penal systems, and technological services, is time variability (see Green et al. 2007a, for instance). Understanding these time-varying dynamics is important for the accurate evaluation of a service system's performance. For example, due to the daily variability of visits in an emergency department, the average wait-time estimate does not adequately represent operational reality (Kang and Park 2015). Further, demand for technological services is highly volatile because it follows electricity consumption patterns affected by pricing and the circadian rhythm (Jang et al. 2016).

The successful management of systems operating in time-varying environments includes the effective allocation of critical resources such as customer service agents, nurses and computing hardware. Although dynamically allocating capacity by reassigning staff in real time is difficult to justify in practice, stationary techniques that allocate resources in anticipation of time-varying demand remain uncommon. In Chapter 2 and 3 of this dissertation, we develop such methods for settings where time-varying demand is an essential feature. Then, in Chapter 4, we predict the usage of Personal Protective Equipment (PPE) by analysing the non-stationary workload of a hospital to aid in the management of scarce resources during the COVID-19 pandemic.

More specifically, in Chapter 2, we investigate the trade-off between acquisition and retention efforts when customers are sensitive to the quality of service they receive, i.e., whether they get timely access to a company's resources when requested. We model the problem as a multi-class queueing network with new and returning customers, time-dependent arrivals, and abandonment. We derive its fluid approximation; a system of ordinary linear differential equations with continuous, piecewise smooth, right-hand sides. Based on the fluid model, we propose a novel approach to determine optimal stationary staffing levels for new and returning customer queues in anticipation of future time-varying dynamics. Using system accessibility as a proxy for service quality and staffing levels as a proxy for investment, we demonstrate how to apply our approach to two families of time-varying arrival functions motivated by real-world applications: an advertising campaign and a clinical setting. In a numerical study, we demonstrate that our approach creates staffing policies that maximize throughput while balancing acquisition and retention efforts more effectively (i.e., equitable abandonment from each customer class) than commonly used near-stationary methods such as variants of square-root staffing policies. Our model confirms that acquisition and retention efforts are intimately linked; this has been found in empirical studies but not captured in the operations literature. We suggest that in time-varying environments, focusing on either alone is not sufficient to maintain high levels of throughput and service quality.

In Chapter 3, we determine the jointly optimal service capacity and retrial intervals between unsuccessful service attempts for a major provider of cloud computing services. Allocating sufficient capacity to cloud services is a challenging task because demand is time-varying. Thus, most firms have been expanding their capacity with little regard to the consequences associated with idle resources, such as excessive energy consumption and excess costs. We model the system as a multi-station queueing network where the arrival rate of jobs is time-varying and the servers represent CPU cores. Jobs are infinitely impatient and those that are not immediately serviced may retry several times before permanently abandoning the system. We introduce an offered load approximation that allows us to construct a recursive representation of the offered load function which describes the fluid dynamics of the system. We develop a calculus-of-variation approach to minimize the total functional variation of the constructed offered load function. We show that an optimal policy can be efficiently obtained and prove that it is similar to maximizing the penalized system throughput. Using a data set of cloud computing requests over a representative 24-hour period from a typical service of our partner organization, we show that our optimal policy results in a 10% reduction in capacity. We also demonstrate that small changes to their service-level agreements may elicit additional savings. Our model can help reduce idle capacity and has implications for managing more sustainable and environmentally friendly cloud computing services. It may also help to explain why so much global cloud capacity is typically idle. That is, in order to satisfy service level agreements encouraging retrial jobs to be processed during off-peak periods while also ensuring that they have short wait times, providers must provision large amounts of capacity.

Finally, in Chapter 4 of the dissertation, we estimate the need for PPE in the general internal medicine department of a hospital affected by the COVID-19 pandemic. Specifically, we determine how much PPE the department has to procure in order to deliver appropriate care to all admitted patients over a specified time-horizon according to the hospital's safety regulations. These estimates are derived by modelling hospital operations as an infinite capacity queue with time-varying arrivals and generally distributed service times. We parametrize this model using a data set spanning over 9 years of clinical operations that includes the records of 22039 patients: their length-of-stay (LoS), demographic data, and clinical information. Under the guidance of medical experts from our partner hospital, we propose a functional relationship between LoS and PPE usage that admits a closed-form expression. This allows us to obtain the estimates of PPE usage without performing extensive simulations. Further, having closed-form expressions equips hospital administrators with an efficient way to evaluate the sensitivity of our predictions to changes in model parameters.

Chapter 2

Customer Acquisition and Retention: A Fluid Approach for Staffing

The work in this chapter won the Queueing SIG student paper competition award from Canadian Operations Research Society in 2019 and has been submitted for publication as the following:

Eugene Furman, Adam Diamant (advisor), Murat Kristal “Customer Acquisition and Retention: A Fluid Approach for Staffing”, in *Production & Operations Management*, [2019]

Dr. Murat Kristal provided financial support for this research, i.e., its presentations in the domestic and international conferences in 2018-2019.

2.1 Introduction

Marketing researchers describe customers as discounted streams of revenue and suggest that retaining clients for longer increases a firm's profit margin (Berger and Nasr 1998, Jain and Singh 2002, Malhotra 2007). Not surprisingly, as part of various customer relationship management (CRM) programs, executives have looked for better ways to retain clients while simultaneously acquiring new ones (Chen and Popovich 2003, Buttle 2004). Acquisition and retention practices are costly, however, and allocating enough resources to support them is critical. For example, BCE invests $110 million into acquiring new clients, and its investments into customer retention programs exceed this amount ten fold (BCE 2016). Further, the CRM industry is large and diverse - it includes hotels, insurance companies, health care organizations, and universities (Milovic 2012) - and its value is expected to reach 40 billion dollars in 2020 (Columbus 2016, Taylor 2018).

Motivated by the task of allocating resources to support CRM initiatives, we investigate the operational trade-off between acquisition and retention efforts when customers are sensitive to the quality of service they receive, i.e., whether they get timely access to a company's resources when requested. The value of capturing the effect of service quality in retention efforts is alluded to both in the literature and industry reports. For instance, Keaveney (1995) identifies quality-of-service as one of the major reasons customers seek a different service provider. Similarly, Oracle claims that service quality substantially contributes to a customers decision to leave (Oracle 2011). We focus on the relationship between a customer's ability to access services and quality. Specifically, increased access to company resources implies higher levels of service quality (Feinberg et al. 2002, Dean 2002) which, in turn, results in fewer unhappy customers who change providers (Reichheld and Sasser 1990, Reichheld and Schefter 2000). There is also an inverse effect, customers who find it more difficult to access services will be more likely to leave.

We model a service system with new and returning customers as a three-station queueing network with feedback and abandonment. After new customers receive service at the first station, they decide whether to join the customer base (the second station). Returning customers periodically request service (station three) and after each service episode, may choose to remain with the company (returning to the second station for a period of time) or defect (leave the system). Customers seeking service are impatient and may decide to leave if they wait sufficiently long. To capture the non-stationary nature of demand, we assume that new customers arrive according to a non-homogeneous Poisson process. Exact analysis of the stochastic system is intractable, and thus, we derive its fluid approximation which results in a system of ordinary linear differential equations (ODE) with continuous, piecewise smooth right-hand sides. Conforming to our goal of balancing acquisition versus retention, we consider the limiting behaviour of the system of ODEs. We formulate a mixed-integer program that determines how much capacity to allocate to each customer type in order to maximize throughput while ensuring no customer type is disproportionally neglected (i.e., abandonment). We focus on the development of appropriate stationary staffing policies in a time-varying environment as frequent staffing changes are costly (Park and Bobrowski 1989, Koole 1997, Kitaev and Serfozo 1999, Batta et al. 2007), can result in added employee stress, and turnover costs exceeding 20% of an employee's annual compensation (Boushey and Glynn 2012).

We apply our theoretical results to two practical scenarios: an advertising campaign (Afèche et al. 2017) and a clinical setting (Yom-Tov and Mandelbaum 2014a). The first scenario assumes that the number of new customers seeking service increases following a promotion. Thus, the firm must decide how much attention to reserve for new customers while also ensuring that existing ones are not neglected. The second scenario assumes that the arrival rate of new customers is periodic, like in the case of patients arriving to a medical facility. The institution must determine the amount of resources to allocate to newly arriving patients versus those that need more frequent, sustained service. In both cases, our model incorporates the time-variability of arrivals and customer abandonment which represent features that affect the relationship between staffing assignments and service quality over extended time periods. Although these represent different operating regimes, our objective is to present the full generality of our methodology by demonstrating that it accounts for a wide range of arrival patterns while proposing stationary staffing policies that are domain specific. Further, it allows a manager to better analyze the long-term effects of resource allocation decisions in a dynamic environment. Indeed, we find that a more realistic model yields additional managerial insight to what can be obtained when demand is assumed to be stationary (e.g., Yom-Tov and Mandelbaum 2014a, King et al. 2016, Afèche et al. 2017).

To evaluate the effectiveness of our approach, we conduct a simulation study and compare the performance of our policy to several benchmarks that are based on the square root staffing (SRS) rule (see, e.g., Feldman et al. 2008a, Janssen et al. 2011, Liu 2018). As compared to the benchmarks, we demonstrate that our approach recommends staffing levels that better balance acquisition and retention efforts over a wide-range of system parameters and arrival functions. We also find that although our objective is to ensure new and returning customers are adequately served, overall throughput does not suffer. Finally, we demonstrate that employing time-varying staffing policies when demand is time-varying may not always be necessary. This is important as there are many settings where dynamically reassigning staff is not feasible (Chan and Sarhangian 2018).

We make several contributions to the literature. First, our research extends previous work on stationary staffing policies for queueing systems by introducing a new method that better accounts for time-dependent demand. Specifically, we show that our policies dominate the near-stationary benchmarks by ensuring both new and returning customers receive timely service while maintaining similar levels of throughput. Further, we identify regimes where our policy results in higher levels of throughput (e.g., high traffic intensity, long service times, high service frequency, low abandonment rate). Thus, our approach performs well without neglecting certain customer classes. An added benefit is that our staffing polices can easily accommodate customers with varying profit contributions and/or lifetime value considerations that are common in the CRM literature.

We add to the literature on customer acquisition and retention by introducing a model that explicitly links these quantities to service quality. Further, in this more realistic setting, we confirm and refine existing results on resource allocation strategies. Similar to King et al. (2016), for instance, we find that acquisition efforts are vulnerable to diminishing returns. However, in contrast, dedicating too few resources to acquisition efforts at any time results in poor performance. In fact, our results suggest that customer acquisition and retention efforts are intimately intertwined, which agrees with the marketing literature (see, for instance, Thomas 2001). Whereas assigning few staff to returning clients results in a performance decrease, too few resources dedicated to acquisition efforts undermines retention activities. Hence, balancing this trade-off requires careful consideration of both acquisition and retention practices in a time-varying environment.

2.2 Literature Review and Contribution

Our paper contributes to the literature on deterministic fluid models for multiserver queueing networks. Specifically, we apply a dynamical systems analysis to determine an optimal stationary staffing policy in a multi-server queueing network with returning customers (feedback), time-dependent arrivals, and abandonment. Many authors investigate queueing systems with arrivals that follow a non-homogeneous Poisson process; see the surveys by Defraeye and Van Nieuwenhuyse (2016), Whitt (2016), and Whitt (2018). The problem of staffing in dynamic service systems has also been an active area of study (e.g., Henderson et al. 1999, Akcali et al. 2006, Bhandari et al. 2008, Robbins and Harrison 2010). Further, several papers investigate staffing in queueing systems with time-dependent arrival processes and abandonment; see Harrison and Zeevi 2005, Bassamboo and Zeevi 2009, Bekker and de Bruin 2010, Defraeye and Van Nieuwenhuyse 2013, Niyirora and Pender 2016 for example. However, few papers examine stationary staffing models in queueing networks with time-varying arrivals and both returning and impatient customers.

Our research proposes a methodology to assign a limited number of flexible servers to customers of different classes given that customers may seek service multiple times. The dynamic and static assignment of flexible servers has been investigated by several researchers; see the paper by Andradóttir et al. (2001). The objective in these papers is to maximize the long-run average throughput or minimize the long-run average costs. For problems where staff can be allocated dynamically, several models have been proposed; single-server queueing systems (Andradóttir et al. 2003), tandem queues (Andradóttir et al. 2007), assembly-type queues (Tsai and Argon 2008), and discrete assignment intervals (Chan and Sarhangian 2018). Our approach involves solving a static optimization problem to determine the number of servers assigned to each customer class. Although several papers investigate the assignment of a fixed-set of servers (e,g., Hillier and So 1989, Futamura 2000, Smith et al. 2010, Lee et al. 2014, Smith and Barnes 2015), few pursue this objective for time-varying systems, i.e., Harrison and Zeevi (2004), Bassamboo et al. (2006).

The square-root staffing (SRS) rule is a well known dynamic staffing model in the operations literature (see, for instance, Borst et al. 2004, Feldman et al. 2008a, Hampshire et al. 2009, Janssen et al. 2011). A challenge for SRS, as described in Green et al. (1991, 2007b), is its application to non-stationary systems. Further, most applications of the SRS rule assume that staffing levels can be changed in real-time, a strong assumption that is typically not true in practice (e.g., health care, unionized shift work). Instead, staffing levels are set at asynchronous intervals by management over relatively long time horizons. In this case, a manager must anticipate the non-stationarity of the arrival process. Our model accounts for this and we, to quantify the benefit of our approach, compare our policy to several variants of the SRS policy (Liu 2018).

As discussed, we analyze the dynamical system corresponding to the fluid limit of the original stochastic queueing network (Halfin and Whitt 1981). Foundational results on fluid approximations can be traced back to Mandelbaum et al. (1998), Whitt (2006), and Kang et al. (2010a) for cases with reneging. The precision of fluid approximations are evaluated in Bassamboo and Randhawa (2010), Daw and Pender (2019) while Pender et al. (2017) use the approach to analyze the impact of delay announcements. Unlike previous work which grounds the evaluation of system performance on a set of distributional assumptions applied to the fluid model (e.g., Jouini et al. 2013, Bassamboo and Randhawa 2015), we focus solely on analyzing the performance of the fluid model.

Our work obtains insights on how to balance customer acquisition versus retention in dynamic service environments. These systems are characterized by a set of new customers (a priori homogeneous) and returning customers (possibly heterogeneous in their service requirements). Although several case studies and empirical research describe this phenomenon in the marketing literature (e.g., Thomas 2001, Reinartz et al. 2005), there are few papers that model and analyze the trade-off. Fruchter and Zhang (2004) formulate differential game, with two competing firms and a fixed market, to investigate how effective acquisition and retention efforts are at generating sales. They find that under various conditions, a focus on either customer retention or acquisition can be an optimal long-run strategy for a firm but do not discuss how to dynamically balance these efforts over time. Dong et al. (2011) develop incentive mechanisms to determine the sales channel (i.e., direct selling versus delegation) for customer retention and acquisition. They find that when acquisition and retention efforts are in conflict, a firm should choose to focus on retention. King et al. (2016) introduce a discrete-time, deterministic dynamic programming model to optimally allocate the resources of a profit maximizing firm. They derive an optimal investment policy which indicates that a firm should shift its focus from acquisition to retention as the size of a firm's customer base grows over time. Finally, Afèche et al. (2017) present a static optimization model to find a set of optimal staffing levels when customers differ in their service request rates, have rewards that may depend on service, and have different return probabilities. They find that an optimal policy has a "bang-bang" structure - certain customers get service or not at all. We extend this literature by incorporating time-varying demand and abandonment. Further, the trade-off between whether to invest in customer acquisition and retention efforts is modeled directly into the dynamics of the problem as the ability of a firm to allocate enough resources to satisfy time-varying demand.

The customer retention and acquisition literature is related to the analysis of service systems with feedback (de Véricourt and Jennings 2011); Jacobson et al. (2012) model an emergency response environment using a closed-queueing network model with feedback and a fixed number of customers who may leave the system. Ding et al. (2015) analyze a call center with unspecialized servers, retrials, and reconnects. Yom-Tov and Mandelbaum (2014a) and Huang et al. (2015) propose a SRS policy to stabilize the performance of a time-varying network with returning customers without abandonment. Liu and Whitt (2017) extend this work by developing an offered-load approximation in a time-varying, many-server queueing system with customer abandonment and returning customers. Chan et al. (2014) present an analysis of a piecewise smooth dynamical system with discontinuous right-hand sides where the service rate is dependent on the number of customers in the system. Our paper takes an alternative approach: we determine an optimal stationary staffing level in anticipation of time-varying dynamics. Further, we demonstrate that our asymptotic analysis and static optimization model can be adapted to a variety of service settings.

2.3 Model Formulation

In this section, we introduce a stochastic queueing network to model a firm's ability to acquire and retain customers, derive its fluid limit, and formulate the corresponding dynamical system.

Let $\mathcal{I} \equiv \{a, b\}$ be a set of customer classes that correspond to new and returning clients respectively. Class-a clients represent new customers that arrive to the system according to a non-homogeneous Poisson process with time-varying intensity $\lambda \equiv \lambda(t)$. Class-$b$ clients are customers that have received service sometime in the past and may require service again. Class-$i \in \mathcal{I}$ clients have service requirements that are exponentially distributed with rate μ_i. There is also a fixed-sized pool of $s > 0$ flexible staff that are able to serve any class of customer and a decision maker that must decide how many staff to dedicate to each customer class. We assume that the staffing policy is stationary, i.e., once decided it cannot be easily changed. Let s_a and s_b be the staffing assignment to new (Q_a) and returning (Q_b) service stations, respectively, such that $s_a + s_b \leq s$. Servers attend to customers of an assigned class under the standard first-in, first-out (FIFO) service policy.

We account for the revenue differential, or the penalty of customer abandonment, by assigning weights ρ_a and ρ_b to new and base customers attempting to receive service in accordance with their relative average lifetime value. That is, we define ρ_i to be relative average lifetime value (LV) of a class-i client seeking service where, without loss of generality, $\rho_b = 1$ and $\rho_a = \frac{LV_a}{LV_b} \in \mathbb{R}_{\geq 0}$. The inclusion of these weights links our queueing model to the CRM literature where balancing acquisition and retention efforts is achieved by maximizing the long-term profit of the firm.

After a class-a customer arrives to the system, she is served immediately if there is an available server at station Q_a, and waits for service otherwise. All customers are impatient and may decide to abandon the queue before receiving service. We assume abandonment times of class-a customers are exponentially distributed with rate τ_a. Then, class-a clients who complete their service either leave the system or join the customer base (which we model as a station Q_c with infinite capacity); they become a class-b client. If a client joins the customer base, she may again seek service, this time at station Q_b, after an exponentially distributed amount of time with mean $1/r$. Customers may also decide to cease their relationship with the service provider; attrition times from station Q_c are exponentially distributed with rate ζ. Class-b clients that seek service at station Q_b are, again, impatient; abandonment times are exponentially distributed with rate τ_b. Class-b customers who abandon the queue rejoin the customer base (i.e., Q_c) with probability θ_c or leave the system.

Let $\mathcal{QN} \equiv (\mathcal{M}, \mathcal{I}, \mathcal{P})$ define the topology of the queueing network where $\mathcal{M} = \{Q_a, Q_b, Q_c\}$ is the set of stations in the network, $\mathcal{I}$ is the set of customer classes, and $\mathcal{P} : (\mathcal{M} \times \mathcal{M}) \to [0, 1]$ is a function defining the routing probabilities among the stations for clients who complete their service requirements. More specifically,

$$\mathcal{P} : (\mathcal{M} \times \mathcal{M}) \to \begin{pmatrix} \mathbb{P}_{aa} & \mathbb{P}_{ab} & \mathbb{P}_{ac} \\ \mathbb{P}_{ba} & \mathbb{P}_{bb} & \mathbb{P}_{bc} \\ \mathbb{P}_{ca} & \mathbb{P}_{cb} & \mathbb{P}_{cc} \end{pmatrix} = \begin{pmatrix} 0 & 0 & \theta_{ac} \\ 0 & 0 & \theta_{bc} \\ 0 & 1 & 0 \end{pmatrix},$$

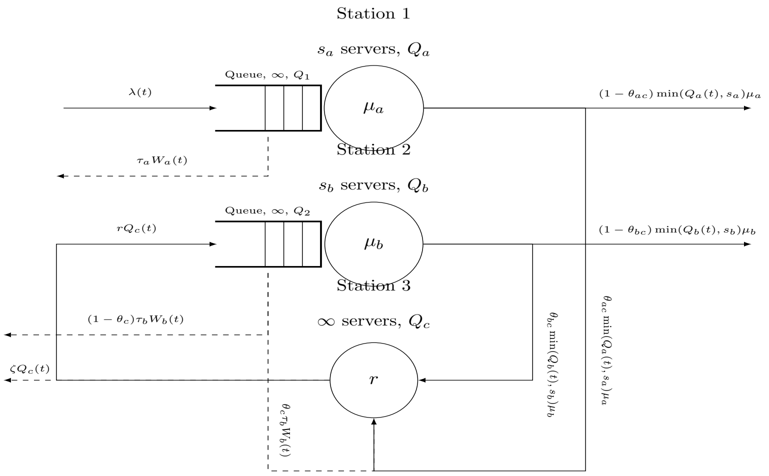
where $\mathbb{P}_{jj'}$ is the probability that a customer who finishes service at station Q_j is routed to station $Q_{j'}$ for $j, j' \in \{a, b, c\}$; θ_{ac} is the probability that a new customer joins the customer base, and θ_{bc} is the probability that a returning customer rejoins the customer base. For simplicity, we refer to a queueing network with specified arrival, service and abandonment rates by its network topology $\mathcal{QN}$.

Let $\mathcal{W} = \{W_a(t), W_b(t), W_c(t)\}$ be a set of headcount stochastic processes corresponding to the number of customers waiting for service in Q_j where $j \in \{a, b, c\}$. Let $\mathcal{X} = \{X_a(t), X_b(t), X_c(t)\}$ be a set of headcount stochastic processes corresponding to the number of busy servers at each station. The system states evolve according to the following equations.

$$\begin{aligned} &Q_a(t) = W_a(t) + X_a(t), \quad W_a(t) = (Q_a(t) - s_a)^+, \quad X_a(t) = (Q_a(t) \wedge s_a); \\ &Q_b(t) = W_b(t) + X_b(t), \quad W_b(t) = (Q_b(t) - s_b)^+, \quad X_b(t) = (Q_b(t) \wedge s_b); \\ &Q_c(t) = X_c(t). \end{aligned} \tag{2.1}$$

With a slight abuse of notation, we denote stations by Q_j and corresponding to them counting stochastic processes by $Q_j(t)$. The states of $\mathcal{QN}$ are described by the stochastic vector $\boldsymbol{Q} = (Q_a(t), Q_b(t), Q_c(t))^T \in \mathbb{R}^3_{\geq 0}$. The dynamics of the system is presented in Figure 2.1.

Figure 2.1: Dynamics of service system



2.3.1 Stochastic Queueing Model

Let N^j_λ, N^j_μ and N^j_τ be the standard Poisson arrival, departure and abandonment processes with time-varying intensities for station j, respectively, and let $\boldsymbol{Q}_0 = (Q_a(t_0), Q_b(t_0), Q_c(t_0))^T$ be a vector of initial conditions, i.e., the number of customers at station j at some starting time t_0. Due to the distributional assumptions introduced in Section 2.3, the system can be modeled as three non-stationary, Erlang-A queues. This, as shown in Mandelbaum et al. (1998), allows us to express the stochastic vector $\boldsymbol{Q}$ in the following functional form.

$$
\begin{aligned}
Q_a(t) &= Q_a(t_0) + N_\lambda^a\left(\int_{t_0}^t \lambda(u)du\right) - N_\mu^a\left(\int_{t_0}^t \mu_a\left(Q_a(u)\wedge s_a\right)du\right) \\
&\quad - N_\tau^a\left(\int_{t_0}^t \tau_a\left(Q_a(u)-s_a\right)^+ du\right), \\
Q_b(t) &= Q_b(t_0) + N_\lambda^b\left(\int_{t_0}^t rQ_c(u)du\right) - N_\mu^b\left(\int_{t_0}^t \mu_b\left(Q_b(u)\wedge s_b\right)du\right) \\
&\quad - N_\tau^b\left(\int_{t_0}^t \tau_b\left(Q_b(u)-s_b\right)^+ du\right), \\
Q_c(t) &= Q_c(t_0) + N_\mu^a\left(\int_{t_0}^t \theta_{ac}\mu_a\left(Q_a(u)\wedge s_a\right)du\right) + N_\mu^b\left(\int_{t_0}^t \theta_{bc}\mu_b\left(Q_b(u)\wedge s_b\right)du\right) \\
&\quad + N_\tau^b\left(\int_{t_0}^t \theta_c\tau_b\left(Q_b(u)-s_b\right)^+ du\right) - N_\mu^c\left(\int_{t_0}^t rQ_c(u)du\right) - N_\tau^c\left(\int_{t_0}^t \zeta Q_c(u)du\right).
\end{aligned}
\tag{2.2}
$$

The stochastic processes in (2.2) describes the evolution of the system. New customers that finish service may leave Q_a and join the customer base, Q_c, or they may leave the system entirely. Returning customers may leave Q_c to seek service at Q_b or may defect. After a returning customer receives service at Q_b, she may either leave the system or rejoin the customer base Q_c. Thus, the non-stationary arrival rate of new customers has a knock-on effect where the arrival, departure, and abandonment processes from all stations in the network $\mathcal{QN}$ are non-stationary.

2.3.2 Fluid Limit and Dynamical System

A cornerstone of constructing the fluid (limiting) approximation of the stochastic equations in (2.2) is to describe the asymptotic dynamics of a corresponding large (scaled-up) queueing system. We employ the heavy traffic limit theorems described in Halfin and Whitt (1981) and Mandelbaum et al. (1998). The procedure is used pervasively in the literature (see, for example, Iglehart 1965, Borovkov 1967, Iglehart 1973a,b). Specifically, we consider a family of queueing networks parametrized by η such that the arrival rate and number of servers of system $\mathcal{QN}^\eta$ are scaled up by a factor of η while the traffic intensity is held constant. We also scale up the processes in $\boldsymbol{Q}$ and the initial conditions in $\boldsymbol{Q}_0$. This transforms (2.2) into

$$\begin{aligned}
Q_a^\eta(t) &= Q_a^\eta(t_0) + N_\lambda^a\left(\int_{t_0}^t \eta\lambda(u)du\right) - N_\mu^a\left(\int_{t_0}^t \mu_a\left(Q_a^\eta(u)\wedge \eta s_a\right)du\right)\\
&\quad - N_\tau^a\left(\int_{t_0}^t \tau_a\left(Q_a^\eta(u) - \eta s_a\right)^+ du\right),\\
Q_b^\eta(t) &= Q_b^\eta(t_0) + N_\lambda^b\left(\int_{t_0}^t rQ_c^\eta(u)du\right) - N_\mu^b\left(\int_{t_0}^t \mu_b\left(Q_b^\eta(u)\wedge \eta s_b\right)du\right)\\
&\quad - N_\tau^b\left(\int_{t_0}^t \tau_b\left(Q_b^\eta(u) - \eta s_b\right)^+ du\right),\\
Q_c^\eta(t) &= Q_c^\eta(t_0) + N_\mu^a\left(\int_{t_0}^t \theta_{ac}\mu_a\left(Q_a^\eta(u)\wedge \eta s_a\right)du\right) + N_\mu^b\left(\int_{t_0}^t \theta_{bc}\mu_b\left(Q_b^\eta(u)\wedge \eta s_b\right)du\right)\\
&\quad + N_\tau^b\left(\int_{t_0}^t \theta_c\tau_b\left(Q_b^\eta(u) - \eta s_b\right)^+ du\right) - N_\mu^c\left(\int_{t_0}^t rQ_c^\eta(u)du\right) - N_\tau^c\left(\int_{t_0}^t \zeta Q_c^\eta(u)du\right).
\end{aligned} \tag{2.3}$$

The relations in (2.3) describe a sequence of queueing systems where the number of servers and the number of arrivals grow as $\eta \to \infty$. The compact convergence of (2.3) to its fluid limit is established in Mandelbaum et al. (1998) and we provide the result, without proof, in the following lemma.

Lemma 1. *Let $Q^\eta(t)$ be a scaled up queueing process parametrized by η that corresponds to the original queueing process $Q(t)$ for all $t \geq 0$. Then, we have*

$$q(t) \equiv \lim_{\eta\to\infty} \sup_{t\in\mathcal{T}} \frac{1}{\eta} Q^\eta(t) \ \ a.s., \tag{2.4}$$

where $\mathcal{T} \subset \mathbb{R}_{\geq 0}$ is a compact set such that $q(t)$ solves the integral equation

$$q(t) = q(t_0) + \int_{t_0}^t \left(\lambda(u) - \mu(u)(q(u)\wedge s) - \tau(u)(q(u)-s)^+\right) du. \tag{2.5}$$

Process $Q^\eta(t)$ is said to converge compactly to $q(t)$, a fluid approximation of $Q(t)$.

Consider the queueing network $\mathcal{QN}$ whose dynamics are governed by the stochastic functional equations in (2.2). If $\boldsymbol{q} = (q_a(t), q_b(t), q_c(t))^T$, then by evaluating a limit of the sequence of queueing networks $\mathcal{QN}^\eta$ as $\eta \to \infty$ and applying Lemma 1, we have that

$$
\begin{aligned}
q_a(t) &= q_a(t_0) + \int_{t_0}^t \lambda(u)du - \int_{t_0}^t \mu_a\left(q_a(u) \wedge s_a\right)du - \int_{t_0}^t \tau_a\left(q_a(u) - s_a\right)^+ du,\\
q_b(t) &= q_b(t_0) + \int_{t_0}^t rq_c(u)du - \int_{t_0}^t \mu_b\left(q_b(u) \wedge s_b\right)du - \int_{t_0}^t \tau_b\left(q_b(u) - s_b\right)^+ du, \qquad (2.6)\\
q_c(t) &= q_c(t_0) + \int_{t_0}^t \left[\theta_{ac}\mu_a\left(q_a(u) \wedge s_a\right) + \theta_{bc}\mu_b\left(q_b(u) \wedge s_b\right) + \theta_c\tau_b\left(q_b(u) - s_b\right)^+\right] du\\
&\quad - \int_{t_0}^t rq_c(u)du - \int_{t_0}^t \zeta q_c(u)du.
\end{aligned}
$$

which implies that $\boldsymbol{q}$ is the fluid approximation of $\boldsymbol{Q}$ component-wise.

The right-hand sides of (2.6) are piecewise smooth functions, and if we differentiate (2.6) piecewise with respect to t, the following relation is obtained

$$\dot{\boldsymbol{q}}(t) = f(\boldsymbol{q}(t)), \qquad (2.7)$$

where $f : \mathbb{R}^3 \to \mathbb{R}^3$ is a continuous, vector-valued, piecewise smooth function. Thus, $\boldsymbol{q}(t)$ is differentiable for any $t \geq 0$. Relation (2.7) with initial condition $\boldsymbol{q}_0 = (q_a(t_0), q_b(t_0), q_c(t_0))^T$ defines the initial value problem (IVP)

$$
\begin{aligned}
&\begin{pmatrix} \dot{q}_a(t) \\ \dot{q}_b(t) \\ \dot{q}_c(t) \end{pmatrix} = \begin{pmatrix} \lambda(t) - \mu_a\left(q_a(t) \wedge s_a\right) - \tau_a(q_a(t) - s_a)^+ \\ rq_c(t) - \mu_b\left(q_b(t) \wedge s_b\right) - \tau_b(q_b(t) - s_b)^+ \\ \theta_{ac}\mu_a\left(q_a(t) \wedge s_a\right) + \theta_{bc}\mu_b(q_b(t) \wedge s_b) + \theta_c\tau_b(q_b(t) - s_b)^+ - (r + \zeta)q_c(t) \end{pmatrix}. \qquad (2.8)\\
&\boldsymbol{q}_0 = (q_a(t_0), q_b(t_0), q_c(t_0))^T
\end{aligned}
$$

This IVP is a piecewise smooth system of ODEs with continuous right-hand sides.

Lemma 2. *A solution to (2.8) exists, is unique, and is piecewise smooth.*

Lemma 2 demonstrates that for any smooth time-varying arrival function, the IVP in (2.8) has a unique solution and the solution is piecewise smooth with a continuous right-hand side. Nevertheless, our analysis of staffing rests on a stronger assumption, namely, that the dynamical system is stable, i.e., converges to a limit as $t \to \infty$. The stability is especially important for systems with monotonous arrival functions. The limit (equilibrium point) and the corresponding speed of convergence provide insight into the behaviour of the solution for small values of t. The next result states a necessary and sufficient condition for the stability of (2.8).

Lemma 3. *If $\lambda(t) \in \mathbb{C}^1$ is a smooth function and $\lim_{t\to\infty} \lambda(t) < \infty$, (2.8) is stable.*

Although stable systems with continuous monotonous arrival functions (e.g., exponential) are mathematically convenient, there are many practical instances where the conditions of Lemma 3 are too strict. For example, we consider a particular instance from the family of periodic arrival functions. To address this, we define the concept of boundedness for any initial condition. Specifically, a solution to (2.8) is bounded if there exists a $\boldsymbol{K} > \boldsymbol{0}$ such that $\boldsymbol{q}(t) < \boldsymbol{K}$ for any $t > 0$. Notice that all vector inequalities and products are defined component-wise.

Lemma 4. *Suppose $\lim_{t\to\infty} \lambda(t)$ is undefined for $\lambda(t) \in \mathbb{C}^1$ but there exists a $K' > 0$ such that $\lambda(t) < K'$ for any $t > 0$. Then, $\boldsymbol{q}(t)$ is bounded.*

Because the solution to (2.8) defines a piecewise smooth, continuous, dynamical system in $\mathbb{R}^3_{\geq 0}$, $q_a(t) = s_a$ and $q_b(t) = s_b$ describe planes (i.e., the boundary Σ) that split the state space into four subspaces, $\mathcal{S}_i$, $i = \{1, 2, 3, 4\}$. The right-hand sides of (2.8) change each time the solution crosses the boundary. We rigorously define the boundary, as well as the subsets $\mathcal{S}_i$, in Appendix A. Nevertheless, because our system has continuous right-hand sides, solutions cross the boundaries and immediately enter into adjacent regions, i.e., the amount of time spent on the boundary has measure zero. This is in contrast to systems with discontinuous right-hand sides (e.g., Chan et al. 2014), where solutions can remain on the boundary for extended periods of time.

For either stable or bounded systems, as characterized by Lemmas 3 and 4, the queue for new customers is described by a non-homogeneous ODE and the complexity of its solution depends on the form of the arrival function. The dynamics governing the queue of returning clients and the customer base requires solving a system of differential equations which take the form

$$\begin{aligned} q_b(t) &= ue^{\psi t} v(\psi) + we^{\chi t} v(\chi) + q_b^*(t), \\ q_c(t) &= ue^{\psi t} + we^{\chi t} + q_c^*(t), \end{aligned} \tag{2.9}$$

where w and u are constants derived from initial conditions, $\psi < 0$ and $\chi < 0$ are the smallest and largest negative eigenvalues associated with the homogeneous solution, respectively, and $v(\psi)$ and $v(\chi)$ are the first coordinates of the eigenvectors corresponding to ψ and χ. Finally, $q_i^*(t)$ for

$i \in \{b, c\}$ is the non-homogeneous term, which converges to a stationary solution as $t \to \infty$. Given a description of the system state $\boldsymbol{q}(t)$, at every point in time, we can calculate the value of the customers in the system using the relation $\boldsymbol{\rho q}(t)$ where $\boldsymbol{\rho} = (\rho_a, \rho_b, \rho_b)^T$.

2.3.3 Model Discussion

Our model captures the trade-off between customer acquisition and retention. Specifically, the number of customers that abandon and the waiting time at each station are functions of the arrival rate and the staffing policy. Determining an optimal staffing assignment, then, involves balancing the length of idle and busy periods for each customer class while also considering their relative importance or average lifetime value. More servers assigned to a customer class implies shorter busy periods in the stochastic regime. Conversely, having fewer severs results in longer busy periods with higher numbers of customers who abandon. Thus, any staffing decision must balance short, new customer waiting times (i.e., better quality-of-service) with a focus on retention and an increased access to company resources. Our model explicitly incorporates the effects of service quality by connecting it to abandonment, a relationship that has not been included in the customer retention literature, while also including time-varying demand and feedback.

Analyzing the behavior of new and returning customers when demand is time-varying is cumbersome in the stochastic regime. Instead, we study the systems fluid limiting behavior. This simplifies the analysis of the dynamics, allows us to capture both overloaded and underloaded regimes, and ensures that optimal control decisions (e.g., staffing) are tractable and implementable. Since a fluid approach assumes a heavy traffic environment, classical application of these results are confined to settings with a relatively large number of arrivals and servers. However, our simulations suggest that our approach does consistently well in environments with a small number of servers and steadily improves as the number of servers increases.

Our work is grounded in classical call center literature and we inherit many of its structural and probabilistic assumptions. For example, we assume that new customers arrive according to a non-stationary Poisson process. Although this may be not an accurate reflection of reality in all settings, the approximation does improve as the system size increases. This also conforms to the heavy traffic assumption used to derive the fluid approximation of the stochastic system. Second, we assume that the service duration and time-to-abandonment are exponentially distributed. Although other more sophisticated distributions can be used, our focus is on how a firm balances acquisition and retention efforts when demand is non-stationary. Using more realistic distributions would increase model complexity and may obfuscate managerial insight. Third, we assume routing probabilities are independent of a customer's service experience. Fourth, following Hu et al. (2016), we consider congestion and fitness abandonment. Customers who wait sufficiently long to obtain service may abandon the system due to congestion (τ_i). This type of abandonment is not related to the quality of service they receive and is dependent only on system accessibility. Those who experience poor service quality may leave the system after a service completion (fitness abandonment). We assume $\theta_{bc} > \theta_c$ as customers are more likely join the customer base after successfully completing service.

For simplicity, we assume two classes of customers, i.e., new and returning clients. Our model, however, can be naturally extended to include several classes of returning clients with a dedicated queue for each additional base customer class. Servers can also be assigned to subsets of clients. Nevertheless, our model assumes that once customers seek service, their rate of abandonment is a function of the queue they enter and not an attribute of the customer class they identify with. The routing probabilities after a service completion follow a similar relationship.

2.4 Staffing Analysis

In this section, we consider two families of arrival functions to better understand the trade-off between acquisition and retention. Theoretically, we capture two broad types of asymptotic behaviour: convergence to a limiting point and convergence to a periodic function with an undefined limit. Practically, the first family is motivated by an advertising campaign, i.e., an event which results in an immediate boost of new customers followed by a gradual decay (e.g., Afèche et al. 2017). The second addresses environments with a periodic pattern of arrivals, such as a clinical setting (e.g., Yom-Tov and Mandelbaum 2014a). In both cases, we study how resources should be allocated given the firms objective is to appropriately serve their customers, i.e, provide timely access to their resources when requested. To this end, we introduce a general four-stage modeling approach that attempts to maximize throughput while ensuring no customer type is disproportionally neglected (i.e., large number of abandonments).

1. Partition the domain into managerially relevant regions.

2. In each region, analyze the limiting behaviour of the dynamical system as $t \to \infty$.

   Systems that converge to limiting points: Establish an ordering relationship amongst the points to determine which piecewise smooth region is most preferable.

   Systems that converge to periodic functions: Find the extrema of the periodic orbits and determine the times when they cross the boundaries of the piecewise smooth regions.

3. Investigate the behavior of the system at finite times (state of disorder).

4. Formulate an mixed-integer program where the objective represents the sum of limiting points (stage 2) weighted by a convex combination of the results from stage 3.

The optimal solution of the optimization model is a stationary staffing policy that accounts for the resource capacity of the system and its time-varying dynamics.

2.4.1 Advertising Campaign: Exponential Decay

We consider a new advertising campaign launched by a service provider. At the start of the campaign the number of customers that request the firm's services rises to its maximum. The increased intensity of arrivals causes an expansion of the customer base, provided the firm's ability to retain clients is not somehow compromised. As the effect of the campaign diminishes, the initial spike in the acquisition of new customers fades and the arrival rate returns to pre-advertisement (or more stable) levels. To capture this behavior, we model the intensity of the arrival process of new customers as a decaying exponential function of the form

$$\lambda(t) \equiv \lambda_1 e^{-\delta t} + \lambda_0, \tag{2.10}$$

where $\lambda_1 > 0$ is the intensity of the campaign, $\delta \in (0, 1)$ controls the steepness of the exponential curve (i.e., the rate of decay), and λ_0 is the intensity of the pre-advertisement arrival process. The arrival function in (2.10) has three intuitive properties: it achieves the global maximum at $\lambda_1 + \lambda_0$, it is monotonously decreasing for all $t > 0$, and $\lim_{t\to\infty} \lambda(t) = \lambda_0$.

Representing the arrival rate of new customers in an advertising campaign as an exponentially decaying function is a convenient choice both practically and mathematically. In Blattberg and

Deighton (1996), for instance, a similar function is used to predict the number of customers acquired over a time-horizon. Further, this choice remains in-line with the marketing literature, which models advertising elasticity (a measure of an advertising campaign's effectiveness in generating new sales) as an exponential function of time (e.g., Parsons 1975, Dant and Berger 1996). Similar considerations are employed in economics where demand is generally assumed to be an exponential function of price (see Thompson and Teng 1984, for instance). Mathematically, an exponentially decaying arrival function assures tractability of the time-varying dynamical system in (2.8).

Substituting (2.10) into (2.8), we obtain the following IVP.

$$\begin{pmatrix} \dot{q}_a(t) \\ \dot{q}_b(t) \\ \dot{q}_c(t) \end{pmatrix} = \begin{pmatrix} \lambda_1 e^{-\delta t} + \lambda_0 - \mu_a(q_a(t) \wedge s_a) - \tau_a(q_a(t) - s_a)^+ \\ r q_c(t) - \mu_b(q_b(t) \wedge s_b) - \tau_b(q_b(t) - s_b)^+ \\ \theta_{ac}(q_a(t) \wedge s_a)\mu_a + \theta_{bc}(q_b(t) \wedge s_b)\mu_b + \theta_{\tau c}\tau_b(q_b(t) - s_b)^+ - (r + \zeta)q_c(t) \end{pmatrix}, \qquad (2.11)$$

$$\boldsymbol{q}_0 = (q_a(t_0), q_b(t_0), q_c(t_0))^T.$$

By Proposition 2 and Lemma 3, the solution of (2.11) is piecewise smooth, and its form changes as the system transitions from one smooth region to another. Further, in each region of the domain (there are 4 such regions), the queue for base customers in (2.11) admits a closed-form solution as in (2.9), and the equation for new customers has a closed-form solution obtained by standard methods. We argue that characterizing the behavior of the system in only a few regions is sufficient for developing a systematic approach to stationary staffing in a time-varying environment. To this end, we partition the domain into three regions characterized by distinct modes of operation. The first mode has high rates of customer acquisition, which in turn, lead to high rates of customer retention. The ability to acquire new customers decreases in the second mode (as compared to the start of the campaign), which results in a slower growth/decline of the customer base. In the third mode, the ability of the firm to acquire new customers returns to pre-advertisement levels and the customer base declines until it reaches a steady-state. Mathematically, we have

Mode 1 - Launch Region: $q_a(t) > s_a$ and $q_b(t) \neq s_b$,

Mode 2 - Loyalty Region: $q_a(t) < s_a$ and $q_b(t) > s_b$,

Mode 3 - Lessening Region: $q_a(t) < s_a$ and $q_b(t) < s_b$.

Given the partition of the domain into modes of operation, our goal is to investigate how a firm can remain in the launch and loyalty regions as long as possible. This is appropriate because in the lessening region, some servers are idle indefinitely, which implies that a new staffing assignment may be more appropriate. Thus, in the second stage of our approach, we examine the limiting behavior of (2.11) and focus specifically on the regions of interest, i.e., the launch and loyalty regions.

Proposition 1 (Limiting Points)**.** *The system state $\boldsymbol{q}(t) = (q_a(t), q_b(t), q_c(t))^T$ is attracted by*

$$(q^*_{a_1}, q^*_{b_1}, q^*_{c_1}) := \begin{cases} \left(\frac{\lambda_0 + s_a(\tau_a - \mu_a)}{\tau_a}, \frac{r\theta_{ac}s_a\mu_a + s_b(\tau_b - \mu_b)(r+\zeta) - s_b r(\theta_{\tau c}\tau_b - \theta_{bc}\mu_b)}{\tau_b(r(1-\theta_{\tau c})+\zeta)}, \frac{\theta_{ac}s_a\mu_a + s_b\mu_b(\theta_{bc}-\theta_c)}{r(1-\theta_c)+\zeta}\right), & q_b(t) \geq s_b, \\ \left(\frac{\lambda_0 + s_a(\tau_a - \mu_a)}{\tau_a}, \frac{r\theta_{ac}s_a\mu_a}{\mu_b(r(1-\theta_{bc})+\zeta)}, \frac{\theta_{ac}s_a\mu_a}{r(1-\theta_{bc})+\zeta}\right), & q_b(t) < s_b, \end{cases}$$

the equilibrium solution in the launch region (i.e., the first mode of operation), and by

$$(q^*_{a_2}, q^*_{b_2}, q^*_{c_2}) := \left(\frac{\lambda_0}{\mu_a}, \frac{r\theta_{ac}q_{a_2}(t_0)\mu_a + s_b(\tau_b - \mu_b)(r+\zeta) - s_b r(\theta_{\tau c}\tau_b - \theta_{bc}\mu_b)}{\tau_b(r(1-\theta_{\tau c})+\zeta)}, \frac{\theta_{ac}q_{a_2}(t_0)\mu_a + s_b\mu_b(\theta_{bc}-\theta_c)}{r(1-\theta_c)+\zeta}\right),$$

the limiting point in the loyalty region (i.e., the second mode of operation) early in the time horizon.

Proposition 1 demonstrates that within each mode of operation, limiting points exist and can be written in closed-form. We note that although $q^*_{b_1}$ is piecewise defined, it does not affect our subsequent analysis. We select $q^*_{c_1}$ corresponding to the case of $q_b(t) \geq s_b$ because it describes the busiest time of operation in the launch phase. If $\boldsymbol{q}(t)$ remained in a particular mode of operation, it would move towards the corresponding limiting point. However, due to the time-varying nature of demand, $\boldsymbol{q}(t)$ may spend time in multiple modes of operation. As a result, we next establish an ordering relationship amongst the limiting points to provide a criterion for determining which mode of operation is better. In particular, we focus on the size of the customer base as $t \to \infty$ as this represents the steady-state number of customers that the firm retains over the long-term.

Lemma 5 (Asymptotic Monotonicity)**.** *Let $\mathcal{QN}_1$, $\mathcal{QN}_2$ and $\mathcal{QN}_3$ be queueing systems with identical parameters but different initial conditions, i.e., they originate in the launch, loyalty, and lessening regions, respectively; and neither system transitions into a higher mode of operation. Then, for any combination of queueing parameters,*

$$q^*_{c_1} > q^*_{c_2} > q^*_{c_3}.$$

Lemma 5 demonstrates that when $\boldsymbol{q}(t)$ is in the launch region, the limiting point that $q_c(t)$ is moving towards is larger than the corresponding limiting point when $\boldsymbol{q}(t)$ is in the loyalty region. Further, when $\boldsymbol{q}(t)$ transitions from the launch region to the loyalty region, $q_c(t)$ is moving to a larger limiting point than a system in which $\boldsymbol{q}(t)$ never reached the launch region. As a result, for long time horizons, $\mathcal{QN}_1$ is more managerially preferable than $\mathcal{QN}_2$.

Although Lemma 5 establishes an asymptotic ordering of the customer base between modes of operation, it does not imply that this relationship holds for any time t. In particular, at finite times the dynamics of the system are governed by two competing exponential functions, as described in (2.9). Monotonicity guarantees are difficult to establish during this period, and we define $\boldsymbol{q}(t)$ to be in a state of disorder. As both exponential terms approach zero, the solution becomes identical to its non-homogeneous value. Thus, the third stage of our approach describes the duration of the disorder period and, again, focuses specifically on the launch and loyalty regions.

Proposition 2 (Duration of Disorder Period)**.** *Let ψ_k be the smallest and χ_k be the largest negative eigenvalues of the homogeneous system of ODEs corresponding to* (2.11) *in mode $k \in \{1,2\}$. If $v(\psi_k)$ and $v(\chi_k)$ represent the first coordinates of the corresponding eigenvectors, then the duration of the disorder period of $q_{c_k}(t)$ is determined by the time required for the constant $\bar{w}_k = |w_k|e^{\chi_k t_0}$ to exponentially reduce to 0 with a given degree of tolerance, where*

$$w_k = \frac{v(\psi_k)[q_c(t_0) - q^*_{c_k}(t_0)] - q_b(t_0) + q^*_{b_k}(t_0)}{e^{\chi_k t_0}[v(\psi_k) - v(\chi_k)]} \quad \textit{for } k \in \{1,2\},$$

and t_0 is the time the system state switches into operating mode k.

Corollary 1. *In the launch region, if $q_{a_1}(t) \to q^*_{a_1}$, then $\bar{w}_0 = |w_0|e^{-\tau_a t_0} \to 0$ where*

$$w_0 = \left[q_a(t_0) - \frac{\lambda_1}{\tau_a - \delta}e^{-\delta t_0} - \frac{\lambda_0 + s_a(\tau_a - \mu_a)}{\tau_a}\right]e^{\tau_a t_0}.$$

For the final stage of our approach, we combine the above results to formulate the objective function for our stationary staffing policy. This considers both the limiting points and the duration of disorder results. Our goal is to formulate an optimization problem to determine s_a and s_b so as to maximize the time spent in the launch and loyalty regions. From Lemma 5, for sufficiently large t, it suffices to determine values s_a and s_b such that the limiting points are as large as possible. Which points to choose follow from the modes of operation. For the launch region, we maximize

$q^*_{a_1}$ and $q^*_{c_1}$ as the advertising campaign has just begun and the focus is on the acquisition of new customers. For the loyalty region, we maximize $q^*_{b_2}$ as, at this point, the focus of the firm has shifted to retention efforts because the rate of new arrivals has As per Proposition 1, the limiting points $q^*_{a_1}$, $q^*_{c_1}$, and $q^*_{b_2}$, are linear in the decision variables s_a and s_b. To combine them, we apply Proposition 2. That is, we multiply each limiting point by a convex combination of the magnitude of the exponential terms governing the length of the disorder period. Larger absolute values of the exponential terms, (i.e., $\bar{w}_0, \bar{w}_1, \bar{w}_2$) indicate that dynamics of the corresponding queues are more time-varying in the vicinity of the initial conditions. To ensure that lengthy queues do not form over these periods, we give more weight to state functions with greater time-variability. As a result, with a slight abuse of notation, we write down the objective of our optimization problem showing the dependency of each of the terms on the decision variables as follows:

$$\begin{aligned} z(s_a, s_b) = & \frac{\bar{w}_0(s_a)}{\bar{w}_0(s_a) + \bar{w}_1(s_a, s_b) + \bar{w}_2(s_a, s_b)|v(\chi_2)|}\rho_a q^*_{a_1}(s_a) + \frac{\bar{w}_1(s_a, s_b)}{\bar{w}_0(s_a) + \bar{w}_1(s_a, s_b) + \bar{w}_2(s_a, s_b)|v(\chi_2)|}\rho_b q^*_{c_1}(s_a, s_b) \\ & + \frac{\bar{w}_2(s_a, s_b)|v(\chi_2)|}{\bar{w}_0(s_a) + \bar{w}_1(s_a, s_b) + \bar{w}_2(s_a, s_b)|v(\chi_2)|}\rho_b q^*_{b_2}(s_b). \end{aligned} \tag{2.12}$$

We note that the correction factor $v(\chi_2)$ that multiplies $\bar{w}_2$ comes from the general solution of the homogeneous system of ODEs. Further, $\boldsymbol{\rho}$ is the relative average lifetime value associated with each customer class. Also, maximizing $z(s_a, s_b)$ ensures that the service system spends the least amount of time in the lessening phase, which is the least managerially desirable. The mixed-integer optimization problem is given by

$$\max_{s_a \in \mathbb{Z}_{\geq 0}, s_b \in \mathbb{Z}_{\geq 0}} z(s_a, s_b) \text{ subject to} \tag{EXP}$$

$$s_a + s_b \leq s, \tag{2.13}$$

$$s_b \geq \left(1 - \frac{\zeta}{r}\right)^+ \theta_{bc} s_a, \tag{2.14}$$

$$\bar{w}_0(s_a) \geq 0, \tag{2.15}$$

$$\bar{w}_k(s_a, s_b) \geq 0, \ k \in \{1, 2\}. \tag{2.16}$$

We note that the $\bar{w}_k(\cdot)$ terms implicitly depend on the staffing level through the initial conditions of the dynamical system. We make this relationship explicit by denoting them as functions of s_a and s_b. As a result, EXP cannot be solved by an off-the-shelf solver tool. Nevertheless, in Section 2.4.3, we prove several structural properties that lead to an efficient solution procedure.

Constraint (3.16) bounds the number of staff that can be assigned to each customer class. Constraint (2.14) defines a lower bound on the staffing level for returning customers. When the attrition rate from the orbit is lower than the rate at which base customers request service ($\zeta \leq r$), the firm benefits by sufficiently staffing station Q_b. The benefit can be quantified; it is equal to the value of having a server dedicated to new clients weighted by the probability that a base customer returns to the orbit after service, i.e., θ_{bc}. When the attrition rate from the orbit is higher than the rate at which base customers request service ($\zeta > r$) the value of assigning a large number of servers s_b to station Q_b diminishes. Finally, constraints (2.15) and (2.16) ensure that the system reaches the launch and loyalty regimes, respectively.

2.4.2 Clinical Setting: Sinusoidal

Assuming that patients arrive uniformly over time is a severe limitation in the modeling of health care systems. For instance, across a single day, many authors have found that the arrival rate in emergency wards is periodic, see e.g., De Bruin et al. (2007) and Green et al. (2007b). Green et al. (1991) model this periodicity by assuming that the arrival rate follows a sine function. This accounts for the cyclical pattern of demand in their study, however, its flexibility in mathematical modeling is particularly attractive as a linear combination of sine functions can be used to approximate any periodic and time-varying arrival process (Yom-Tov and Mandelbaum 2014a). Systems with periodic arrivals have also been analyzed by several researchers in the operations community (e.g., Whitt 2014, Liu and Whitt 2017) who use the form

$$\lambda(t) \equiv \bar{\lambda} + \bar{\lambda}\sigma \sin\left(\omega t + \phi\right), \tag{2.17}$$

where $\bar{\lambda} > 0$, $\sigma \in [0, 1]$, $\omega > 0$ and $\phi \geq 0$ are the average arrival rate, relative amplitude, frequency, and phase, respectively. We use a standard notation for the frequency of the sine function, i.e., $\omega = \frac{2\pi}{T}$, and assume the phase equals zero without loss of generality.

Substituting (3.12) into (2.8), we obtain the following IVP.

$$\begin{pmatrix} \dot{q}_a(t) \\ \dot{q}_b(t) \\ \dot{q}_c(t) \end{pmatrix} = \begin{pmatrix} \bar{\lambda} + \bar{\lambda}\sigma \sin(\omega t) - \mu_a(q_a(t) \wedge s_a) - \tau_a(q_a(t) - s_a)^+ \\ rq_c(t) - \mu_b(q_b(t) \wedge s_b) - \tau_b(q_b(t) - s_b)^+ \\ \theta_{ac}(q_a(t) \wedge s_a)\mu_a + \theta_{bc}(q_b(t) \wedge s_b)\mu_b + \theta_{\tau c}\tau_b(q_b(t) - s_b)^+ - (r + \zeta)q_c(t) \end{pmatrix}, \tag{2.18}$$

$$\boldsymbol{q}_0 = (q_a(t_0), q_b(t_0), q_c(t_0))^T.$$

Similar to Section 2.4.1, $q_a(t)$ can be written in closed-form while the solution to $q_b(t)$ and $q_c(t)$ remains as in (2.9). In contrast, however, the dynamical system described by (2.18) is bounded as per Lemma 4 and the asymptotic solutions may not converge to limiting points. Thus, we apply the four-stage approach, except now, we examine a system that converges to a periodic function.

For the first stage, we again argue that characterizing the behavior of the system in only a few regions is sufficient for developing a systematic approach to stationary staffing in this time-varying environment. That is, we partition the domain into three regions characterized by distinct modes of operation. The first mode has high rate of employee utilization. The second mode is the opposite, employees are underutilized. Finally, the third mode represents all other regions.

Mode 1 - Overloaded Region: $q_a(t) > s_a$ and $q_b(t) > s_b$,

Mode 2 - Underloaded Region: $q_a(t) \leq s_a$ and $q_b(t) \leq s_b$.

Mode 3 - Mixed Region: $q_a(t) > s_a$ and $q_b(t) \leq s_b$, $q_a(t) \leq s_a$ and $q_b(t) > s_b$.

The institutions goal is to provide service to as many customers as possible while ensuring servers are sufficiently utilized. Given the partition of the domain into modes of operation, the highest utilization of servers is achieved in the first mode. Thus, it makes sense to investigate how the institution can remain in the overloaded region while avoiding the second mode, i.e., the underloaded region. As a result, in the second stage of our approach, we examine the limiting behavior of (2.18) and find the extrema of these periodic orbits in these two regions of interest.

Lemma 6 (Asymptotic Dynamics)**.** *Suppose α_j, β_j and γ_j are real numbers for $j \in \{a, b, c\}$ whose values are dependent on the initial conditions of* (2.18)*. Then,*

1. *If* $\min q_a(t) \geq s_a$ *then there exists a set of equilibrium points*

$$(q_b^*, q_c^*) := \begin{cases} \left(\frac{r\theta_{ac}s_a\mu_a + s_b(\tau_b - \mu_b)(r+\zeta) - s_b r(\theta_{\tau c}\tau_b - \theta_{bc}\mu_b)}{\tau_b(r(1-\theta_{\tau c})+\zeta)}, \frac{\theta_{ac}s_a\mu_a + s_b\mu_b(\theta_{bc}-\theta_c)}{r(1-\theta_c)+\zeta}\right), & q_b(t) \geq s_b \\ \left(\frac{r\theta_{ac}s_a\mu_a}{\mu_b(r(1-\theta_{bc})+\zeta)}, \frac{\theta_{ac}s_a\mu_a}{r(1-\theta_{bc})+\zeta}\right), & q_b(t) < s_b, \end{cases} \tag{2.19}$$

while $q_a(t)$ *converges to a periodic orbit of the form*

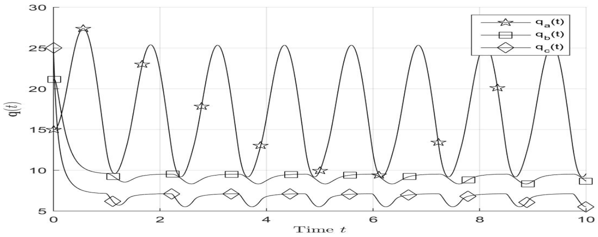
$$q_a^*(t) = \alpha_a + \beta_a sin(\omega t) + \gamma_a cos(\omega t).$$

2. *If* $\max q_a(t) \leq s_a$ *then* $\boldsymbol{q}(t)$ *converges to a set of periodic orbits of the form*

$$q_j^*(t) = \alpha_j + \beta_j sin(\omega t) + \gamma_j cos(\omega t) \text{ for } j = \{a, b, c\}.$$

Notice that Lemma 6 describes the asymptotic behavior of the system in two extreme cases. However, if $\max q_a(t) > s_a$ and $\min q_a(t) < s_a$, the dynamics cannot be so easily characterized. Figure 2.2 illustrates one such example; $q_b(t)$ and $q_c(t)$ do not converge to a sinusoid, rather, they are more complex periodic functions. Nevertheless, in all cases, including those where the dynamics cannot easily be characterized, the *extrema* of the periodic orbits can be derived.

Figure 2.2: Dynamics of $\bar{\boldsymbol{\rho}}\boldsymbol{q}(t)$ when $\max q_a(t) > s_a$ and $\min q_a(t) < s_a$



Lemma 7 (Asymptotic Bounds)**.** *The limiting solutions of* $\boldsymbol{q}(t)$ *obey the following relations*

$$q_a^*(t) \geq \underline{q}_a^* = \frac{\bar{\lambda}}{\mu_a} - \frac{\bar{\lambda}\sigma}{\mu_a^2 + \omega^2}\left(\mu_a sin\left[arctan\left(\frac{\mu_a}{\omega}\right)\right] + \omega cos\left[arctan\left(\frac{\mu_a}{\omega}\right)\right]\right),$$

$$q_b^*(t) \geq \underline{q}_b^* = \alpha_b + \beta_b sin\left[arctan\left(\frac{\beta_b}{\gamma_b}\right)\right] + \rho_b\gamma_b cos\left[arctan\left(\frac{\beta_b}{\gamma_b}\right)\right],$$

$$q_c^*(t) \leq \bar{q}_c^*.$$

Upper bounds for $q_a^*(t)$ and $q_b^*(t)$, and a lower bound for $q_c^*(t)$, can also be derived although we omit the results as they do not play a role in subsequent theory. Note that in order to describe when the bounds become tight, the second stage of the modeling approach also requires that the time points corresponding to when $q_a(t)$ is underloaded or overloaded be known.

Lemma 8 (Crossing Points)**.** *The state function $q_a^*(t)$ crosses the boundary Σ at times*

$$t = \frac{2arctan\left(\frac{\beta_a \pm \sqrt{\beta_a^2-(s_a-\alpha_a+\gamma_a)(s_a-\alpha_a-\gamma_a)}}{s_a-\alpha_a+\gamma_a}\right)}{\omega} + \frac{2\pi}{\omega}n, \text{ where } n = 0, 1, 2 \ldots.$$

Although similar expressions for $q_b^*(t)$ and $q_c^*(t)$ can be derived, only the crossing points of $q_a^*(t)$ are important as they drive the asymptotic behavior of the system. An implication of Lemma 8 is that the proportion of time $q_a^*(t)$ spends underloaded or overloaded per period can be obtained. Specifically, if $t_1 < t_2$ are two adjacent time points when $q_a^*(t)$ crosses Σ from above and below respectively, $\kappa := (t_2 - t_1)/T$ is the proportion of time $q_a^*(t)$ is underloaded. Thus, $\boldsymbol{q}(t)$ spends κT time units per period of $q_a^*(t)$ in regions where the lower bounds in Lemma 7 are tight. Similarly, $\boldsymbol{q}(t)$ spends $(1-\kappa)T$ time units per period in regions where the upper bound is tight. Note that to compute the crossing points for the value of the new customers in the system, $\rho_a q_a(t)$, we need only multiply the staffing level s_a by $\frac{1}{\rho_a}$ in Lemma 8. We denote the proportion of time that $\rho q_a(t)$ is underloaded each period by κ_ρ.

The third stage of our approach accounts for the behaviour of $\boldsymbol{q}(t)$ at finite times. As in Section 2.4.1, the dynamics of the system are initially governed by two competing exponential functions and during this period, $\boldsymbol{q}(t)$ is said to be in a state of disorder. The exponential functions approach zero as $t \to \infty$ and $\boldsymbol{q}(t)$ converges to its non-homogeneous term, which by Lemma 6, is a periodic function or, in some cases, may be an equilibrium point. Because the duration of the disorder period is determined by the homogeneous part of the general solution in (2.9), Proposition 2 still holds, i.e., the homogeneous systems corresponding to (2.11) and (2.18) are identical.

Thus, the final stage of our approach combines the results from stage two and three by assigning weights, derived in Proposition 2, to the asymptotic bounds from Lemma 7. The goal is to formulate an optimization problem to determine s_a and s_b so as to maximize the time spent in the overloaded region and minimize the time spent in the underloaded region. Using similar logic as in Section 2.4.1, we develop an objective function using $\underline{q}_b^*$ and $\bar{q}_c^*$ (i.e., the asymptotic bounds) multiplied by a convex combination of weights $\bar{w}_2(s_a, s_b)$ and $\bar{w}_1(s_a, s_b)$. Because $\bar{q}_c^*$ is reached when $q_a^*(t)$ is overloaded, we multiply $\bar{q}_c^*$ by the proportion of time per period that $q_a^*(t)$ spends in the overloaded state, $(1 - \kappa)$. Similarly, we multiply $\underline{q}_b^*$ by κ to capture the amount of time per period $q_a^*(t)$ spends in the underloaded state. With a slight abuse of notation, the objective is

$$z(s_a, s_b) = (1 - \kappa_\rho(s_a)) \frac{\bar{w}_1(s_a, s_b)}{\bar{w}_1(s_a, s_b) + \bar{w}_2(s_a, s_b)} \rho_b \bar{q}_c^*(s_a, s_b) + \kappa_\rho(s_a) \frac{\bar{w}_2(s_a, s_b)}{\bar{w}_1(s_a, s_b) + \bar{w}_2(s_a, s_b)} \rho_b \underline{q}_b^*.$$

Notice that maximizing $z(s_a, s_b)$ is akin to serving as many customers as possible while ensuring that servers are highly utilized. That is, maximizing the first term increases the number of customers in the orbit while maximizing the second reduces the number of idle servers by increasing the load of the system in the underloaded state. The mixed-integer optimization model is

$$\max_{s_a \in \mathbb{Z}_{\geq 0}, s_b \in \mathbb{Z}_{\geq 0}} z(s_a, s_b) \;\text{ subject to} \tag{SINE}$$

$$s_a + s_b \leq s, \tag{2.20}$$

$$s_b \geq \left(1 - \frac{\zeta}{r}\right)^+ \theta_{bc} s_a, \tag{2.21}$$

$$\bar{w}_k(s_a, s_b) \geq 0, \;\; k \in \{1, 2\}. \tag{2.22}$$

where (2.20) and (2.21) remain as in Section 2.4.1 and (2.22) is the analog of the constraints (2.15)-(2.16). We again note that the $\bar{w}_k(s_a, s_b)$ terms implicitly depend on the staffing level through the initial conditions of the dynamical system.

2.4.3 Solution Approach

In both EXP and SINE, the objective functions are convex combinations of terms associated with the length of the disorder period. Unfortunately, these terms are implicit functions of the decision variables and thus, a simple mixed-integer linear program cannot be solved. Nevertheless, in this section, we show that the problems exhibit structural properties that are conducive to formulating a solution procedure that efficiently determines the optimal staffing levels.

First, the objective functions are neither convex nor concave. Thus, a straightforward relationship between the initial conditions of the dynamical system and the associated weights in the objective function cannot be exploited in an efficient solution approach. Instead, we prove that there exists a small interval within the feasible region that contains the optimal solution. As a result, we only need to perform an exhaustive search within this small interval to find globally optimal solutions.

For EXP, notice that in Proposition 3, by requiring that $q_a(t)$ transitions into the overloaded regime, we ensure that the optimal staffing level s_a^* cannot be greater than the maximum of $q_{a_2}(t)$ which corresponds to the loyalty phase. Thus, we obtain the following result.

Proposition 3. *The optimal staffing level s_a^* for EXP is contained in the set*

$$\mathcal{S}_{exp} := \left\{ s_a \in \mathbb{Z}_{\geq 0} \middle| s_a \geq \frac{\lambda_a}{\mu_a}, s_a \leq \begin{cases} q_{a_2}(t^*), & \text{if } t^* \geq 0, \\ q_a(t_0), & \text{otherwise.} \end{cases} \right\},$$

where

$$t^* = \frac{log\left(-\mu_a\left[q_{a_2}(t_0) - \frac{\lambda_0}{\mu_a} - \frac{\lambda_1}{\mu_a-\delta}e^{-\delta t_0}\right]e^{\mu_a t_0}\right) - log\left(\frac{\delta\lambda_1}{\mu_a-\delta}\right)}{\mu_a - \delta}.$$

Further, if $\lambda_0 \to \infty$ as $\lambda_1 \to \infty$, then $|\mathcal{S}_{exp}| \to 0$.

Proposition 3 provides a range of values where the optimal solution is guaranteed to be found. Notice that this set is much smaller than the size of the feasible region corresponding to EXP. It also indicates that as λ_0 and λ_1 increase, the number of candidate solutions decreases, i.e., $\mathcal{S}_{exp}$ becomes empty.

Similarly for SINE, notice that in Proposition 4, we require that $q_a(t)$ must periodically reach an overloaded state. As a result, if s_a^* is greater than the upper bound of the underloaded queue $q_a(t)$, then the overloaded state is never reached. Further, if s_a^* is less than the lower bound of the underloaded queue $q_a(t)$, then $q_a^*(t)$ will be overloaded indefinitely. According to the following result, if $\bar{\lambda}$ increases as fast as σ decreases, the number of candidate solutions decreases, i.e., $\mathcal{S}_{sine}$ becomes empty.

Proposition 4. *The optimal staffing level s_a^* for SINE is contained in the set*

$$\mathcal{S}_{sine} := \left\{ s_a \in \mathbb{Z}_{\geq 0} \middle| s_a \geq \min \underline{q}_a^*(t), s_a \leq \max \underline{q}_a^*(t) \right\},$$

where $\underline{q}_a^*(t)$ corresponds to $q_a^*(t)$ in the underloaded state. If $\sigma \to 0$ as fast as $\bar{\lambda} \to \infty$, $|\mathcal{S}_{sine}| \to 0$.

Proposition 3 and 4 represent necessary conditions for determining the optimal solutions to EXP and SINE. They also suggest that as the total number of servers gets large and the requirements of the propositions are satisfied, the number of candidate solutions decreases to 0. Thus, interestingly, as the system size grows, relatively fewer points need to be evaluated to find the optimal solution.

Our solution approach begins by computing the upper and lower bounds of the sets $\mathcal{S}_{exp}$ and $\mathcal{S}_{sine}$. Note that because assigning less than s servers is suboptimal, constraint (3.16) and (2.20) are always binding. Then, starting from the upper bound, we iteratively decrease s_a and simultaneously increase s_b until constraint (2.15) and (2.16) for EXP and constraint (2.22) for SINE is no longer satisfied or we reach the lower bound of the sets $\mathcal{S}_{exp}$ and $\mathcal{S}_{sine}$. Constraint (2.14) and (2.21) may also reduce the upper bound of the feasible solution region if they are violated. During each iteration, we solve (2.11) for EXP and (2.18) for SINE using the given staffing level. We then compute the objective function value and determine whether it is the maximum value observed so far. If EXP or SINE are infeasible, then we assign the value of the upper bound $\mathcal{S}_{exp}$ or $\mathcal{S}_{sine}$ to s_a^*.

2.4.4 Model Discussion

The proposed four-stage procedure only requires that the arrival process be modeled by a smooth function. Thus, the approach is general, and its implementation is straightforward: to derive a staffing policy, it suffices to analyse the asymptotic behaviour of (2.8), weight its limiting points and solve a simple optimization problem. Further, notice that we do not assign waiting costs or penalties to customers that abandon: doing so results in complex optimal control problems (e.g., Anderson Jr et al. 2006) with limited tractability in cases with piecewise smooth systems. Massey and Pender (2018), for instance, propose a Lagrangian profit maximization based stationary staffing procedure for Erlang-A systems. Instead, we identify preferable operating regimes (e.g., higher values of q_c^*, $\bar{q}_c^*$ and $\underline{q}_b^*$) and determine the staffing policies that induce them. The technique is similar to identifying parameter values that result in desirable steady-state dynamics (see, e.g, Pender et al. 2017). The inclusion of relative weights for staffing generalizes the overall premise of lifetime value that is present in the CRM literature and makes the approach valid for environments featuring either equally valuable or prioritised classes of customers. For example, if base customers are more profitable (i.e., spend more time per visit or pay a subscription fee), it might be better for the firm to focus more on their abandonment. Conversely, if their switching cost is high, and thus their loyalty, the firm can focus more on serving new customers.

We assign larger weights to the limiting points of state functions with greater variability during early stages of the time horizon. That is, the weights in (2.9) characterize the deviation of the solution from its steady-state caused by the initial conditions: larger weights correspond to more variation prior to the solution approaching its limiting behaviour. The result is that queues with relatively high variation will be prioritized by EXP and SINE when assigning servers. Weighting the points in this fashion is similar to keeping safety stock in inventory systems with demand uncertainty, i.e., products with higher demand variability have a larger amount of safety stock.

2.5 Numerical Results

In this section, we present a comprehensive numerical study of the staffing policies generated by optimization problems EXP and SINE. We compare their performance to modifications of the tail probability of delay (TPoD) staffing policy developed in Liu (2018), a generalized version of the square root staffing (SRS) policy, by varying customer patience levels, the speed of service, and the frequency of service requests. We also evaluate the magnitude of deviation of EXP and SINE from the optimal staffing policy (OPT) determined by simulating all valid staffing levels in the stochastic regime and choosing the policy that maximizes the throughput or minimizes the total number of class a and b clients who abandon the system. Finally, we identify cases when EXP and SINE outperform these benchmarks and discuss its implications for management theory.

We fix the attrition rate of base customers from the orbit to $\zeta = 1$ and set the probabilities of congestion and fitness abandonment to $\theta_c = 0.5$ and $\theta_{ic} = 0.9$ for $i = \{a, b\}$, respectively. This ensures that there is a high-likelihood of retention after a successful service completion while a substantial chance that we will be unable to retain clients in a congested system; our results are qualitatively similar as θ_{ic}, $i \in \{a, b, \tau\}$, is varied between 0.5 and 1 (see Afèche et al. 2017, for instance). Relative to the attrition rate ζ, we introduce fast, intermediate, and slowly varying forms of the exponential (Section 2.4.1) and sinusoidal (Section 2.4.2) arrival functions (see Appendix

A "Simulation Parameters" for more details). The parameters are selected to ensure that the maximum intensity of arriving customers is significantly larger than the attrition rate.

We choose a pool of twenty servers ($s = 20$) and assign them to new and returning customers. This experimental setup is similar to Chevalier and Tabordon (2003), for instance, who consider a pool of 6 specialized servers. We also conduct simulations of larger systems ($s = 40$ and $s = 60$) to study how performance scales with the size of the system. We vary the service rate (μ_i) and abandonment rate (τ_i) for $i = \{a, b\}$ and the rate that base customers request service from the orbit (r). The rates are chosen so that higher values are larger than the attrition rate and the smallest values remain close to it. We consider scenarios of high and low abandonment, service and arrival rate of returning clients (see Appendix A "Simulation Parameters"). We repeat these trials for two time horizons; long ($t = 25$) and short ($t = 10$). This dichotomy is consistent with the literature (see Liu and Xie 2018, for instance). Each experiment begins with a burn-in period of $t = 5$ time units after which we record performance measures associated with service quality. This includes the average number of new and returning customers served (throughput) and the average total number of customers who abandon. Both metrics are commonly used in the literature to evaluate the performance of fluid (deterministic) queueing systems (e.g., Buzacott and Shanthikumar 1993, Bassamboo and Randhawa 2015). All reported values represent the average over 10 stochastic simulations. The staffing policies and discrete-event simulation were programmed in Matlab R2015a.

We compare our stationary staffing policies, EXP and SINE, to four benchmarks that are stationary or near-stationary. The first, $GSRS_0$, is a generalized square root staffing policy that does not allow any reassignment of staff over the time horizon. The second, $GSRS_t$, is similar except that it allows staff to be reassigned a fixed number of times. Both are modifications of the TPoD staffing policy. More specifically, for the stationary benchmark, we compute staffing levels according to

$$s_a \coloneqq s_h + \xi\sqrt{s_h}, \qquad s_b \coloneqq s - s_a, \tag{2.23}$$

and

$$\xi \coloneqq z_\chi\sqrt{\frac{\tau_a}{\mu_a}}, \quad s_h \coloneqq \frac{\bar{\lambda}_a}{\mu_a}e^{-\tau_a h}$$

where $\bar{\lambda}_a$ is the average arrival rate to Q_a, μ_a is the average service rate of new customers, and ξ

is the desired level of service. We compute z_χ by inverting the standard normal distribution for a specified probability of delay χ. Following Gans et al. (2003), Brown et al. (2005), Aksin et al. (2007), and Liu (2018), we adhere to the rule that 80 percent of service requests are served within 1/3 of a time unit. Thus, we fix the probability of delay χ and delay parameter h to 0.2 and 1/3, respectively. We then implement (2.23) by computing $\bar{\lambda}_a$ over the entire time horizon.

For our non-stationary benchmark, the staffing levels follow the equations in (2.23) once again. However, for the advertisement campaign, $\bar{\lambda}_a$ is computed by splitting the time horizon into two equal intervals. For the clinical setting, we compute $\bar{\lambda}_a$ at times when $\lambda(t)$ is above and below its symmetry line. This ensures that the staffing policy switches once per period of $\lambda(t)$. For both the advertisement campaign and the clinical setting, our non-stationary benchmarks re-allocate the staff over the intervals characterized by a higher and lower intensity of arrivals of new customers, accordingly. We implement two versions of $GSRS_0$ and $GSRS_t$, with $\tau_a = 0$ and $\tau_a \neq 0$, corresponding to the Erlang-C and Erlang-A versions of the TPoD, respectively.

2.5.1 Advertising Campaign: Results

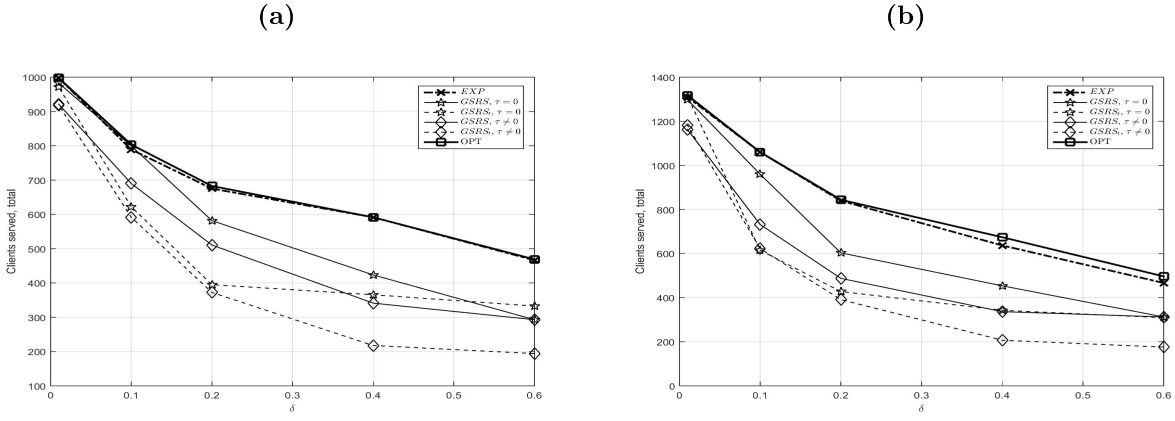
In this setting, the amount that demand varies with time is determined solely by the decay rate of the exponential (i.e., δ) in (2.10). Systems that slowly converge to a stationary state have small values of δ and are characterized by low levels of non-stationarity whereas systems that quickly reach a steady state solution are more time-varying and have large values of δ. Because our results are qualitatively similar for both short and long horizons, we only discuss cases where $t = 25$.

For high levels of non-stationarity ($\delta \geq 0.6$) and varying levels of traffic intensity, EXP outperforms $GSRS_0$ and $GSRS_t$, i.e., throughput is better and total abandonment is similar or lower. For example, when the abandonment rate is low (clients are patient), the throughput of EXP is 1.3 and 1.25 times larger than the $GSRS_0$ and $GSRS_t$, respectively. Remarkably, $GSRS_0$ and $GSRS_t$ with $\tau_a \neq 0$ assign too few servers to new clients which results in poor performance (for additional results, see Appendix A "Summary of Results for EXP, $s = 20$").

As the system becomes less congested, the accuracy of the fluid approximation naturally decreases which reduces the performance gap among the policies. This explains why their throughput, for example, equalizes in the "slow server" setting. In systems with faster servers, the traffic intensity is significantly higher due to the small number of servers assigned to new clients.

We note that as the system becomes more stationary (i.e., δ decreases) the performance of the benchmarks improve and the gap between all staffing policies decrease. We demonstrate this in Figure 2.3 where the performance of the staffing policies for the larger systems are presented. In both settings, $s = 40$ and $s = 60$, the throughput equalizes as the value of δ becomes small whereas the gap widens as δ increases. Moreover, the gap amongst all policies increase in the system size, i.e., the performance of EXP improves as the system becomes larger. Finally, we observe that the performance of EXP is close to OPT and the gap remains small as the system gets larger.

Figure 2.3: Total throughput for EXP with $t = 25$ for the systems where base customers require frequent service: (a) $s = 40$; and (b) $s = 60$.



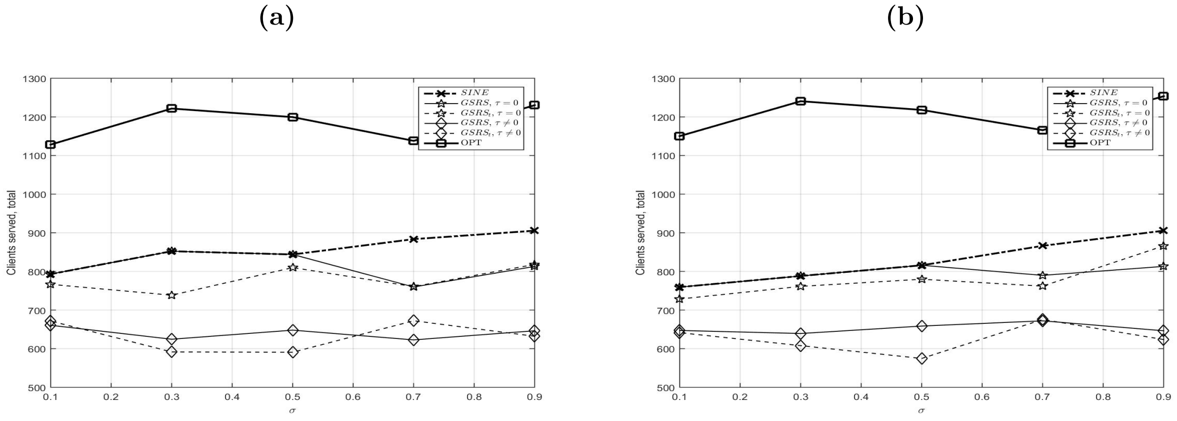
Because Figure 2.3 displays the typical performance of the policies, it also demonstrates that adjusting a staffing policy partway through the time horizon does not increase throughput. In fact, doing so decreases throughput, i.e., it reduces service accessibility. This is because, as observed in our simulations, EXP does not singularly focus on new customers. Instead, it attempts to strike an equitable balance between the acquisition of new customers and the retention of existing ones. Similar results are observed across all simulation experiments which indicates that, in contrast to the benchmarks, EXP ensures high levels of service quality for all customer types.

2.5.2 Clinical Setting: Results

When demand is periodic, the degree of non-stationarity is determined by the cycle length T and the relative amplitude σ of the sinusoid function. Following Green et al. (1991), systems that exhibit high levels of time-varying behavior have short cycle lengths and large amplitude levels.

Our results are qualitatively similar in this setting as compared to Section 2.5.1. We again confirm that for the most time-varying regimes, when values of σ and T are large and small, respectively, SINE outperforms the benchmarks in all queueing regimes (for additional results, see Appendix A "Summary of Results for SINE, $T = 2, 8$, $s = 20$'). This effect is demonstrated in Figure 2.4. In this setting, for the systems $s = 40$ and $s = 60$, the performance of SINE improves as σ increases. Once again, as the system becomes more stationary, i.e., σ decreases or T increases, the performance of all the policies equalizes. According to this figure, there is a gap between OPT and the rest of policies but it does not grow as the system size increases.

Figure 2.4: Total throughput for SINE with $t = 25$ and $T = 2$ for the systems where base customers require frequent service: (a). $s = 40$; and (b). $s = 60$



Finally, we again observe that re-assigning staff partway through the period of the arrival function does not improve throughput and performs worse than SINE . This result is consistent with our findings from the previous section in that the performance of $GSRS_t$ is not better than $GSRS_0$.

2.6 Summary of Contributions

We contribute to the operations research literature by showing that our approach to staffing succeeds in cases where standard near-stationary methods fail (see, e.g., Green et al. 1991, Green and Kolesar 1991). That is, our policies perform as well as the benchmarks in more stationary settings and substantially better in environments characterized by high levels of time variability and traffic intensity. In particular, our methodology ensures that neither new nor returning customers are neglected - a common problem that is present in the existing literature, e.g., King et al. (2016), Afèche et al. (2017) - while high throughput levels are maintained. Our method is also more time-efficient as compared to computer-based approaches that simulate multiple trajectories for each staffing level and then use the results to derive the optimal server split.

Our analysis demonstrates that stationary policies may perform as well as time-dependent alternatives when the time-varying dynamics of the system are accounted for. Specifically, the results from our numerical study suggest that there is little benefit to re-assigning staff partway through an advertising campaign. Further, re-assigning staff partway through a period, in a periodic environemtn such as a clinic, is not useful either. In both cases, the benchmarks tend to severely overestimate or underestimate the demand for either new or base clients and the quality of service deteriorates as a result. Obviously, if more adjustments are allowed, higher throughput values can be achieved. However, frequently re-assigning staff is costly - both cognitively and financially - and becomes less operationally feasible. Thus, our work has important practical implications: managers who want to minimize the number of staffing changes can be empowered to make well-performing assignments using our stationary alternative while still accounting for time-varying dynamics.

We contribute to the literature on customer acquisition and retention by providing new insights into the field of customer relationship management (CRM). For example, King et al. (2016) suggest that companies with a larger customer base should begin to invest less in acquisition activities. This is because retaining too many customers becomes expensive and acquiring new ones is no longer essential; growth is not a priority for larger firms. We refine this argument by demonstrating that either customer acquisition or retention efforts, if neglected, can reduce a customer's ability to access services. Thus, cost savings may come at the expense of service quality, and this affects customers' overall perception of the company's value. This perception, in turn, may have implications for the future: customers may begin to leave for no specific reason (Griffin 2001, Gee et al. 2008) and the ability to acquire new and retain base clients may be hindered.

Moreover, the need to serve both classes of customers at all times confirms a link between customer acquisition and retention efforts found in empirical studies but not captured in the operations literature. According to marketing researchers, assuming that these processes are independent leads to incorrect estimations of the duration of the relationship between a customer and the firm, which is important for retention analysis (Thomas 2001). Our approach explicitly models this dependence and suggests that focusing on acquisition or retention efforts alone is not sufficient to maintain high levels of throughput and service quality (e.g., Afèche et al. 2017). Instead, we provide a decision support tool for staffing that balances these competing priorities.

In addition, the high-traffic environments where our staffing approach performs well are important regimes for customer retention. Specifically, successful services attract many customers and it is natural that they, in operationally efficient organizations, will experience some effects of congestion (e.g., customers who must wait). According to Lu et al. (2013), although periods of high congestion are correlated with high sales, a moderate increase in the number of queued customers generates sales reduction equivalent to a 5% price increase. Campbell and Frei (2011) similarly show that longer waiting times affect customer satisfaction and retention. We demonstrate that our approach gives customers more access to company services which reduces abandonment and as a result, long wait times. Thus, it follows that our approach to staffing, in addition to better balancing acquisition and retention efforts, can lead directly to increased revenues.

Finally, We note that our approach is amenable to profit and/or lifetime value considerations that are common in the CRM literature. In particular, when all clients are equally valuable, our staffing policy maximizes the common good, i.e., it improves the service quality for all customers equally. Conversely, if classes are not equally valuable, our method can easily account for this by assigning greater weight to more profitable customers classes. We achieve this by circumventing classical approaches that use optimal control theory, which are notoriously difficult to solve efficiently.

Chapter 3

Optimal Capacity Planning for Cloud Service Providers with Periodic, Time-Varying Demand

The work in this chapter has been submitted for publication as the following:

Eugene Furman, Adam Diamant (advisor) "Optimal Capacity Planning for Cloud Service Providers with Periodic, Time-Varying Demand", in *Manufacturing & Service Operations Management* , [2020]

3.1 Introduction

Data-driven decision-making and the explosion in operational processes guided by artificial intelligence requires significant amount of data and large amounts of processing capacity. For these reasons, the amount of data collected by global enterprises has been growing rapidly; it will reach 175 trillion gigabytes by 2025 which is 5.3 times larger than in 2018 (Reinsel et al. 2018). Much of this data must be processed in real-time, and thus, cloud computing infrastructure with near zero-latency has become an essential service for many corporations (Bruckner and Tjoa 2002). Unsurprisingly, to support these challenging service requirements, the worldwide cloud service market is expected to grow by 15% year-over-year to reach $354 billion in 2022 (Gartner 2019).

Many firms, however, have been expanding their cloud computing capacity with little regard to the consequences associated with idle resources, i.e., the housing, provisioning, maintenance, powering, and cooling of servers. Specifically, idle capacity can result in a massive amount of wasted energy which is both a financial and environmental concern (Koomey et al. 2007, Uchechukwu et al. 2014). The scale is astonishing; 40% of cloud service providers pay for a larger capacity than needed while some companies find that their average CPU utilization is only around 5% (Chapel 2019). On a global scale, it is estimated that there are approximately 10 million idle servers which translates into $30 billion worth of idle infrastructure (Forbes 2015). Alarmingly, idle servers represent over 35% of all cloud computing capacity (FLEXERA 2019).

Motivated by the systemic underutilization of servers for cloud computing, we consider a major provider of such services that has become acutely aware of this issue. Cloud computing capacity is allocated to different services for internal company use and to external subscribers on-demand. Dynamically changing this allocation is not feasible as provisioning additional CPU units either requires purchasing expensive hardware or re-allocating computing resources within the organization; such changes cannot be processed immediately and require several hours before they can take effect. Promptly analysing large volumes of data is critical for their operation, however, they also believe that a more data-driven approach to provisioning resources can reduce their idle capacity, decrease costs, and promote a more sustainable operations (e.g., reduce energy consumption).

To address this issue, we represent the aggregate capacity of the company's cloud computing infrastructure by the total number of homogeneous cores that it maintains, i.e., individual processors within a CPU. Customers arrive stochastically to the system to gain access to the company's cloud capacity. As indicated by our industry partner, arriving customers request one CPU core at any instant with a random service duration associated with the size of the job. If the job does not immediately begin processing upon the customer's arrival to the system, it leaves and requests capacity once again after some time interval has passed. Thus, customers are infinitely impatient and requests that are not immediately serviced may retry several times before permanently abandoning the system. To capture the non-stationary nature of demand, we assume that customers arrive according to a non-homogeneous Poisson process. Exact analysis of the stochastic system is intractable, and thus, we analyze the behaviour of a corresponding system with infinite capacity, i.e., the offered load (Zychlinski et al. 2018). By analysing its fluid approximation and applying a novel calculus-of-variations approach, we determine the jointly optimal service capacity and retrial interval between unsuccessful attempts given a pre-specified service level. In addition, we demonstrate that our methodology can incorporate constraints that give preference to certain types of workload when the system is critically loaded, i.e., utilization exceeds a desired service-level.

Using a data set of cloud computing requests over a representative 24-hour period from a typical service of our partner organization, we conduct a case study to demonstrate the practicality of our methodology. We show that demand for cloud computing capacity is time-varying and exhibits several peaks throughout the course of a single day. Using our approach to estimate the optimal service capacity and retrial interval of unsuccessful requests, we argue that our industry partner can significantly reduce their aggregate cloud computing capacity without affecting service. We show that this reduction has significant cost implications and also contributes to their vision of becoming a more sustainable organization and reducing their carbon footprint.

We contribute to the operations research literature by extending previous work on stationary staffing policies for queueing systems with a finite number of retrials, and time-varying, periodic demand. To this end, we introduce a new method that allows us to gain insight into the dynamics of an otherwise intractable system. In particular, our approach of minimizing the variation in the offered load can be applied to many systems with throughput-based objectives and/or service level agreements; it can also accommodate a broad class of piecewise smooth arrival functions that are general enough to describe most periodic demand processes in service settings. We contribute to the literature on sustainable operations by proposing a quantitative technique that reduces the idle capacity of a cloud service provider without compromising service quality. As high energy consumption results in operational costs, it also leads to high carbon emissions (Chang et al. 2010, Garg and Buyya 2012, Chang et al. 2016). Thus, our optimal policy is an energy-efficient solution that can be used to reduce the environmental impact of a cloud computing service.

3.2 Literature Review and Contribution

Our paper introduces a deterministic fluid model for a multi-server queueing system with a finite number of retrial opportunities, i.e., customers who enter the system and find that all servers are busy do not queue. Instead, they re-enter the system at a later time and attempt to obtain service. Pioneering work in the study of retrial queues dates back to Cohen (1900), Wilkinson (1956), and Falin and Templeton (1997). Since then, most papers assume an infinite number of retrial attempts, i.e., customers can attempt to obtain service until capacity becomes available (Sung and Chae 2000, Artalejo and Falin 2002, Krishnamoorthy and Ushakumari 2002, Artalejo 2010). From a theoretical perspective, these analyses typically focus on approximations and state-space truncation methods. Few papers derive fluid approximations for these systems (e.g., Kang et al. 2010b, Kang 2015) as they involve the use of indicator functions and it is hard to motivate the resulting dynamical system as the limit of the stochastic network. There is, however, a large body of literature that analyzes queueing systems with feedback (i.e., customers who have multiple service interactions) by incorporating infinite capacity orbits (de Véricourt and Jennings 2011, Jacobson et al. 2012, Yom-Tov and Mandelbaum 2014a, Ding et al. 2015, Huang et al. 2015, Furman et al. 2019). Our analysis builds on this work in that we introduce an offered load approximation for a multi-server stochastic system with a finite number of retrial attempts using these infinite capacity orbits. Thus, the fluid limit is well-motivated and widely accepted in the literature (Halfin and Whitt 1981). For instance, Zychlinski et al. (2018) use an offered load approximation for capacity allocation decisions in hospitals affected by bed blocking. We examine the workload dynamics of the original system without limiting its size or imposing unrealistic retrial behaviour.

We determine the optimal stationary staffing policy for a queueing system in anticipation of time-varying dynamics. Many authors investigate queueing systems with arrivals that follow a non-homogeneous Poisson process; see the survey papers by Defraeye and Van Nieuwenhuyse (2016), Whitt (2016), and Whitt (2018). Several studies investigate stationary or near stationary staffing of queueing systems with time-dependent arrival processes; see Harrison and Zeevi 2005, Bassamboo and Zeevi 2009, Bekker and de Bruin 2010, Defraeye and Van Nieuwenhuyse 2013 and Niyirora and Pender (2016) for example. The problem of staffing for a dynamic service system has also been an active area of research (e.g., Henderson et al. 1999, Akcali et al. 2006, Bhandari et al. 2008, Robbins and Harrison 2010). Further, as shown in Aguir et al. (2008) and Pustova (2010), the impact of retrial behavior on service capacity is difficult to model. To this end, Janssen and van Leeuwaarden (2015) propose a modification of the square-root staffing (SRS) (see, for instance, Borst et al. 2004, Feldman et al. 2008a, Hampshire et al. 2009, Janssen et al. 2011) that considers a queue with an infinite number of retrial attempts. To the best of our knowledge, our work is unique in that it determines the jointly optimal static service capacity and retrial interval for a system with time-varying arrivals, a finite number of retrial attempts, and pre-specified service levels. That is, we solve an optimization problem that minimizes the variation in total workload subject to a set of operational restrictions during peak demand periods. We note that our proposed solution approach is quite general; it can accommodate a large class of piecewise smooth arrival functions, and thus, can be applied to other management science applications.

Our solution approach minimizes the variation in the offered load. Mathematical applications that employ techniques from the family of "calculus-of-variations" methods are broad. For instance, Bioucas-Dias and Figueiredo (2010), Mattingley and Boyd (2010), and Zhao et al. (2014) minimize the total variation of a functional in order to remove noise for signal processing. Shu (1988), Ryu et al. (1993), and Gottlieb and Shu (1998) apply the diminishing total variation approach to develop numerical solvers for ordinary and partial differential equations. Zhu et al. (2008) use the total variation technique to propose a cancer treatment planning algorithm for intensity modulated radiation therapy. Minimizing the total variation, although suitable for a variety of applications, is typically onerous (see, for instance, Li et al. 2013 and Goldstein et al. 2014). Our work applies this approach to a management science problem and demonstrates that variation minimization can be naturally linked to more familiar queueing concepts, i.e., throughput maximization.

We contribute to research on capacity allocation for technological services by applying our methodology to the operations of a cloud service provider in order to address their excess cost and sustainability concerns. Research on cloud computing has been growing as the industry continues to expand. Many studies determine the appropriate cloud capacity levels when demand is volatile. For instance, Jiang et al. (2012), Dorsch and Häckel (2014), and de Assunção et al. (2016) propose real-time capacity allocation schemes that can dynamically adjust depending on the utilization of servers. The objective of these methods is to maximize revenue or minimize costs by solving a stochastic dynamic program. Several studies model cloud computing operations as a queueing system (for example, Khazaei et al. 2011, Vilaplana et al. 2014, Chiang and Ouyang 2014). This literature focuses on classical queueing-theoretic approaches which model cloud services as Jackson networks. They impose Markovian assumptions on the arrival process, often substitute the retrial feature with feedback, and determine the service capacity of the cloud by minimizing long-run costs. Other work focuses on establishing dynamic pricing policies to maximize revenue (e.g., Wang et al. 2010, Jin et al. 2014, Chen et al. 2019). Finally, a large body of literature is dedicated to the growing sustainability concern related to the operations of cloud services. For example, Pan et al. (2010), Garg and Buyya (2012), and Kalange Pooja (2013) describe features of the cloud that could be implemented to reduce its energy footprint. Dorsch and Häckel (2012), for instance, connect the sustainability discussion to optimal service capacity by suggesting that idle servers be employed for other tasks. Our work extends these studies by proposing a queueing-based methodology that incorporates more realistic assumptions (e.g., time-varying demand, stationary capacity provisioning, and generally distributed retrial times) and provides a cloud service provider guidance when determining the optimal trade off between the amount of idle capacity to provision (excess cost and energy consumption) and the operational protocols to implement during peak periods.

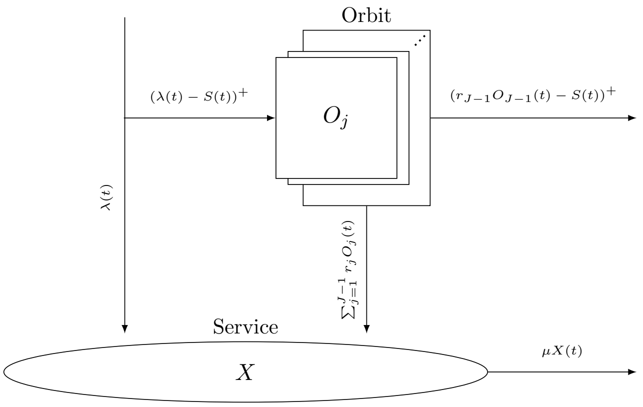
3.3 Model Formulation

In this section, we introduce a general stochastic queueing network and describe its suitability in modeling the cloud service operation of our partner organization. In particular, every few weeks our partner organization estimates demand for the upcoming period and cloud capacity $s \in \mathbb{R}_{>0}$ is provisioned. Once capacity is determined, jobs arrive and are serviced by the processing cores of remote machines. While the peak number of computing requests during a one-hour period can exceed four mullion, demand for cloud services is highly variable and exhibits a periodic pattern that repeats daily. When there is idle capacity, jobs are serviced immediately upon entering the system. At times when capacity is fully utilized, a job may make multiple computing requests in order to obtain service; the time interval between retrials is set by our partner organization.

We model this system as a multi-station queueing network with homogeneous, fully flexible servers (processing cores) and exogenous, time-varying arrivals (jobs). In this setting, customers submit jobs that require a single CPU core. New customers (Class-0) arrive to the system (station X) according to a non-homogeneous Poisson process with periodic, time-varying intensity $\lambda \equiv \lambda(t)$ that repeats daily. We assume that the jobs they submit have service requirements which are exponentially distributed with rate μ. If upon arrival to the system all servers are busy (i.e., all cores are processing jobs), customers do not queue, but instead, join a retrial orbit (station O). There are $J-1$ retrial orbits; class-j customers represents jobs that have unsuccessfully attempted service $j \in \mathcal{J} \equiv \{1, \ldots, J-1\}$ times. Jobs leave retrial orbit O_j and attempt service according to a general process with rate r_j; this value is set by our partner organization After $J-1$ unsuccessful attempts, jobs leave the system unserved. We assume that, in order to conform to the decision-making timeline of our partner organization, the provisioning policy to determines s is stationary.

Let $\{X(t), O_j(t) | t \geq 0\}$ be a set of headcount stochastic processes corresponding to the number of busy servers and jobs in the orbit, respectively. In addition, let $\{S(t) = (s - X(t))^+ | t \geq 0\}$ denote the number of idle servers at time t. The dynamics of this system is presented in Figure 3.1.

Figure 3.1: Dynamics of the Service System



3.3.1 Capacity Minimization Problem and Offered Load Approximation

To reduce the long-term financial and environmental costs of idle capacity, while also maximizing system throughput, our partner organization would like to determine the minimum capacity s, and retrial rates $\boldsymbol{r} := (r_1, r_2, \ldots, r_{J-1})$, so that for service-level parameters $\alpha \in (0,1)$ and $\beta \in (0,1)$, a fraction β of the system capacity is busy less than a proportion α of the planning horizon $T > 0$. The service-level requirement also specifies that an access policy will take effect once the system becomes critically loaded, i.e., $X(t) > \beta s$. More specifically, our partner organization has decided to implement a policy such that once the workload from new customers and the first $j-1$ orbits enters into the critically loaded regime, on average, the number of jobs attempting service from the j^{th} orbit should be zero. Thus, when the system workload exceeds the threshold βs, jobs will be restricted from accessing the server in descending order of the number of previous service attempts; no service restrictions are imposed on the arrival of new jobs. The purpose of this policy is to schedule retrial jobs so that the critically loaded regime does not persist for too long - it reduces the potential for unexpected mechanical issues to arise - while being simple to implement (it is parametrized by $\boldsymbol{r}$) and not unreasonably rigid (e.g., an all-or-nothing policy). Finally, notice that choosing a higher threshold is similar to increasing β while choosing a lower threshold is suboptimal, i.e., it results in larger capacity levels and longer waiting times for retrial jobs.

Let $\boldsymbol{r}_j := (r_1, r_2, \ldots, r_j)$ denote the retrial rates for the first j orbits; we define the upper and lower bounds on r_j to be $\overline{r}_j(s, \boldsymbol{r}_{j-1})$ and $\underline{r}_j(s, \boldsymbol{r}_{j-1})$, respectively, for $j \in \{2, 3, \ldots, J-1\}$; similarly, $\overline{r}_1(s)$ and $\underline{r}_1(s)$ are the respective bounds on r_1. These bounds ensure that, in expectation, retrial jobs do not attempt service when $X(t) > \beta s$; their functional form depends on the specification of the arrival function. As a result, the throughput-based objective and operational constraints can be formulated as a non-linear stochastic optimization model with J decision variables:

$$\max_{s \in \mathbb{R}_{>0}, \boldsymbol{r} \in \mathbb{R}^{J-1}_{\geq 0}} \int_0^T \left(\mu \mathbb{E}[X(t)] - \gamma s\right) dt, \text{ subject to} \tag{QOPT}$$

$$\frac{1}{T} \int_0^T 1_{\{X(t) > \beta s\}} dt \leq \alpha, \tag{3.1}$$

$$\underline{r}_1(s) \leq r_1 \leq \overline{r}_1(s), \tag{3.2}$$

$$\underline{r}_j(s, \boldsymbol{r}_{j-1}) \leq r_j \leq \overline{r}_j(s, \boldsymbol{r}_{j-1}), \qquad j \geq 2, \tag{3.3}$$

where γ is the amortized cost of purchasing and maintaining one CPU core over the horizon T given that revenue is normalized to a unit value per customer. As specified by our partner organization, we fix γ small enough so that servicing clients is always profitable, i.e., no job is refused.

The objective of QOPT is to jointly determine the minimum capacity level and optimal retrial intervals in order to maximize throughput less a penalty that deters the selection of more capacity than needed in order to satisfy demand. Including such penalty term in the objective is a well-known technique that reduces a constrained problem to an unconstrained one. (Fletcher 1975, Viswanathan and Grossmann 1990, Yeniay 2005). Constraint (3.1) represents a generalised service level agreement (SLA). Although its form is induced by the SLA of the partner organization, it is not new to the operations literature. Soh and Gurvich (2017), for instance, examine a probabilistic variant of the minimization problem with service capacity as a single decision variable. Constraints (3.2)-(3.3) bound average retrial times so that jobs attempt service only when total system workload is less than βs. Notice that the bounds on r_1 are functions of the service capacity only while the bounds on r_j for $j \geq 2$ also account for the dynamics in the orbits that precede them.

Problem QOPT cannot be solved using techniques from classical queueing theory (e.g., Markov chains). This is because the service-level restrictions (3.1)-(3.3) require that $X(t)$ be known or approximated; such approximation requires that the temporal ordering among arrivals be tracked. Further, replacing the stochastic process $X(t)$ by its fluid approximation is non-trivial due to the general distribution of retrial times (Mandelbaum et al. 1998). To overcome these obstacles, we construct an offered load approximation of the system, i.e., we increase the service capacity to infinity. This removes the burden of including $O_j(t)$ in the analysis (there can be no retrials because servers are never fully utilized) and preserves information about the workload in the original system.

Let $\Pi(\cdot)$ be a standard headcount Poisson process and let $\{U(t)|t > 0\}$ be a stochastic process that represents the number of jobs in service at time t. If $U(0)$ is the number of jobs that are initially being processed, the stochastic equation governing the dynamics of the uncapacitated system is

$$U(t) = U(0) + \Pi\left(\int_0^T \lambda(t)dt\right) - \Pi\left(\int_0^T \mu U(t)dt\right).$$

Define $u(t) \in \mathbb{C}^1$ to be the fluid approximation (offered load) of $U(t)$ obtained by using either the functional strong law of large numbers or the Chapman-Kolmogorov forward equations (Massey and Pender 2018). That is, $u(t)$ is the solution to the following ordinary differential equation:

$$\dot{u}(t) = \lambda(t) - \mu u(t). \tag{3.4}$$

We also define $x(t)$ to be the fluid approximation of $X(t)$, i.e., the state descriptor of the original capacitated system with s servers. For convenience, we reformulate QOPT with respect to $x(t)$:

$$\max_{s \in \mathbb{R}_{>0}, \boldsymbol{r} \in \mathbb{R}^{J-1}_{\geq 0}} \int_0^T (\mu x(t) - \gamma s)\, dt, \ \text{ subject to} \tag{OPT}$$

$$\frac{1}{T} \int_0^T 1_{\{x(t) > \beta s\}} dt \leq \alpha, \tag{3.5}$$

$$\underline{r}_1(s) \leq r_1 \leq \overline{r}_1(s), \tag{3.6}$$

$$\underline{r}_j(s, \boldsymbol{r}_{j-1}) \leq r_j \leq \overline{r}_j(s, \boldsymbol{r}_{j-1}), \qquad j \geq 2. \tag{3.7}$$

Notice that OPT is equivalent to QOPT except that the expectation is removed in the objective; the fluid approximation describes the system dynamics of the average stochastic path.

3.3.2 Modified Offered Load Function and Total System Workload

Because our partner organization is primarily concerned with the long-term dynamics of their cloud service operation, we consider the asymptotic periodic orbit of $u(t)$, i.e., $v(t) \in \mathbb{C}^1$, and the time points when it crosses a fixed level s. If there exist instances where $v(t) > s$ for some times t, then in the capacitated system, some of the fluid is immediately served while the rest joins the orbit.

We define the amount of fluid that is immediately served upon arrival at time t by the system load function (SLF) $\bar{v}(t; s) := v(t) \wedge s$ and introduce two piecewise smooth, continuous functions $z(t; \boldsymbol{r}, s)$ and $y(t; \boldsymbol{r}, s)$. The first function, $z(t; \boldsymbol{r}, s)$, is the modified OLF representing the total system workload for any service capacity s and the corresponding feasible stationary retrial rates $\boldsymbol{r}$. It represents the sum of fluid entering the system from both external and internal (i.e., retrial orbits) sources at all times t. The second function, $y(t; \boldsymbol{r}, s)$, represents the aggregate amount of retrial fluid from all orbits that attempt to obtain service at time t. Thus, we have that

$$z(t; \boldsymbol{r}, s) = \bar{v}(t; s) + y(t; \boldsymbol{r}, s). \tag{3.8}$$

Given a capacity level s, there exist stationary retrial rates that satisfy (3.8). Our goal, then, is to select s^* and $\boldsymbol{r}^*$ so as to optimally solve OPT. In the following proposition, we derive an explicit form of the retrial fluid $y(t; \boldsymbol{r}, s)$ that, given an optimal selection of retrial rates and service capacity, maximizes system throughput by optimally redistributing the retrial fluid.

Proposition 5 (Explicit Representation of $y(t; \boldsymbol{r}, s)$)**.** *Let $\upsilon(t)$ be a continuous function. For $j \in \{1, 2, \ldots, J-1\}$, denote the modified OLF of a system with j orbits by $z_j(t; \boldsymbol{r}_j, s)$ such that*

$$z_1(t; \boldsymbol{r}_1, s) = \bar{\upsilon}(t; s) + (\upsilon(t - r_1^{-1}) - s)^+, \tag{3.9}$$

$$z_j(t; \boldsymbol{r}_j, s) = z_{j-1}(t; \boldsymbol{r}_{j-1}, s) \wedge s + (z_{j-1}(t - r_j^{-1}; \boldsymbol{r}_{j-1}, s) - s)^+ \qquad \forall j \geq 2. \tag{3.10}$$

Then, the modified OLF for the aggregate retrial fluid and the total system workload are

$$y(t; \boldsymbol{r}, s) = z_{J-1}(t; \boldsymbol{r}_{J-1}, s) - \bar{\upsilon}(t; s), \quad z(t; \boldsymbol{r}, s) = z_{J-1}(t; \boldsymbol{r}_{J-1}, s). \tag{3.11}$$

Proposition 5 describe a recursive relationship between the fluid leaving each orbit and the total system workload. In particular, at each stage, (3.9)-(3.10) decompose the workload into two flows: fluid serviced after j or fewer attempts and fluid serviced after the $(j+1)^{\text{st}}$ attempt. By definition, in a system with $J-1$ orbits, the modified OLF is given by $z(t; \boldsymbol{r}, s) = z_{J-1}(t; \boldsymbol{r}_{J-1}, s)$. Thus, if $\upsilon(t)$ admits a closed-form expression, the aggregate retrial function $y(t; \boldsymbol{r}, s)$ and the corresponding OLF for the total system workload, $z(t; \boldsymbol{r}, s)$, can also be written in closed-form. Notice that Proposition 5 only requires that $\upsilon(t)$ be a continuous function. Thus, the recursive relationship holds in settings where $\upsilon(t)$ is a smooth or piecewise smooth function (see Section 3.4.2).

3.4 Variation-Based Capacity and Retrial Rate Optimization

In this section, we leverage the recursive definition of $z(t; \boldsymbol{r}, s)$ to develop a calculus-of-variations approach to solving OPT. We first define the total variation of a function and use the concept to formulate two optimization problems that minimize the variation in $z(t; \boldsymbol{r}, s)$. The first formulation, presented in Section 3.4.1, assumes that arrivals are represented by a sinusoidal function, a common modelling choice in the extant literature (Green et al. 2007a, Yom-Tov and Mandelbaum 2014b, Furman et al. 2019). In practice, however, the arrival rate may be described by more complicated periodic functions. Thus, in Section 3.4.2, we extend our results and present a formulation that assumes arrivals can be represented by any periodic smooth or piecewise smooth arrival process defined such that the length of its smooth intervals equals the period of $\lambda(t)$. Finally, in Section 3.4.3, we discuss the implications of our approach and its connection to the original stochastic system.

Definition 1 (Total Variation of a Real-valued Function)**.** *Let $f(t)$ be a bounded, continuous, real-valued function for $t \in \mathcal{D} \subset \mathbb{R}_{\geq 0}$. If $\mathcal{E}$ is a countable set of all points t where $\dot{f}(t)$ does not exist, then the total variation of $f(t)$ is given by*

$$V(f) = \int_{\mathcal{D}\setminus\mathcal{E}} |\dot{f}(t)| dt.$$

By minimizing the variation in the modified OLF for system workload, we find parameters s^* and $\boldsymbol{r}^*$ that optimally redistributes fluid from peaks above s^* to cavities that are below it, i.e., throughput will be maximized. Note that to avoid the complexity associated with integrating $\upsilon(t)$ over incomplete periods, without loss of generality, we fix the planning horizon to $\tilde{T} \coloneqq \Omega^{-1}T\Omega$.

3.4.1 Sinusoidal Arrival Functions

To illustrate our approach using a common specification for demand that is both time-varying and periodic, we first analyze the setting where arrivals are represented by a sinusoid. That is,

$$\lambda(t) \coloneqq \bar{\lambda} + \bar{\lambda}\sigma \sin\left(\omega t + \phi\right), \tag{3.12}$$

where we define $\bar{\lambda} > 0$, $\sigma \in [0, 1]$, $\omega > 0$ and $\phi \geq 0$ to be the average arrival rate, relative amplitude, frequency, and phase parameters, respectively.

Recall that $u(t)$ is the OLF for the uncapacitated system. Because $u(t)$ is affected by the periodicity of $\lambda(t)$, it approaches the periodic orbit $\upsilon(t)$ for sufficiently large t. If there exist instances where $\upsilon(t) > s$ for some times t, then in the capacitated system, some of the fluid is immediately served while the rest joins the orbit. Thus, for the demand function given by (3.12), we first derive closed-form expressions for $u(t)$, $\upsilon(t)$, and the time points when $\upsilon(t)$ crosses a level s.

Lemma 9. *Let the dynamics of $u(t) \in \mathbb{C}^1$ be governed by (3.4) and $\lambda(t) \in \mathbb{C}^1$ be given by (3.12). If u_0 and υ_k are real numbers for $k \in \{1, 2, 3\}$, then the following three statements hold:*

(i) The solution to (3.4) admits a closed-form expression of the form

$$u(t) = u_0 e^{-\mu t} + \upsilon_1 + \upsilon_2 \cos(\omega t) + \upsilon_3 \sin(\omega t). \tag{3.13}$$

(ii) The OLF $u(t)$ approaches a periodic curve of the form

$$\upsilon(t) = \upsilon_1 + \upsilon_2 \cos(\omega t) + \upsilon_3 \sin(\omega t). \tag{3.14}$$

(iii) For $n \in \mathbb{Z}$, the equation $\upsilon(t) = s$ admits closed-form, periodic solutions

$$t(n,s) = \frac{2arctan\left(\frac{\upsilon_3 \pm \sqrt{\upsilon_3^2 - (s-\upsilon_1+\upsilon_2)(s-\upsilon_1-\upsilon_2)}}{s-\upsilon_1+\upsilon_2}\right)}{\omega} + \frac{2\pi}{\omega} n, \tag{3.15}$$

where we have explicitly noted the dependence of the time points on the capacity level s.

Lemma 9 indicates that, for sinusoidal demand, a solution to the dynamical system can be easily determined and the limiting OLF $\upsilon(t)$ can be written in closed-form. This is important if we wish to find an exact expression for $z(t; \boldsymbol{r}, s)$. In order to determine when flow from the orbits can re-enter the system, it's also crucial to characterize the time points when $\upsilon(t)$ moves from an overloaded to an underloaded system. Lemma 9 demonstrates that these periods can be determined exactly.

Define $\mathcal{Z}(n; \beta s)$ be a set of time points where $z(t; \boldsymbol{r}, s) > \beta s$ and $\upsilon(t) \leq \beta s$; these points can be calculated in closed-form based on the results of Proposition 5 and Lemma 9. Using the definition of total variation, we formulate the following optimization problem:

$$\min_{s \in \mathbb{R}_{>0}, r_j \in \mathbb{R}_{\geq 0}} V(z(t; \boldsymbol{r}, s)) \quad \text{subject to} \tag{VAR}$$

$$\frac{t_2(n, \beta s) - t_1(n, \beta s) + |\mathcal{Z}(n; \beta s)|}{\Omega} \leq \alpha, \tag{3.16}$$

$$z_1(t; \boldsymbol{r}_1, s) \leq \bar{\upsilon}(t) \qquad \forall t \in \{t | \upsilon(t) \geq \beta s\}. \tag{3.17}$$

The objective of VAR is to minimize the total variation in system workload while ensuring that the pre-specified service level is achieved and the retrial workload does not enter service in periods where the system is critically loaded. Constraint (3.16) specifies that the SLA must be satisfied; the left-hand side tracks the time $z(t; \boldsymbol{r}, s)$ spends above βs. When $\lambda(t)$ is given by (3.12), the bounds $\overline{r}_1(s)$ and $\underline{r}_1(s)$ can be simplified and replaced with (3.17) which guarantees that jobs from the first orbit re-enter service only after the system workload drops below βs. Finally, in this setting, constraint (3.7) becomes redundant and, as a consequence, is omitted from the formulation.

Lemma 10. *Let $t_2(n,s) > t_1(n,s) > t_0(n,s) > 0$ be consecutive solutions to the equation $\upsilon(t) = s$ for $n \in \mathbb{Z}$ where $\dot{\upsilon}(t_0(n,s)) < 0$, $\dot{\upsilon}(t_1(n,s)) > 0$ and $\dot{\upsilon}(t_2(n,s)) < 0$, respectively. Define the amount of excess fluid, $e_\Omega(s)$, and the total area of the cavity under level s, $d_\Omega(s)$, per period as*

$$e_\Omega(s) = \int_{t_1(n,s)}^{t_2(n,s)} \upsilon(t)dt - s(t_2(n,s) - t_1(n,s)), \qquad d_\Omega(s) = s(t_1(n,s) - t_0(n,s)) - \int_{t_0(n,s)}^{t_1(n,s)} \upsilon(t)dt.$$

Then, constraint (3.17) *implies that the following relation must hold*

$$e_\Omega(s) \leq d_\Omega(\beta s). \tag{3.18}$$

Notice that (3.17) limits when retrial fluid can re-enter the system, i.e., when the workload drops below βs. If the workload is too large, retrial fluid continuously accumulates and the orbits never empty. However, if the amount of retrial fluid that has accumulated over Ω does not exceed the area of the subsequent cavity, all of the retrial flow can be served. This logic is formalized in Lemma 10 and is particularly important when determining the optimal solution to VAR.

Proposition 6 (Constrained Variation)**.** *Let s_1 and s_2 be the capacity levels obtained by equating the left-hand and right-hand sides of* (3.16) *and* (3.18)*, respectively, and solving for s. Then,*

(i) The optimal capacity level of VAR is given by

$$s^* = \max\{s_1, s_2\}.$$

(ii) Given s^*, the optimal retrial rates are such that $r_j^* \geq 0$ for $j \in \{2, 3, \dots, J-1\}$ and

$$(t_1(n, \beta s^*) - t_0(n, s^*))^{-1} \leq r_1^* \leq (t_2(n, \beta s^*) - t_1(n, s^*))^{-1}. \tag{3.19}$$

The proposition specifies that the optimal capacity level and retrial rates are selected such that throughput in the original capacitated system matches throughput in the system where fluid arriving when $x(t) = s^*$ is redistributed, i.e., $x(t) = z(t; \boldsymbol{r}^*, s^*) \wedge s^*$. In general, smaller capacity values result in smaller total variation values. As capacity decreases, the left-hand side of constraint (3.16) increases until it reaches its upper bound given by α (the right-hand side of this constraint). Further, recall that $y(t; \boldsymbol{r}, s)$ represents the aggregate amount of retrial fluid from all orbits that attempts to re-enter the system at time t. Constraint (3.18) specifies that over the entire time horizon, this retrial fluid must fit into the corresponding area under βs that is available to receive it. Higher values of capacity reduce the amount of excess fluid and increase the size of the cavity that can be used to serve retrial fluid. As a result, even though the objective function is non-linear, we have that s^* equates the left-hand and right-hand sides of (3.16) or (3.18). We also find that all retrial rates satisfying (3.19) are optimal solutions to OPT. Consequently, without loss of generality, we select r_1^* so that jobs re-enter service as soon as possible and fix $r_j^* \geq 0$ for $j \in \{2, 3, \dots, J-1\}$.

We now demonstrate that minimizing the total variation in system workload is akin to maximizing the penalized throughput objective in OPT. We note that VAR is a much simpler optimization problem. Directly solving OPT is intractable; it is a non-linear optimization problem where the objective function and constraints do not have closed-form expressions.

Proposition 7 (Equivalency)**.** *The optimal solution to VAR is the optimal solution to OPT.*

3.4.2 General Periodic Arrival Functions

In the previous section, we formulated a general calculus-of-variations approach to solve OPT for settings where customer arrivals $\lambda(t)$ can be described by a sine function, as in (3.12). Such arrival function has properties that make VAR analytically tractable, i.e., point symmetry and exactly one maximum and minimum per period Ω. As a result, if the SLA is satisfied, $z(t; \boldsymbol{r}, s)$ does not exceed

βs level when $\upsilon(t) \leq \beta s$. This implies that at most one orbit is required to serve all jobs and the time $z(t; \boldsymbol{r}, s)$ spends above βs level can be tracked given the closed-form expression of $\upsilon(t)$.

The results extend to more complicated periodic, smooth functions such as higher order polynomials or a combination of sine functions. Although these arrival processes can fit many patterns, they do not necessarily guarantee that the fitted $\lambda(t)$ agrees with the observed data. This is because polynomial functions are not periodic and one may not be able to estimate the frequency values of sinusoids with sufficient precision (Chen et al. 2018). However, observed arrivals with a specified period $\Omega \geq 0$ may be modelled using a piecewise smooth, continuous, periodic function of the form

$$\lambda_\Omega(t) \coloneqq \lambda(t \mod \Omega) \tag{3.20}$$

defined such that $\lambda(t)$ has a polynomial representation. Although other function classes may suffice, polynomials are the simplest analytical form that allows for direct evaluation. Thus, in this section, we extend our main results to accommodate this broad class of arrival functions. Specifically, we show that solving OPT optimally using the variation approach generalizes to cases where periodicity and the piecewise smoothness of arrivals are essential features (see Section 3.5).

Contrary to (3.12), however, $\lambda_\Omega(t)$ may not have point symmetry and there may be several local extrema over Ω. Thus, in the following lemma, we generalize the results corresponding to the OLF in the uncapacitated system and its periodic orbit.

Lemma 11. *Let $\lambda_\Omega(t)$ be defined as in (3.20). Then, for a real number u_0 and a polynomial function $\upsilon(t)$, the following two statements hold.*

(i) The solution to $\dot{u}_\Omega(t) = \lambda_\Omega(t) - \mu u_\Omega(t)$ admits a closed-form expression of the form

$$u_\Omega(t) = u_0 e^{-\mu t} + \upsilon(t \mod \Omega). \tag{3.21}$$

(ii) The OLF $u_\Omega(t)$ approaches a periodic, piecewise smooth function $\upsilon_\Omega(t) \coloneqq \upsilon(t \mod \Omega)$.

Lemma 11 extends Lemma 9 and indicates that for any arrival function that takes the form of (3.20), $\upsilon_\Omega(t)$ is piecewise smooth, continuous, and admits a closed-form expression. In contrast to Lemma 9, however, we do not derive explicit expressions for the time points when $\upsilon_\Omega(t)$ crosses the capacity level s. Since this involves solving for the roots of a higher order polynomial - which, in general, does not have closed-form expressions - one must determine them numerically.

For the nth period of $\upsilon_\Omega(t)$, let $M_\Omega \in \mathbb{Z}_{>0}$ be the number of times $\upsilon_\Omega(t)$ exceeds the capacity level s and define $\{\mathcal{A}_m(n,s)\}_{m=1}^{M_\Omega}$ and $\{\mathcal{B}_m(n,s)\}_{m=1}^{M_\Omega}$ to be consecutive, non-overlapping time intervals corresponding to when $\upsilon_\Omega(t) \geq s$ and $\upsilon_\Omega(t) < s$, respectively. To redistribute the retrial fluid over the intervals where $\upsilon_\Omega(t) < \beta s$, we follow the same recursive procedure as in Section 3.4.1. Because $\upsilon_\Omega(t)$ admits a closed-form expression, existence of a suitable modified OLF $z(t;\boldsymbol{r},s)$ is guaranteed by Proposition 5. Thus, defining $\mathcal{Z}_m(n;\beta s)$ to be the set of time points where $z(t;\boldsymbol{r},s) > \beta s$ and $\upsilon_\Omega(t) \leq \beta s$ for all m and $n \in \mathbb{N}$, we generalize VAR by presenting the following optimization problem.

$$\min_{s\in\mathbb{Z}_{>0}, r_j\in\mathbb{R}_{\geq 0}} V(z(t;\boldsymbol{r},s)) \ \text{ subject to} \tag{GVAR}$$

$$\frac{\sum_{m=1}^{M}\left(|\mathcal{A}_m(n,\beta s)| + |\mathcal{Z}_m(n;\beta s)|\right)}{\Omega} \leq \alpha, \tag{3.22}$$

$$z_1(t;\boldsymbol{r}_1,s) \leq \bar{\upsilon}_\Omega(t) \qquad \forall t \in \{t|u_\Omega(t) \geq \beta s\}, \tag{3.23}$$

$$z_j(t;\boldsymbol{r}_j,s) \leq z_{j-1}(t;\boldsymbol{r}_{j-1},s) \quad \forall t \in \{t|z_{j-1}(t;\boldsymbol{r}_{j-1},s) \geq \beta s\}, j \geq 2, \tag{3.24}$$

$$z(t;\boldsymbol{r},s) - s \leq 0 \qquad \forall t. \tag{3.25}$$

The objective of GVAR is identical to VAR; constraint (3.22) generalizes the SLA constraint given by (3.16), i.e., its left-hand side tracks the time $z(t;\boldsymbol{r},s)$ spends above βs, and (3.23) generalizes (3.17) by ensuring that jobs from the first orbit re-enter service only after the system workload drops below βs. Constraints (3.6)-(3.7) reduce to the relation between $z_j(t;\boldsymbol{r}_j,s)$ and $z_{j-1}(t;\boldsymbol{r}_{j-1},s)$ as given by (3.24); these constraints ensure that the retrial workload from orbits $j \geq 2$ do not attempt service when the system is critically loaded. Finally, because $\upsilon_\Omega(t)$ may have multiple peaks above the capacity level βs, contrary to the simpler sinusoidal case, minimizing the total variation in system workload involves balancing the distribution of fluid over multiple cavities. Since it is possible that some demand may not be processed, we include (3.25) which rectifies the issue.

Corollary 2 (Constrained Variation)**.** *If the fluid that accumulates per time interval $\mathcal{A}_m(n,s)$ is less than the total area of the cavity of $\mathcal{B}_m(n,s)$ for each n and m such that $|\mathcal{Z}_m(n;\beta s)| = 0$, then*

(i) The capacity level s^* that optimally solves GVAR is the solution to (3.22) *at equality.*

(ii) Given s^*, the optimal retrial rates are such that $r_j^* \geq 0$ for $j \in \{2, 3, \ldots, J-1\}$ and

$$\frac{1}{r_1^*} \geq \max\{|\mathcal{A}_m(n, \beta s^*) \cap \mathcal{B}_m(n, s^*)| + |\mathcal{A}_m(n, s^*)|\}_{m=1}^M, \tag{3.26}$$

$$\frac{1}{r_1^*} \leq \min\{|\mathcal{A}_m(n, \beta s^*) \cap \mathcal{B}_m(n, s^*)| + |\mathcal{B}_m(n, \beta s^*)|\}_{m=1}^M. \tag{3.27}$$

In general, solving GVAR is challenging. However, under the assumption that the width, depth, and area of each cavity is larger than the amount of retrial fluid entering the system, (3.22) holds with equality. This assumption, which is expressed mathematically in Corollary 2, ensures that the optimal capacity level and retrial rates are determined by equating throughput in the original capacitated system with throughput in the system where fluid arriving when $x(t) = s^*$ is redistributed, i.e., $x(t) = z(t; \boldsymbol{r}^*, s^*) \wedge s^*$. Then, notice that given s^*, any stationary retrial rate r_1 that satisfies (3.26) and (3.27) is optimal. Without loss of generality, we select r_1^* so that jobs re-enter service as soon as possible and fix $r_j^* \geq 0$ for $j \in \{2, 3, \ldots, J-1\}$.

Corollary 3 (Equivalency)**.** *The optimal solution to GVAR is the optimal solution to OPT.*

Corollary 3 generalizes Proposition 7 and demonstrates that our approach of minimizing the variation in system workload can be applied to a broad class of arrival functions. This is particularly important as, although generalised sine functions are a common choice when modeling time-varying, periodic arrival functions, real-world circumstances (e.g., Section 3.5) may be far more complex.

3.4.3 Model Discussion

Our calculus-of-variations approach introduces a new technique for determining the optimal stationary capacity and set of retrial intervals for a system with time-varying demand, a finite number of retrial attempts, and pre-specified service levels. It applies to any setting where a decision maker wishes to determine the minimum capacity level that maximizes system throughput while adhering to various company requirements on how service should be administered. As demonstrated in Section 3.4.2, our model accommodates a large class of periodic patterns of time-varying demand and generally distributed retrial times. Whereas directly optimizing OPT is intractable, minimizing the total variation yields optimal solutions that can be derived realitvely easily.

Solving GVAR may lead to its own set of challenges. If the width, depth, and area of each cavity exceeds the amount of retrial fluid waiting to re-enter the system, any retrial rate satisfying (3.26) and (3.27) can return such fluid into the system; a corresponding s^* can be obtained by solving a single non-linear equation. However, if there exists cavities that cannot fit the accumulated retrial fluid up to that point, a more exhaustive search is required. Fortunately, as the SLA becomes stricter, more capacity will be provisioned. By increasing capacity levels, the total amount of retrial fluid will decrease while the area in the cavities of the corresponding long-term OLF will increase. Thus, it should be easier to solve GVAR for SLAs that encourage high-quality service.

Our analysis suggests that in order to minimize the total variation in workload, only one orbit is required. This is because the optimal retrial rate r_1 is bounded for any service capacity s and any value within these bounds serves the entire retrial workload. However, because we are approximating the original stochastic system by a fluid, the rest of the orbits are useful in reducing abandonment from the system in the stochastic regime. As previously discussed, the fluid approximation describes the system dynamics of the average stochastic path. Thus, our approach captures well the average service experience, but there can be considerable variability in the sample paths of customers in the original stochastic system. Thus, a decision maker should exercise caution when setting the retrial rates. Changing r_j for $j = 2, \ldots, J-1$ may impact the performance of the original service system while remaining undetected by the fluid model.

3.5 Case Study

In this section, we apply our calculus-of-variations approach to determine the optimal cloud capacity level and set of retrial rates for a single service within our partner organization. We estimate the operational costs associated with this policy and evaluate its sensitivity with respect to changes in the SLA. Our analysis suggests that substantial savings may be possible; further savings can be achieved by slightly modifying the company's current SLA without undermining service quality.

We use three performance measures to compare optimal policies under different service level agreements: (i) operational costs; (ii) the probability of waiting before accessing service; and (iii) the expected waiting time. As advised by our partner organization, operational costs are associated with housing, maintaining, powering, and cooling one CPU core. Since this is equal to \$0.03 per hour, the total annual operational costs $C(s)$ of capacity can be calculated as

$$C(s) = 262.8s.$$

To estimate the latter two performance metrics, i.e., the probability that a computing request waits before accessing service and the expected waiting time, we introduce fluid-based estimates of the long-term probability of entering service after at least $j \in \{1, 2, \dots, J\}$ attempts. Because the fluid model tracks the average path of the original stochastic system, we can estimate the probability that a request accesses service after at least two attempts by computing the proportion of time $\upsilon_\Omega(t)$ is above level s divided by the total amount of fluid in the system. The probability that a request accesses service after at least $j > 2$ attempts can be computed in a similar fashion, i.e., it is the proportion of time $z_{j-2}(t; \boldsymbol{r}_{j-2}, s)$ is above level s divided by the total amount of fluid in the system. The logic is formally expressed in the following definition.

Definition 2 (Fluid-based Probability)**.** *Let $\Xi \in \{1, 2, \dots, J\}$ be a discrete random variable representing the long-term average number of attempts before a job is serviced or lost. Define let $\mathrm{e}_\Omega(f, s)$ to be the amount of fluid above level s for an OLF $f(t)$ during a single period Ω. Then,*

(i) The long-run probability of entering service after at least $j \in \{1, 2, \dots, J\}$ attempts is

$$\hat{\mathbb{P}}(\Xi \geq 1) = 1, \tag{3.28}$$

$$\hat{\mathbb{P}}(\Xi \geq 2) = \frac{e_\Omega(\upsilon_\Omega, s)}{e_\Omega(\upsilon_\Omega)}, \tag{3.29}$$

$$\hat{\mathbb{P}}(\Xi \geq j) = \frac{e_\Omega(z_{j-2}, s)}{e_\Omega(\upsilon_\Omega)} \qquad j > 2. \tag{3.30}$$

(ii) The fluid estimate of the probability mass function of Ξ is

$$\hat{\mathbb{P}}(\Xi = j) = \hat{\mathbb{P}}(\Xi \geq j) - \hat{\mathbb{P}}(\Xi \geq j + 1) \qquad \forall j. \tag{3.31}$$

Let $W \geq 0$ be a random variable that represents the waiting time before accessing service. According to Definition 2, $\hat{\mathbb{P}}(\Xi \geq 2) = \hat{\mathbb{P}}(W > 0)$ and the fluid estimate of the long-term probability of waiting in the original stochastic system can be computed. Notice that jobs are serviced after at most J attempts or lost, i.e., leave the system without service. Thus, we have that $\Xi \leq J$. Finally, notice that a job serviced after exactly $j > 1$ attempts spends $\frac{1}{r_1} + \frac{1}{r_2}, \ldots, \frac{1}{r_{j-1}}$ time waiting in the orbits on average. As a result, we can write the estimated expected waiting time in accordance with the definition of the expectation of a discrete random variable. That is, for retrial rates $r_j \neq 0$, the fluid estimate of the expectation of W, or the expected waiting time before a job enters service, is

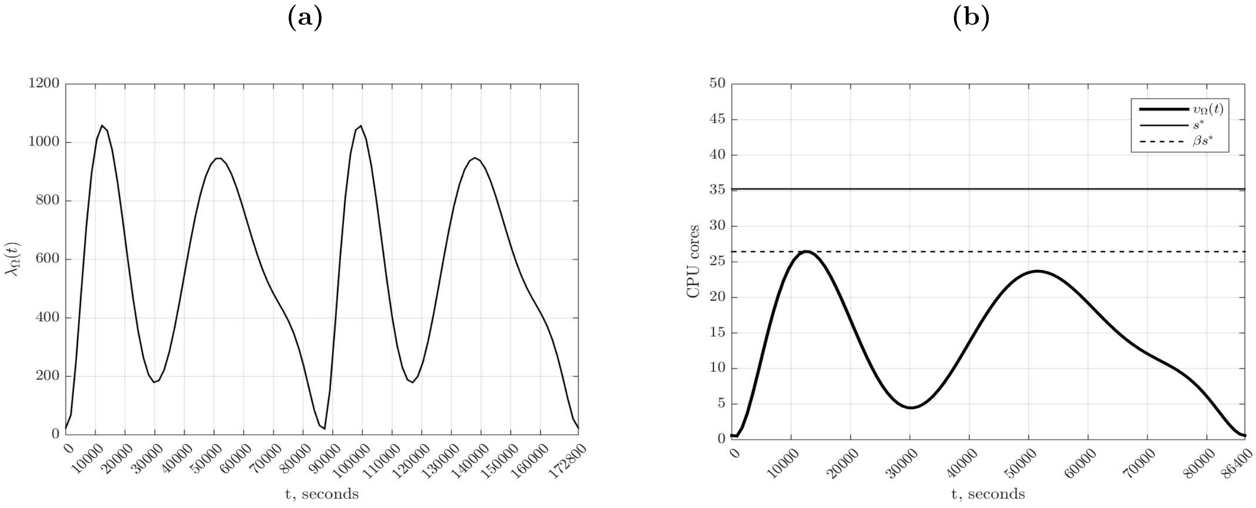
$$\hat{\mathbb{E}}[W] = \sum_{j=2}^{J} \sum_{k=2}^{j} \frac{1}{r_{k-1}} \hat{\mathbb{P}}(\Xi = j).$$

Our partner organization is composed of many different services. Each service has a specific business focus and is allocated dedicated cloud computing capacity. Although each service may differ in the functional form of the arrival rate and the service requirements of cloud computing requests, they have similar features, i.e., demand is time-varying and jobs are homogeneous. A typical service is characterised by a large number of arrivals with short service times. Our data set contains per second counts of arriving jobs over a typical 24-hour period for one such service. It accounts for approximately 50 million arrivals, and as indicated by our partner organization, the average service time equals 25 milliseconds. In accordance with our partner organization's practices, we assume that the arriving pattern repeats daily and fix $J = 3$ so that requests not serviced after 3 attempts are considered lost. Further, our partner organization implements an SLA with $\alpha = 0.01$ and $\beta = 0.75$, i.e., more than 75% of the system capacity should be busy less than 1% of each day.

We convert the data into instantaneous arrival counts and fit the resulting sample using a curve fitting toolbox in Matlab R2015a. Figure 3.2a represents the fitted arrival process over two periods Ω, i.e., 48 hours or 172,800 seconds. Without loss of generality, we assume $t = 0$ corresponds to 2am. According to this figure, the system experiences a high volume of arrivals from 5am to 6am and from 4pm to 5pm; the intensity of arrivals is lowest at approximately 2am and 10:30am. Thus, $\lambda_\Omega(t)$ is a piecewise smooth, periodic function that captures large fluctuations in the workload over each day of operation. Because the length of the smooth intervals of $\lambda_\Omega(t)$ equals its period, the arrival function satisfies the requirements in Section 3.4.2 and $\lambda(t)$ is represented by an eighth degree polynomial function (see Appendix B for more details).

As per Lemma 11, we derive a closed-form expression of $\upsilon(t)$ that can also be represented by

Figure 3.2: Fluid estimates: (a) $\lambda_\Omega(t)$ with $\tilde{T} = 48$ hours; and (b) the optimal solution with $\tilde{T} = 24$ hours.

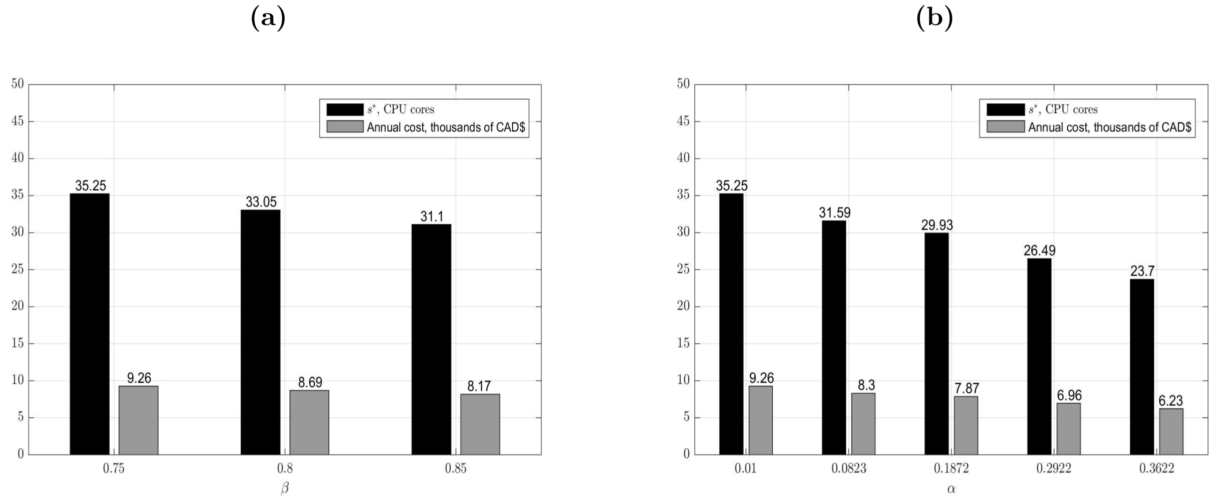


an eighth degree polynomial function (see B for more details). By Corollary 2, we solve GVAR and present its solution in Figure 3.2b (the solid line). The analysis indicates that our partner organization needs 35.25 CPU cores to meet their service requirements (we refer to this service capacity as the reference level and note that managers can round s^* up or down in accordance with the recommendations of their hardware engineers). Notice that, due to the SLA, the minimum capacity that satisfies (3.22) is greater than the maximum of $\upsilon_\Omega(t)$. As a result, $\hat{\mathbb{P}}(W > 0) = 0$ and $\hat{\mathbb{E}}[W] = 0$. Further, in the fluid setting, the reference level is sufficient to serve all jobs and selecting $\boldsymbol{r}$ has no effect on system performance. The optimal capacity s^* produced by our model is approximately 10% less than the partner organization's current allocation for this service. Because our partner organization allocates cloud computing capacity to thousands of services, a 10% reduction in capacity translates into a company-wide savings of tens of millions of dollars annually.

In Figures 3.3 and 3.4, we perform a sensitivity analysis with respect to the SLA. Each group of bars represents a capacity level and its annual costs, respectively. For Figures 3.3b, 3.4a, and 3.4b, we select the range of α such that each figure includes a set of representative scenarios (from left to right): (i) the optimal capacity levels are less than the reference level but have $\hat{\mathbb{E}}[W] = 0$; (ii) the minimum capacity level where $\hat{\mathbb{E}}[W] = 0$; and (iii) capacity levels where $\hat{\mathbb{E}}[W] > 0$.

In Figure 3.3a, we fix $\alpha = 0.01$ and vary β from 0.75 (first group of bars), 0.8 (second group of

Figure 3.3: Sensitivity analysis of the SLA: (a) We fix $\alpha = 0.01$ and vary β; and (b) We fix $\beta = 0.75$ and vary α.

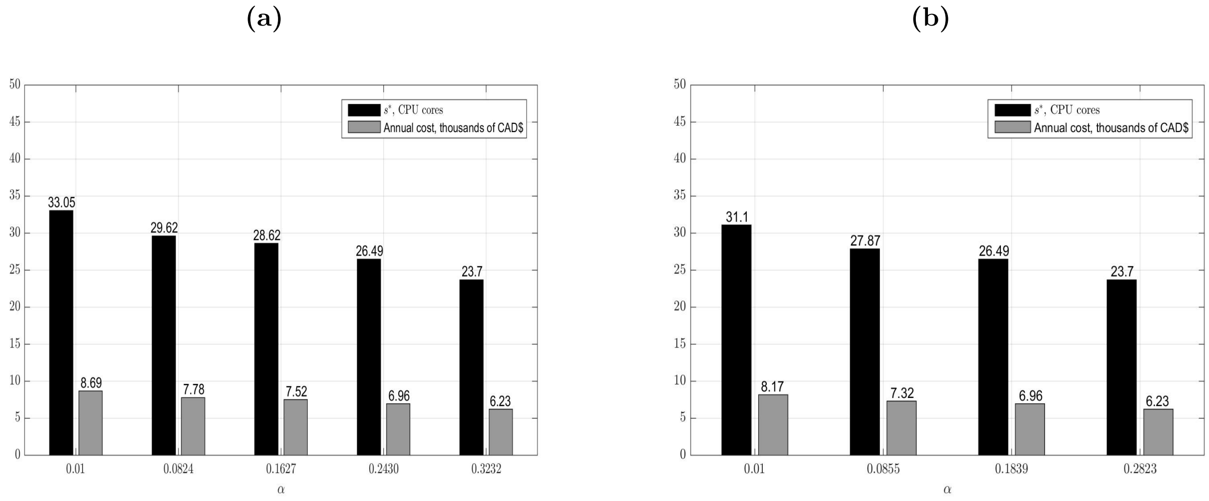

(a)

(b)

bars), and 0.85 (third group of bars). We find that increasing β to 0.8 results in an additional 6.25% increases annual savings when compared to the reference level with no change in $\hat{\mathbb{E}}[W]$. The SLA with $\beta = 0.85$ implies savings of 11.8% as compared to the reference level; again, there is no change in $\hat{\mathbb{E}}[W]$. In Figure 3.3b, we fix β and vary α noting that our partner organization has advised us $\beta = 0.75$ is an engineering standard. We find that on average, increasing α by 0.01 reduces the capacity level by approximately 0.33 servers. This does not change the expected waiting time or the probability of waiting until $\alpha = 0.2922$ where 26.49 CPU cores are provisioned. Beyond this point, the expected waiting is positive and it is imperative $\boldsymbol{r}$ be selected optimally to obtain the minimum waiting for that capacity. Nevertheless, in this regime, we observe that the expected waiting time increases quickly. For $\alpha = 0.3622$, three fewer servers are purchased as compared to $\alpha = 0.2922$. However, the fluid estimate of the probability of waiting increases to 14.5% while the expected waiting time is 24.4 minutes (see Table B.1 in Appendix B for more details).

In Figures 3.4a-3.4b, we fix $\beta = 0.8$ and $\beta = 0.85$, respectively, and vary α. Although these values of β are larger than the engineering standard, we find that we obtain lower values of α for the same capacity level as compared with $\beta = 0.75$. As before, the expected waiting time and probability of waiting does not change until the number of CPU cores exceeds 26.49. Beyond this value, the expected waiting is positive (see Tables B.2-B.3 in Appendix B for more details). Nevertheless, when $s < 26.49$, we observe that increasing β results in shorter expected waiting times because jobs are allowed to re-enter service faster. Because the expected waiting time of 25 minutes is considered long by the partner organization, we do not review capacity levels that result in a longer wait.

Figure 3.4: Alternative SLAs by varying α fixing (a) $\beta = 0.8$; and (b) $\beta = 0.85$.



(a)

(b)

Our analysis indicates that there are many SLA parameters that produce an expected waiting time of zero for arriving jobs. Thus, a 10% reduction in capacity, as suggested by our initial analysis using our partner organizations current SLA, may be a conservative estimate. In particular, the results from Figures 3.3 and 3.4 suggest that our partner organization has two operational levers with which to generate additional savings. By modifying their existing SLA using a combination of α and β, capacity levels can be further reduced while still providing very high levels of service quality. The reduction in capacity means that less electricity will be consumed and this addresses the sustainability issues that have become an increasing concern in the cloud computing industry.

3.6 Summary of Contributions

Our analysis suggests that in settings where the arrival rate can be characterized by a sinusoid, jobs will be serviced after at most two attempts. That is, the remaining retrial orbits do not impact the fluid dynamics but perform a safeguarding role in the stochastic regime against abandonment due to repeated service denials. Thus, managers that observe sinusoidal demand should be focused on selecting the retrial rate parameter for the first orbit. When the demand function is more complex, as in our case study, we show that our approach is flexible in that it can accommodate complicated arrival functions. Nevertheless, this complexity is not inherited by the corresponding optimization problem as the optimal solution can be efficiently obtained. We confirm the utility of our methodology by showing that capacity at our partner organization can be reduced by approximately 10%. We then introduce easy-to-compute performance measures that link the system performance of the fluid model to the stochastic regime. We use the fluid-based performance measures to illustrate that the proposed down-scaling of CPU capacity does not substantially affect service quality for a large range of SLAs, i.e., the probability of waiting and the expected waiting time are zero. Thus, additional savings are achievable with virtually no discernible service impact.

Although further investigation is required, our study suggests that cloud computing providers may actually operate closer to optimality than has been previously thought. That is, the large amounts of idle capacity (see Columbus 2016, FLEXERA 2019, for instance) may be a byproduct of satisfying challenging service level agreements rather than a systemic undervaluation of idle resources. To this end, our results encourage managers to carefully review their SLAs as small changes may result in a substantial reduction in costs. Our partner organization shares these conclusions. They agree that reviewing their SLA is an important step in making their operations more efficient and sustainable. Nevertheless, our analysis does confirm that cloud computing providers determine capacity levels by ensuring that most of the workload during peak demand periods can be immediately serviced. This explains reports (e.g., Chapel 2019) claiming that, for the most part, cloud capacity around the globe is usually idle. It follows that, in order to satisfy strict SLAs and accommodate access policies that encourage retrial jobs to enter service during off-peak periods while preserving short waiting times, providers must provision large amounts of capacity.

Chapter 4

Prediction of PPE in Hospitals During COVID-19

The work in this chapter has been submitted for publication as the following:

Eugene Furman, Alex Cressman (medical practitioner), Saeha Shin (data analyst from a partner organization), Alexey Kuznetsov (supervisory committee member), Fahad Razak (medical practitioner), Amol Verma (medical practitioner) and Adam Diamant (advisor) "Prediction of PPE in Hospitals During COVID-19", in *Health Care Management Science*, [2020]

4.1 Introduction

Personal protective equipment (PPE) includes items such as surgical masks, face shields, gloves, eye protection, and gowns (Canadian Centre for Ocupational Health and Safety 2018). They are designed to protect the wearer, and individuals they come in contact with, from potential exposure to infectious diseases (FDA 2020). Although PPE is typically used in clinical settings, it has become an essential commodity following the recent outbreak of Coronavirus Disease (COVID-19). That is, to combat the spread of the virus, many governments are mandating the use of PPE in public spaces such as retail stores, restaurants, community centers, and on public transit (e.g., Ontario Health 2020a). Wearing a mask to conduct activities outside the home is now recommended by the World Health Organization (WHO 2020), the Centers for Disease Control and Prevention (CDC 2020), and the Government of Canada (Canada.gov 2020). This non-pharmaceutical intervention is designed to slow the spread of COVID-19, however, it has also resulted in large surges in demand for PPE and, correspondingly, critical supply shortages (Gondi et al. 2020). This has had a detrimental effect on the ability of hospitals to source PPE (Livingston et al. 2020) and outfit their staff (Ranney et al. 2020). In some cases, the inability to provide adequate PPE to frontline health care workers has led to higher rates of infection and death amongst patients (Balmer and Pollina 2020).

In hospitals, PPE has traditionally been used to protect healthcare workers when performing various types of medical procedures (Akduman et al. 1999, Benson et al. 2013). During the pandemic, however, PPE has become a requirement for all patient-practitioner interactions; any time a health worker enters a patient's room or physically interacts with a patient, they are required to wear PPE. As a result, although patient volumes initially decreased with the onset of the pandemic as many non-emergent procedures were postponed, there has been a large increase in the use of PPE to manage urgent and non-elective patient care (Daly 2020). For instance, in response to the COVID-19 pandemic, the Canadian government has ordered approximately 395 and 154 million surgical and N95 masks, respectively, to distribute directly to hospitals (Public Services and Procurement Canada 2020). As acute care facilities resume normal operations (e.g., diagnostic testing, elective surgery, ambulatory care), all staff, employees, and visitors will likely be required to wear PPE at all times (UHN 2020) while additional PPE requirements will be mandated during medical procedures (Ontario Health 2020b). This will put even more pressure on PPE supply chains which, in some health care systems, face estimated delays of up to 6 months and have had major distributors unable to fill orders (Mehrotra et al. 2020). Since one of the biggest obstacles to restarting normal hospital operations is the consistent and timely supply of PPE, these statistics are particularly troubling (Daly 2020).

Given the importance of PPE in acute care centers, proactive PPE management has become an essential component in hospital operations (Crawley 2020). Successful administration of PPE inventory is directly linked to accurately predicting the demand for medical services, and in particular, the number and nature of all patient-practitioner interactions (see Barrett et al. 2020, for instance). Doing so is challenging due to the large number of diagnoses, clinical procedures, and surgical interventions as well as the time-dependent nature of patient arrivals (e.g., Yom-Tov and Mandelbaum 2014b). While various simulation studies have been used to estimate hospital workload during the pandemic (Calafiore et al. 2020, Toda 2020, Wangping et al. 2020), they are hard to replicate, time-consuming to build, difficult to use effectively, and are not conducive to performing a comparative analysis that is required for prescriptive managerial decision-making.

In this work, we develop a time-varying queueing model to predict the amount of PPE required in a clinical inpatient setting over a specified time horizon. As has been well-established in the literature (e.g., Whitt and Zhang 2017, Yom-Tov and Mandelbaum 2014a), we assume that the process governing when patients arrive to the hospital is time-dependent. We then cluster patients with similar hospital experiences (e.g., diagnosis, expected treatment plan) into classes and estimate their length-of-stay (LoS) in the hospital as well as the PPE requirements for each interaction with a practitioner. We show that these dynamics can be modeled using multiple independent $M_t/G/\infty$ queues (see Massey and Whitt 1993, for instance), one for each patient class, and as a consequence, derive closed-form estimates for the expected amount of PPE required during the time horizon.

Using a large data set of clinical, demographic, and operational attributes from 22,039 patients admitted to the general internal medicine (GIM) service at St. Michael's Hospital (a primary care facility in Toronto, Canada) from April 2010 to November 2019, we demonstrate the practicality of our approach. We first validate the assumption that time-varying demand is an appropriate modelling choice. We then describe how to group patients into classes depending on the nature of their medical interactions as well as their LoS values. Note that this is an important step to ensure that patients in the same class have similar hospital experiences. Next, we use our model to predict the yearly PPE requirements of the GIM service at St. Michael's Hospital when it returns to normal operations excluding those patients who are diagnosed with COVID-19. Using the current regulations governing PPE use at the hospital and leveraging pre-pandemic patient volumes, we show that the GIM service will need approximately 225,000 gloves, 11,500 gowns, 181,500 surgical masks, 7500 N95 masks, and 4000 face shields. Thus, gloves and surgical masks represent approximately 90% of the predicted PPE usage. We also find that while demand for gloves is driven entirely by patient-practitioner interactions, 86% of the predicted demand for surgical masks can be attributed to the requirement that medical practitioners will need to wear masks when not interacting with patients. In addition, we show that our approach provides upper and lower bounds for the amount of PPE predicted to be used. We also perform an analysis to determine the sensitivity of the predictions to the number of patient classes chosen by the modeller.

We contribute to the operations research and medical literature by applying a queueing theoretical framework to a high-impact medical problem. To the best of our knowledge, our work is the first to obtain closed-form expressions for PPE usage in a hospital setting. Our method is analytical, computationally efficient, and does not require that a hospital develop an extensive simulation study. By deriving closed-form expressions, the sensitivity of the predictions to changes in the model's parameters can be evaluated. This helps hospital administrators gain practical insight into the dynamics of PPE usage which is especially valuable for the effective management of a scarce resource in a rapidly changing environment. Finally, we note that our approach is easily scalable; it can be used to make predictions for a single department, an entire hospital, or be deployed at the regional or provincial level.

4.2 Literature Review and Contribution

To predict PPE consumption, we introduce a stochastic queueing framework with multiple independent $M_t/G/\infty$ queues to model the dynamics of distinct patient classes that are admitted to the hospital, receive clinical care, and interact with practitioners. Pioneering theoretical work in the study of $M_t/G/\infty$ systems date back to Palm (1943) and Khintchine (1955) who show that the number of jobs in the system at any time instant follows a Poisson process with a time-varying rate. Since then, the extant literature has shown that departures from such queues also follows a non-homogeneous Poisson process (see Brown 1969, Foley 1982, 1986). More recent work derives the expected number of jobs remaining in the system after each departure, i.e., the number of busy servers, for specific service distributions Eick et al. (1993a,b). Further, several studies derive the steady-state distribution and fluid limit of systems with a periodic arrival rate Dong and Whitt (2015a,b), Whitt (2015). For a review of queueing systems with non-stationary demand, see the survey papers by Defraeye and Van Nieuwenhuyse (2016) and Whitt (2018).

From a practical perspective, the number of applications that use $M_t/G/\infty$ queues to model service systems is vast: they have been employed, for instance, to evaluate the adequacy of storage systems (Crawford 1977), determine the readiness of military equipment (Hillestad and Carrillo 1980, Crawford 1981), and model the arrival of customers to in-bound call centers (Khudyakov et al. 2010, Vizarreta et al. 2018). Specific to healthcare, several studies have used the model to analyse practitioner staffing and capacity management problems (Yom-Tov and Mandelbaum 2014a, Pender 2016, Furman et al. 2019, Razak et al. 2020). Due to the assumption of infinite capacity, $M_t/G/\infty$ queues are particularly useful in situations where service delay is near zero Green and Kolesar (1998). The principle of zero waiting time is common in the estimation of total workload for staffing analyses and is also known as the offered load approximation (Feldman et al. 2008a, Janssen et al. 2011, Liu 2018, Furman and Diamant 2020). For instance, de Véricourt and Jennings (2011) model a medical unit as a closed queueing network and determine optimal nurse-to-patient ratios. There are also several papers that analyze the supply of hospital beds and derive expressions to promote better management strategies in settings with time-varying demand Green and Nguyen (2001), Green et al. (2007b), Zeltyn et al. (2011), Zychlinski et al. (2018). We contribute to this literature by using multiple $M_t/G/\infty$ queues to derive closed-form expressions to predict PPE consumption from an offered load estimate of hospital workload.

Our work is related to the literature that develops best-practices for supply chain disruptions. Tang (2006), Stecke and Kumar (2009) and Carbonara and Pellegrino (2018) provide insight into how a supply chain can respond to natural disasters, terrorist attacks, and other unforeseeable emergencies. Logistics networks can be built with redundant transportation routes (Dash et al. 2013), suppliers are encouraged to invest in more robust infrastructure (Dolinskaya et al. 2018), and inventory postponement can be employed to better understand the changing demand-supply relationship (Chiou et al. 2002, Yeung et al. 2007, Choi et al. 2012). Nevertheless, especially in demand-driven supply chains, these approaches are not always useful in situations with extreme demand volatility unrelated to infrastructure damage or logistical disturbances (Chan and Chung 2004, Chen and Xiao 2009, Verdouw et al. 2011). Instead, effective inventory management and accurate demand predictions are crucial (Chen et al. 2001, Milner and Rosenblatt 2002, Qi et al. 2004, Xu et al. 2003). We add to this literature by proposing an analytical demand prediction tool for PPE usage that can be employed in settings with supply chain disruptions where consumption is a function of the length of a customers interaction with an organization.

Specific to research on COVID-19, our analysis is related to studies that predict future demand for medical services; see the surveys by Sahin et al. (2020), Workman et al. (2020) and Harapan et al. (2020). Since the onset of the pandemic, this literature has grown substantially. Some studies employ deterministic compartmental modifications of Susceptible-Infected-Recovered (SIR) models which are parameterized by empirical studies (Tuite et al. 2020, Calafiore et al. 2020, Biswas et al. 2020). Such methods result in systems of differential equations that must be solved numerically to obtain predictions or insight related to possible public health initiatives (Liu et al. 2020). Other studies combine dynamic SIR models with Bayesian inference techniques (see Chen and Qiu 2020, for example) or propose stochastic Markov models to predict the spread of the disease (see Zhang et al. 2020, for example); solutions are obtained by performing a simulation analysis. Stochastic implementations of SIR models are also common in the literature (Bardina et al. 2020, Karako et al. 2020, Simha et al. 2020). We provide an approach that can be used alongside these models. In particular, given a PPE policy and a (potentially) time-varying demand curve for hospital services using one of the above methods, our model derives a closed-form expression for PPE usage and can be employed during a COVID-19 outbreak or after regular operations have resumed.

Finally, our work contributes to the literature on critical shortages of PPE during the COVID-19 pandemic. While many studies leverage COVID-19 transmission models to evaluate the effectiveness of non-pharmaceutical containment strategies (e.g., Evgeniou et al. 2020, Flaxman et al. 2020, Zhang and Enns 2020), the literature predicting demand for PPE is scarce. Some authors propose qualitative techniques to manage PPE in a medical setting (Rowan and Laffey 2020, Ranney et al. 2020). These strategies are consistent with practices that are used when there are demand and/or supply disruptions in the pharmaceutical industry (see Fox et al. 2009, for example). Other papers use simulation-based frameworks to derive PPE usage (Barrett et al. 2020). These approaches are difficult to reproduce, and thus, their estimation error is hard to quantify. Our work is the first to propose an analytical predictive model of PPE demand in a clinical setting that can be deployed at multiple scales (departmental, hospital, regional), settings (outbreaks or regular operations), and can also be independently used by administrative personnel for operational planning and supply management.

4.3 Model Formulation and Workload Estimation

In this section, we introduce a general stochastic queueing model and describe its suitability in estimating the amount of PPE required for a hospital department. Let $\mathcal{I}$ be a set of patient classes defined using managerially-relevant features, for instance, demographic characteristics, patients with varying acuity levels, clinical diagnoses, and length-of-stay. Classes should be chosen such that all patients in class-$i \in \mathcal{I}$ have similar care paths, i.e., a sequence of medical investigations and interventions, and LoS values. Class i patients are assumed to arrive to the hospital and be admitted according to a non-homogeneous Poisson process $\Lambda_i(t)$ with time-varying intensity rate $\lambda_i \equiv \lambda_i(t)$. Further, each class-$i$ patient stays at the hospital for a random time S_i which represents their length-of-stay (LoS); we define the corresponding stochastic vector $\boldsymbol{S} \coloneqq (S_1, S_2, ..., S_I)$. The LoS for each patient within each class is independent and identically distributed where class-i patients have cumulative distribution function G_i. Finally, we assume that S_i is independent from $\Lambda_i(t)$ for any time $t \in \mathbb{R}$.

Our goal is to estimate the total clinical workload of a hospital department, which in turn, will allow us to predict the PPE required. Thus, we do not restrict hospital capacity and instead, assume that practitioners can provide medical care to any admitted patient as soon as they arrive. As a result, we estimate the total workload of a hospital department by aggregating the workload from $I = |\mathcal{I}|$ independent $M_t/G/\infty$ queues leveraging the merging/splitting property of a Poisson process. Inferring the workload from such systems is a standard modelling technique in the operations literature (Eick et al. 1993b, Massey and Whitt 1993, Feldman et al. 2008b). In addition, patients transferred from one clinical service to another are considered discharged by the former and newly admitted by the latter. Such events are rare and thus, we can consider these individuals as new arrivals for estimation purposes. Note that the intensive care unit (ICU) constitutes an exception to this rule: between 5 to 10 percent of GIM patients are transferred to the ICU at least once over the duration of their treatment. In this study, we consider the ICU as an external service, and, thus, subtract the times patients spend there from their total length-of-stay.

Let $\{\Delta_i(t) | t \in \mathbb{R}\}$ be a headcount stochastic process corresponding to the number of class i patients being discharged over the interval $[0, t]$. Applying Theorem 1 in Eick et al. (1993b), we obtain the steady-state probability distribution of $\Delta_i(t)$. Because the GIM service has been continuously operating for a long time, the steady-state assumption is appropriate in our setting. Specifically, the number of class i patients discharged over the interval $[0, t]$ is given by $\{\Delta_i(t) | t \in \mathbb{R}\}$ which is a non-homogeneous Poisson process with mean

$$\mathbb{E}[\Delta_i(t)] := \int_0^t \int_0^\infty \lambda_i(u - s) dG_i(s) du \qquad \forall i \in \mathcal{I}. \tag{4.1}$$

Notice that, following the framework of Eick et al. (1993b), we assume $t \in \mathbb{R}$ but only consider the dynamics of the system at times $t \geq 0$.

Unfortunately, for most LoS distributions, (4.1) must be computed numerically as closed-form expressions do not exist unless, for example, G_i is exponentially distributed. In addition, the departure process $\Delta_i(T)$ is dependent on the LoS of class-i patients. As a result, we condition on the individual quantiles of the LoS distribution for each class $i \in \mathcal{I}$. More specifically, let σ_i be the desired quantile value for class-i patients where we define $\boldsymbol{\sigma} := (\sigma_1, \ldots, \sigma_I)$ and let $\Delta_i(t; \sigma_i)$ denote the departure process of class-i patients conditioned on $S_i = \sigma_i$ for each $i \in \mathcal{I}$. Thus, $\{\Delta_i(t; \sigma_i) | t \in \mathbb{R}\}$ is a headcount stochastic process that represents the number of class i patients discharged over the interval $[0, t]$ with LoS value equal to σ_i. This corresponds to a non-homogeneous Poisson process with mean

$$\mathbb{E}[\Delta_i(t, \sigma_i)] := \int_0^t \lambda_i(u - \sigma_i) du \qquad \forall i \in \mathcal{I}. \tag{4.2}$$

4.3.1 Prediction of Demand for PPE

Multiple types of PPE are used in clinical settings, such as surgical masks, N95 respirators, gloves, face shields, etc. Further, demand for different kinds of PPE varies depending on the nature of the interaction between patients and practitioners as well as current public health regulations and institutional guidelines (see Ontario Health 2020b, Barrett et al. 2020, for instance). Thus, we assume that a hospital uses N different types of PPE in its daily operations.

Total demand of PPE comprises all protective equipment used by employees, i.e., medical staff, and patients. Although, in this study, we assume that patients admitted to the hospital occupy separate rooms and do not need to wear PPE while on their own, our model can be naturally extended to account for patients with shared accommodations. Further, hospital policy dictates that clinicians wear a surgical mask and a face shield for all interactions with hospitalized patients. Additional precautions may be used by hospital staff and clinicians when performing particular procedures and/or assessments. There may also be separate regulations for patients who are placed in a higher level of isolation, such as those diagnosed with COVID-19. As a result, we define Q_n^m to be the total quantity of type $n \in \{1, 2, ..., N\}$ PPE used by employees when no interaction with patients takes place and $Q_{i,n}^u$ to be the amount of type n PPE used by medical staff during interactions with class i patients. Thus, the total demand for type n PPE is given by

$$Q_n := Q_n^m + \sum_{i=1}^{I} Q_{i,n}^u \qquad \forall n \in \{1, 2, \ldots, N\}. \tag{4.3}$$

We assume, without loss of generality, that PPE is not reused but discuss this extension in Section 4.4.3.

Define $\boldsymbol{m} := (m_1, m_2, \ldots, m_N)'$ to be a vector such that element m_n represents the average number of type n PPE items used daily by an employee when not interacting with patients. Then,

$$Q_n^m = m_n W(T) \qquad \forall n \in \{1, 2, \ldots, N\}, \tag{4.4}$$

where $W(T)$ is the number of estimated work days of all medical employees over the planning horizon. In particular, note that we assume Q_n^m increases linearly in the workload $W(T)$. Discussions with medical practitioners indicate that this is the most appropriate model.

Suppose there are J different categories of clinical interactions such as nursing (e.g., vital signs measurement, medication administration), physician visits, medical testing, and surgical procedures. Define an $I \times J$ matrix $\boldsymbol{C}$ where element $c_{i,j}$ is the average daily number of clinical interactions from category j that are required by a class i patient (note: median values can also be used to reduce the effect of outliers although we did not observe any appreciable difference in our results). We also define an $I \times J$ matrix $\boldsymbol{U}_n$ such that element $u^n_{i,j}$ represents the average number of type n PPE items used during each category j interaction with a patient of class i (see Table C.2 in Appendix C). Then,

$$Q^u_{i,n} = \sigma_i \Delta_i(T; \sigma_i) \sum_{j=1}^{J} c_{i,j} u^n_{i,j} \tag{4.5}$$

$$\forall i \in \{1, 2, \ldots, I\}, \forall n \in \{1, 2, \ldots, N\},$$

where the estimate is conditioned on the LoS value σ_i.

Notice that $c_{i,j}u^n_{i,j}$ represents the average daily number of type n PPE used by class i patients during all medical interactions belonging to category j. Aggregating over each j and multiplying by the stochastic quantity S_i gives the average number of type n PPE used by a class i patient during their length-of-stay in the hospital. Finally, multiplying these terms by the integral of the headcount stochastic process gives the average amount of type n PPE used by all class i patients discharged over the specified time horizon T.

As noted above, $\Delta_i(T)$ and S_i are dependent, i.e., the number of discharged patients at time t is a function of the LoS of class-i patients. This makes deriving the marginal expectation of Q_n cumbersome to obtain. Instead, in the following lemma, we leverage (4.2) and derive the conditional expectation of Q_n given that the LoS of class-i patients is fixed to a given quantile.

Lemma 12 (Conditional Expectation)**.** *For every i, suppose $\sigma_i > 0$ and $T > \sigma_i$. Then,*

$$\mathbb{E}[Q_n | \boldsymbol{S} = \boldsymbol{\sigma}] = \sum_{i=1}^{I} \sigma_i \sum_{j=1}^{J} c_{i,j} u^n_{i,j} \int_0^T \lambda_i(u - \sigma_i) du \tag{4.6}$$

$$+ m_n W(T), \ \forall n \in \{1, 2, \ldots, N\}.$$

Equation (4.6) is derived by conditioning on a particular quantile of the LoS distribution. For example, if $\sigma_i = \mathbb{E}[S_i]$ for all $i \in \mathcal{I}$, then for a class-i patient, (4.6) considers the dynamics of the average stochastic path of the departure process $\Delta_i(t; \sigma_i)$ as the total number of paths grows to infinity. Further, as the variances of the hospital LoS and the average daily counts of medical interactions decrease, the gap between the conditional and unconditional expectation of the demand for type n PPE (Q_n) also decreases. Thus, (4.6) provides a better approximation to the demand for PPE if the classes of patients are selected such that their LoS and treatment requirements are relatively similar; this motivates why patients should first be clustered into I classes.

Table 4.1: Summary of the notation.

$\lambda_i(t)$	class i patient's rate of admission to the hospital
S_i	random variable corresponding to the hospital length-of-stay of a class i patient
$\boldsymbol{S}$	$I \times 1$ stochastic vector of length-of-stay random variables
σ_i	desired quantile value chosen for the length-of-stay distribution of a class i patient
$\boldsymbol{\sigma}$	$I \times 1$ vector of length-of-stay quantile values
$G_i(t)$	cumulative distribution function for the length of stay of class i patient
$\Delta_i(t)$	stochastic process counting the number of class i patients discharged over $[0, t]$
$\Delta_i(t; \sigma_i)$	stochastic process counting the number of class i patients discharged over $[0, t]$ conditional on $S_i = \sigma_i$
Q_n	total stochastic demand for type n PPE
Q_n^m	total stochastic demand for type n PPE by all hospital employees while not interacting with patients
$Q_{i,n}^u$	total stochastic demand for type n PPE by all hospital employees during their interactions with class i patients
$\boldsymbol{C}$	$I \times J$ matrix of average daily counts of medical interactions j required by class i patient
$\boldsymbol{m}$	$N \times 1$ vector of average daily counts of type n PPE used by hospital employees while not interacting with patients
$\boldsymbol{U}_n$	$I \times J$ matrix of average number of type n PPE used during medical interaction j with a class i patient

4.4 Data Description and Results

We apply our approach to estimate the PPE needs for the GIM service at St. Micheal's hospital. The GIM accounts for approximately 40% of all emergency department admissions to the hospital Verma et al. (2017) and cares for patients with a broad range of diseases Verma et al. (2018) while focusing on cases with complex medical needs. Because the operations at St. Michael's Hospital is directly affected by the COVID-19 pandemic, effective prediction of PPE usage is critical to their inventory planning and their ability to deliver adequate medical care.

To parameterize our predictive model, we used 9 years of data from April 2010 to November 2019 collected from St. Michael's Hospital by the General Medicine Inpatient Initiative (GEMINI) Verma et al. (2017). The data set includes both administrative and clinical records of discharged patients. GEMINI data sets have been rigorously validated and are demonstrated to be highly reliable Pasricha et al. (2020). Our data set comprises of 37,492 hospital admissions for 22,039 unique patients whose median age is 66 years old (52, 79), where values in brackets correspond to the first and third quartile, respectively. Approximately 43% of hospital admissions to the GIM are by female patients and the five most common clinical diagnoses are chronic obstructive pulmonary disease and bronchiectasis (6%), pneumonia (5%), acute cerebrovascular disease (5%), urinary tract infections (5%), and gastrointestinal hemorrhages (4%).

The median value for LoS is 4.83 days (2.58, 9.54) which suggests an asymmetrical probability distribution. We determine the average daily counts of medical interactions per patient as well as the corresponding type of interaction and PPE usage from the data set and by interviewing medical experts in the partner hospital (see Appendix C for details on the semi-structured interview protocol). Notice that Table C.2, provided in Appendix C, displays the average amount of PPE used during all medical interactions in addition to the items already worn by clinical staff when not interacting with patients. Thus, in cases where no additional PPE is required, the value in the table is equal to zero. Alternatively, some medical interactions are conducted by multiple practitioners which means that a larger amount of PPE is required. For example, surgical procedures typically require two porters, a surgeon and one or two trainees, an anesthesiologist, and two nurses.

When not interacting with patients, medical staff require two surgical masks per shift (which is approximately 12 hours in length) and one face shield per week. The GIM service at St. Michael's Hospital requires 50 nurses, 4 phlebotomists, 10 porters, 20 doctors, 3 physiotherapists, 3 occupational therapists, 2 dietitians, 2 language pathologists, and 3 discharge planners each day. For simplicity, we assume that shifts of all medical staff are of the same length. Notice that this assumption is easy to relax. Finally, we consider seven types of PPE ($N = 7$): gloves, gowns, surgical masks, N95 masks, face shields, bouffants, and boot covers.

We use equation (4.6) to derive an annual estimate of PPE usage by clustering all patients into classes based on the nature of their medical interactions as well as their length-of-stay within the hospital. To account for the aforementioned asymmetry in the LoS distribution, and since (4.6) computes a conditional expectation, we evaluate PPE usage assuming that LoS remains at one of its quantile values for each class. We fix our planning horizon (T) to one year (365 days) and estimate the value of $\int_0^{T-\sigma_i} \lambda_i(u)du$ by calculating the number of class-i discharges that occur during a typical year prior to the pandemic. In the remainder of this section, we confirm that a non-homogeneous Poisson distribution best describes the arrival process. We then discuss how we cluster patients into classes and present estimates of the projected annual PPE usage.

4.4.1 Testing the Non-homogeneous Poisson Assumption

Because our data set contains the arrival times and discharge times of each patient, the number of discharges from the GIM over a planning horizon can be computed without evaluating the integral in (4.6). However, there are many cases where such fine-grained data is not available. In such settings, only arrival times and/or LoS values may be accessible. In other cases, the prediction interval set by the modeller may be sufficiently short (e.g., daily or weekly) which necessitates the evaluation of a functional form of departing patients at time t. In these scenarios, computing the integral is essential. Therefore, both for completeness and to ensure that the analytical representation of the demand for PPE in (4.6) is valid, we test the assumption that the arrival process follows a non-homogeneous Poisson distribution.

We closely follow the procedure described in Brown et al. (2005), i.e., we test the null hypothesis (H_0) that admissions to the GIM follow a Poisson distribution with a piecewise constant rate. To do this, we break up the planning horizon into progressively smaller non-overlapping time intervals. Note that, for this analysis, we consider admissions to the GIM from the two most recent years in order to account for possible changes in the demand for GIM services. We then continue to decrease the length of these intervals until the arrival rate remains stationary over at least 90% of the constructed intervals. We test the hypothesis of stationarity by applying the Kolmogorov-Smirnov (KS) test and confirming that, for each time interval, the logarithmically transformed arrival times can be modeled by independent standard exponential random variables.

Table 4.2: Testing the non-homogeneous Poisson assumption for different time intervals.

Number of Intervals	Length (days)	% Not Rejected By KS Test
10	90.8	0.00
20	43.0	35.00
30	28.2	63.30
40	21.0	80.00
80	10.3	88.75
800	1.00	90.38

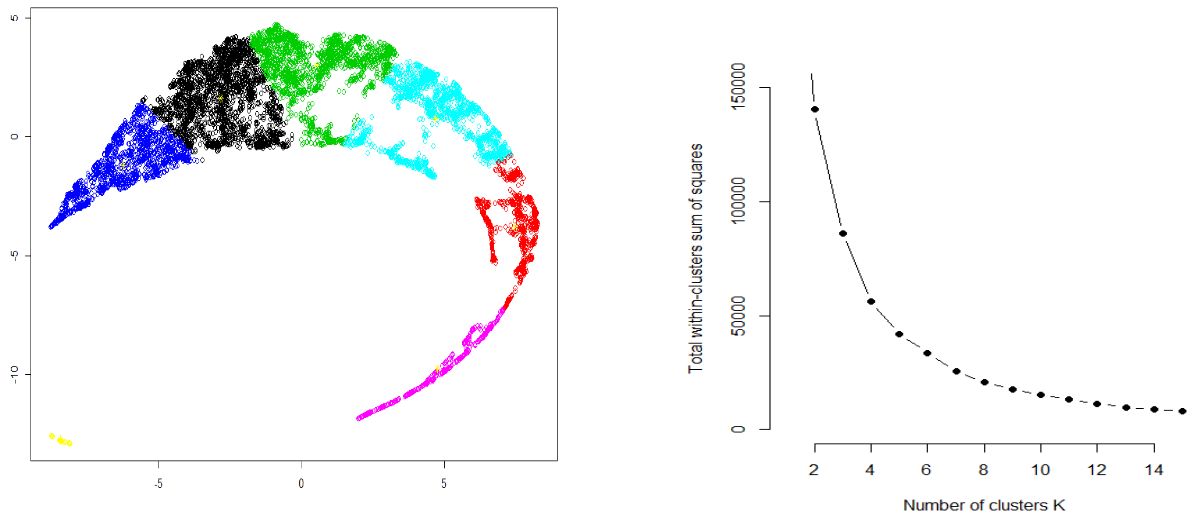
According to Table 4.2, as the length of each interval reduces to one day, the arrival rates over 90% of the intervals follow a Poisson distribution with a stationary rate according to the KS test (0.05 significance level). This implies that a non-homogeneous Poisson distribution best describes the arrival rate and that a $M_t/G/\infty$ modelling framework is appropriate for this application.

4.4.2 Clustering Results

To ensure patient classes have similar care paths and LoS values, we cluster patients into groups based on the nature of their medical interactions (15 types) and LoS (see Table C.2 in Appendix C). In our data set, each medical intervention is captured by a set of timestamps. To avoid counting the same patient-practitioner interaction multiple times, we assume that all timestamps within a one hour interval are related to a single interaction. This assumption is critical as some interactions between patients and practitioners result in multiple timestamps that are minutes apart (e.g., vital signs, the administration of drugs, and laboratory test collections) and, thus, reflect a single episode of PPE use.

We use the Uniform Manifold Approximation and Projection (UMAP) algorithm paired with the k-means clustering algorithm to group patients into classes. UMAP is a dimensionality reduction technique based on Riemannian geometry and algebraic topology that projects high-dimensional data (15 types of medical interactions and LoS) onto a two-dimensional space; see Figure 4.1a for a visual representation. The smaller total squared error within a cluster implies that there is a high similarity of patients assigned to that class. It also improves the quality of our conditional estimate of demand for PPE. Thus, we determine the optimal number of clusters by applying the k-means clustering algorithm which minimizes the total squared error within each cluster. We use the elbow method to determine the best value of k Joshi and Nalwade (2013).

Figure 4.1: Clustering results.



(a) Cluster Visualization

(b) Elbow Plot

As demonstrated in Figure 4.1b, the within cluster error decreases slowly as the number of clusters exceeds 7; adding more clusters does not model the data significantly better. For more information on the clustering approach, please see Appendix C and Table C.1.

To illustrate the effect of our clustering procedure, we present a quantile summary of the LoS (days) distribution by comparing non-clustered patients to the clustered results. Having relatively similar LoS values in each cluster is important as we would like its within-cluster variation to decrease so that the gap between our conditional estimates and their corresponding marginal quantities is small. If all patients are assigned to a single class, one quarter stay in the GIM between 0 and 1.9 days (first quartile); similarly, 25% of patients remain in the GIM more than 7.9 but less than 354.2 days (fourth quartile). The clustered patients, however, have more similar LoS ranges. In particular, cluster one contains patients who remain in the GIM for a very short period of time, cluster two and three are assigned patients who stay in the hospital less than one week, cluster four includes patients who stay in care less than 11 days, cluster five includes patients with LoS shorter than 20 days, and cluster six includes patients who stay in the facility significantly longer. Cluster seven, which contains approximately 1% of patients, represent the departments heaviest users. We note that some clusters have overlapping LoS ranges because other factors describing their care path differ.

Table 4.3: The effect of clustering on the different quantiles for the LoS distribution (days).

	0%	25%	50%	75%	100%
Cluster 1 of 1 (100%)	0.0	1.9	3.9	7.9	354.2
Cluster 1 of 7 (18%)	0.0	0.5	0.8	1.4	4.8
Cluster 2 of 7 (27%)	0.1	1.7	2.3	2.9	6.4
Cluster 3 of 7 (22%)	0.4	3.7	4.5	5.2	7.1
Cluster 4 of 7 (17%)	5.3	6.9	7.9	9.3	11.7
Cluster 5 of 7 (10%)	10.8	12.7	14.3	16.6	20.9
Cluster 6 of 7 (6%)	20.4	24.2	29.2	35.8	57.0
Cluster 7 of 7 (1%)	59.0	65.6	82.0	128.2	354.2

4.4.3 PPE Estimation Results

We apply equation (4.6) to compute the total demand for type n PPE using 5, 6, 7, and 8 cluster partitions. To describe its distribution, we condition our estimates on the quartiles of the LoS and present the results in Table 4.4, where the first and third rows per each cluster quantity correspond to the lower and upper bounds of PPE usage. According to the seven-cluster estimates in Table 4.4, gloves and surgical masks are the most prevalent items as they constitute 90% of the total PPE predicted. Further, the annual usage of gowns represents only 3% (similarly to bouffants and boot covers) of the total (454,324) amount of PPE used, while N95 masks constitute only 2%. As a reminder, due to the nature of our data, these estimates account for non-COVID-19 patients only, i.e., those patients who are not under investigation for the Coronavirus. However, our model is flexible enough and can accommodate these patients as separate classes if the data becomes available.

Table 4.4: Prediction of PPE usage as a function of the number of clusters.

LoS Quartile	Gloves	Gowns	Surgical Masks	N95 Masks	Face Shields	Bouffants	Boot Covers
Five Clusters ($I = 5$)							
Q1	122,771	6,422	169,193	4,094	3,906	6,422	6,422
Median	206,459	10,748	180,093	6,891	3,906	10,748	10,748
Q3	264,107	13,785	187,208	8,787	3,906	13,785	13,785
Six Clusters ($I = 6$)							
Q1	134,232	6,935	169,917	4,385	3906	6,935	6,935
Median	219,111	11,348	180,954	7,239	3906	11,348	11,348
Q3	279,440	14,517	188,221	9,203	3906	14,517	14,517
Seven Clusters ($I = 7$)							
Q1	129,216	6,779	169,233	4,229	3,906	6,779	6,779
Median	226,007	11,721	181,774	7,476	3,906	11,721	11,721
Q3	277,995	14,433	187,989	9,161	3,906	14,433	14,433
Eight Clusters ($I = 8$)							
Q1	151,878	7,839	171,964	4,980	3,906	7,839	7,839
Median	229,751	11,850	182,296	7,610	3,906	11,850	11,850
Q3	274,123	14,163	187,491	9,051	3,906	14,163	14,163

Table 4.4 also helps to understand the sensitivity of our results to the number of patient classes specified by the modeller. In general, we observe higher predictions in the amount of PPE as the number of clusters increases. This is because the average and median values of features included in the clustering procedure are more heavily influenced by larger-valued observations. However, the increase in predicted PPE usage with the number of clusters is sample specific; data sets with fewer outliers may have a decreasing pattern. Although using more clusters decreases the total squared error, fewer data points contribute to the length-of-stay estimate. This may lead to an inaccurate prediction for the LoS distribution even though patients may have similar care plans. Furthermore, the estimates may overfit to the data in the sample. Thus, we advise that a modeller

does not increase the number of clusters too far beyond the point that is recommended by the elbow method.

We find that some types of PPE, such as surgical masks and face shields, show little variation in forecasted demand. That is, their quartile estimates are similar regardless of the number of patient classes chosen. This is because the majority of the annual need for these types of PPE occur when practitioners are not interacting with patients; the estimate is 156,220 and 3,906 for surgical masks and face shields, respectively. Thus, while demand for gloves is solely driven by the number of medical interactions, 86% of surgical mask use is driven by the requirement that medical employees must wear a mask whilst in the hospital.

Finally, we note that the above approach can be adapted to address situations where PPE can be reused. In particular, let γ_n be the proportion of type n PPE which can be reused over r_n interactions. Then, $(1-\gamma_n)\mathbb{E}[Q_n|\boldsymbol{S}=\boldsymbol{\sigma}]+\frac{\gamma_n}{r_n}\mathbb{E}[Q_n|\boldsymbol{S}=\boldsymbol{\sigma}]$ represents the total predicted demand of type n PPE.

4.5 Conclusions

In this paper, we leverage results time-varying queueing models to present a prediction framework that can be used to forecast the amount of PPE required over a specified time horizon. To this end, we first cluster patients with similar hospital experiences into classes and estimate their LoS in the hospital as well as the PPE requirements of each patient-practitioner interaction. By demonstrating that the dynamics of each patient class can be modelled using an $M_t/G/\infty$ queue, we present closed-form estimates for the expected amount of PPE required for each patient class and aggregate the results together to generate a prediction of PPE usage.

We contribute to the pandemic and supply chain disruption literature by helping practitioners mitigate unexpected changes in demand when disruptions do not affect the operation of a service, but instead, prompt new mandatory regulations that affect the equipment used in its performance. Moreover, our analysis provides bounded estimates that anticipate the time-variability in the system. In particular, using current PPE-usage guidelines under COVID-19, we find that the general internal medicine department at our partner hospital must anticipate much higher demand for gloves and surgical masks than gowns. The former comprises 90% of the total 454,324 items predicted while the latter accounts for only 3% of the annual PPE usage. In addition, our analysis suggests that only 14% of demand for surgical masks in a hospital setting is caused by interactions with patients. Thus, an annual estimate of usage for this type of PPE is expected to be less volatile than the anticipated demand for gloves.

As suggested in Section 4.3, our approach is versatile and computationally efficient. A simple application of Lemma 12 admits a back-of-the-envelope calculation. In this case, the aggregate number of departures from the system per patient type as well as a quantile estimate for the LoS are sufficient to derive bounded conditional estimates of PPE usage. Contrary to Barrett et al. (2020), for instance, our predictions do not require that an extensive simulation study be constructed; the technique we develop is not restricted to estimates of PPE during a quarantine and can be applied to other settings such as normal hospital operations. In addition, our approach may be used for a comparative analysis. For example, if patient classes are pre-specified by a medical practitioner, the demand for PPE can be estimated and compared for multiple choices of arrival functions and LoS distributions over a planning horizon of arbitrary length. Our time-varying queueing framework naturally accommodates this exploratory approach by providing an analytical way of estimating the total number of departures conditioned on a carefully selected LoS value.

Although our PPE prediction tool can be applied to a wide variety of clinical settings, our study includes a number of data-specific limitations. In particular, the guidelines governing the use of PPE for each type of medical interaction, as summarized in Table C.2 in Appendix C, is distinct to St. Michael's Hospital. These estimates may vary depending on the location and clinical focus of the medical institution under consideration. Further, we estimate the clinical workload generated by typical medical interactions based on the data collected prior to the COVID-19 pandemic, i.e., we exclude both confirmed COVID-19 patients and patients who are under investigation for the virus. As this data becomes available, the PPE needs for these patient categories can be estimated and added to the prediction model. While 15 important types of clinical interactions are captured in the data set, some are represented more crudely than others. For example, a nurse who assists a patient with toileting or bathing is not captured. As a result, our approach may underestimate the hospital's overall PPE needs. However, these limitations may be addressed by collecting additional data fields or by consulting a patients electronic health record. To this end, future research should seek to validate the predicted estimates of PPE usage against real-world demand.

Despite these limitations, our methodology complements ongoing efforts that help to manage supply chains during the COVID-19 pandemic. For instance, using an arrival function estimated by SIR models, we can derive the corresponding PPE requirements over a planning horizon of arbitrary length. Our study also shows good synergy with emerging platforms that connect PPE suppliers to consumers (University of Wisconsin 2020, Afèche et al. 2020) as consumers can more accurately predict their PPE usage and liaise with suppliers that have the requisite capacity.

Chapter 5

Conclusions

In this dissertation, we propose computationally efficient techniques for stationary resource allocation in anticipation of time-varying dynamics. We apply these methods to the settings of customer acquisition and retention, cloud computing and healthcare.

In Chapter 2, we introduce a new methodology to assign a fixed pool of servers to multiple classes of homogeneous clients within an environment with time-varying demand. Leveraging a fluid and dynamical systems approach, our methodology proposes a stationary staffing policy that performs well in anticipation of non-stationary dynamics. We also extend the literature on customer acquisition and retention by incorporating accessibility as a measure of service quality. This provides a link between staffing decisions and the effect they have on service quality and customer defections. We conduct an extensive numerical study to show that our staffing policy outperforms modifications of the commonly-used square-root staffing rule (SRS) by providing better access to service for all customers while maintaining high levels of throughput. Thus, by finding the optimal allocation of servers to customer classes, our approach endogenously balances acquisition and retention efforts.

In Chapter 3, we develop a new calculus-of-variations approach to determine the jointly optimal stationary capacity level and retrial rates for an environment with time-varying demand and jobs that have a finite number of retrial attempts. The proposed methodology overcomes challenges associated with the analysis of non-Markovian retrial queues where models must be restricted to only a few servers and/or arriving customers are assumed to be infinitely patient. Leveraging an offered load approximation that decomposes the workload into new versus retrial jobs, we construct a recursive representation of the offered load function (OLF) that describes the fluid dynamics of the original stochastic system. This OLF is piecewise smooth and has a closed-form expression for a large class of arrival patterns. To derive the optimal policy, we minimize the total variation of the OLF and show that the optimal capacity level and retrial rates can be explicitly derived for a variety of cases - albeit, with some assumptions - such as when the arrival function is sinusoidal or is piecewise smooth, continuous, and periodic. Finally, we prove that minimizing the functional variation of the constructed OLF is equivalent to maximizing system throughput.

In Chapter 4, in response to the increasing demand for PPE induced by the ongoing COVID-19 pandemic, we predict the annual requirements for multiple types of PPE in the GIM department at St. Michael's hospital in Toronto, Canada. Our estimation approach features a queueing framework with time-varying exogenous arrivals and generally distributed service times. We use this model to predict the workload of patients admitted to the GIM and determine a functional relationship mapping this workload to PPE used by medical staff. Further, we parametrize our predictive model by conducting interviews with medical practitioners of the partner institution and by collecting operational and clinical data over a 9-year period. This allows us to estimate a patient's length-of-stay (LoS), the number and nature of different patient-practitioner interactions, and the number of PPE used per interaction. Using current PPE-usage guidelines under COVID-19, we find that the general internal medicine department at our partner hospital must anticipate much higher demand for gloves and surgical masks than gowns. The former comprises 90% of the total 454,324 items predicted while the latter accounts for only 3% of the annual PPE usage. In addition, our analysis suggests that only 14% of demand for surgical masks in a hospital setting is caused by interactions with patients. Thus, an annual estimate of usage for this type of PPE is expected to be less volatile than the anticipated demand for gloves.

.

Bibliography

Afèche P, Baron O, Hu M, Krass D (2020) COVID PPE help Accessed from `http://www.covidppehelp.ca/`.

Afèche P, Araghi M, Baron O (2017) Customer acquisition, retention, and service access quality: Optimal advertising, capacity level, and capacity allocation. *Manufacturing & Service Operations Management* 19(4):674–691.

Aguir S, Akşin Z, Karaesmen F, Dallery Y (2008) On the interaction between retrials and sizing of call centers. *European Journal of Operational Research* 191(2):398–408.

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

Akduman D, Kim LE, Parks RL, L'Ecuyer PB, Mutha S, Jeffe DB, Evanoff BA, Fraser VJ (1999) Use of personal protective equipment and operating room behaviors in four surgical subspecialties: personal protective equipment and behaviors in surgery. *Infection Control and Hospital Epidemiology: The Official Journal of the Society of Hospital Epidemiologists of America* 20(2).

Aksin Z, Armony M, Mehrotra V (2007) The modern call center: A multi-disciplinary perspective on operations management research. *Production and operations management* 16(6):665–688.

Anderson Jr EG, Morrice DJ, Lundeen G (2006) Stochastic optimal control for staffing and backlog policies in a two-stage customized service supply chain. *Production and Operations Management* 15(2):262–278.

Andradóttir S, Ayhan H, Down DG (2001) Server assignment policies for maximizing the steady-state throughput of finite queueing systems. *Management Science* 47(10):1421–1439.

Andradóttir S, Ayhan H, Down DG (2003) Dynamic server allocation for queueing networks with flexible servers. *Operations Research* 51(6):952–968.

Andradóttir S, Ayhan H, Down DG (2007) Dynamic assignment of dedicated and flexible servers in tandem lines. *Probability in the Engineering and Informational Sciences* 21(4):497–538.

Artalejo J, Falin G (2002) Standard and retrial queueing systems: a comparative analysis. *Revista matemática complutense* 15(1):101–129.

Artalejo JR (2010) Accessible bibliography on retrial queues: Progress in 2000–2009. *Mathematical and computer modelling* 51(9-10):1071–1081.

Balmer C, Pollina E (2020) Italy's Lombardy asks retired health workers to join Coronavirus fight. *World Economic Forum, Reuters*.

Bardina X, Ferrante M, Rovira C (2020) A stochastic epidemic model of COVID-19 disease. *Preprint* Available from arXiv: 2005.02859.

Barrett K, Nakamachi Y, et al. (2020) Estimated demand for personal protective equipment for Ontario acute care hospitals during the COVID-19 pandemic. Accessed from `https://drive.google.com/file/d/1LEW0irL6426OIYtVYo-s-XMHrvxalVh1/view?usp=sharing`.

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

Bassamboo A, Randhawa RS (2010) On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers. *Operations research* 58(5):1398–1413.

Bassamboo A, Randhawa RS (2015) Scheduling homogeneous impatient customers. *Management Science* 62(7):2129–2147.

Bassamboo A, Zeevi A (2009) On a data-driven method for staffing large call centers. *Operations Research* 57(3):714–726.

Batta R, Berman O, Wang Q (2007) Balancing staffing and switching costs in a service center with flexible servers. *European journal of operational research* 177(2):924–938.

BCE (2016) Annual report. Accessed from `http://www.bce.ca/investors/AR-2016/2016-bce-annual-report.pdf`.

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

Benson SM, Novak DA, Ogg MJ (2013) Proper use of surgical n95 respirators and surgical masks in the OR. *AORN journal* 97(4):457–470.

Berger PD, Nasr NI (1998) Customer lifetime value: Marketing models and applications. *Journal of interactive marketing* 12(1):17–30.

Bhandari A, Scheller-Wolf A, Harchol-Balter M (2008) An exact and efficient algorithm for the constrained dynamic operator staffing problem for call centers. *Management Science* 54(2):339–353.

Bioucas-Dias JM, Figueiredo MA (2010) Multiplicative noise removal using variable splitting and constrained optimization. *IEEE Transactions on Image Processing* 19(7):1720–1730.

Biswas K, Khaleque A, Sen P (2020) COVID-19 spread: Reproduction of data and prediction using a SIR model on Euclidean network. *Preprint* Available at arXiv: 2003.07063.

Blattberg RC, Deighton J (1996) Manage marketing by the customer equity test. *Harvard business review* 74(4):136.

Borovkov A (1967) On limit laws for service processes in multi-channel systems. *Siberian Mathematical Journal* 8(5):746–763.

Borst S, Mandelbaum A, Reiman MI (2004) Dimensioning large call centers. *Operations research* 52(1):17–34.

Boushey H, Glynn SJ (2012) There are significant business costs to replacing employees. *Center for American Studies* 16.

Brown L, Gans N, Mandelbaum A, Sakov A, Shen H, Zeltyn S, Zhao L (2005) Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American statistical association* 100(469):36–50.

Brown M (1969) An invariance property of poisson processes. *Journal of Applied Probability* 6(2):453–458.

Bruckner RM, Tjoa AM (2002) Capturing delays and valid times in data warehouses – towards timely consistent analyses. *Journal of Intelligent Information Systems* 19(2):169–190.

Buttle F (2004) *Customer relationship management* (Routledge).

Buzacott JA, Shanthikumar JG (1993) *Stochastic models of manufacturing systems*, volume 4 (Prentice Hall Englewood Cliffs, NJ).

Calafiore GC, Novara C, Possieri C (2020) A modified SIR model for the COVID-19 contagion in Italy. *Preprint* Available at arXiv: 2003.14391.

Campbell D, Frei F (2011) Market heterogeneity and local capacity decisions in services. *Manufacturing & Service Operations Management* 13(1):2–19.

Canadagov (2020) Non-medical masks and face coverings: About Accessed from `https://www.canada.ca/en/public-health/services/diseases/2019-novel-coronavirus-infection/prevention-risks/about-non-medical-masks-face-coverings.html`.

Canadian Centre for Ocupational Health and Safety (2018) Personal protective equipment (PPE) Accessed from `https://www.ccohs.ca/teach_tools/phys_hazards/ppe.html`.

Carbonara N, Pellegrino R (2018) Real options approach to evaluate postponement as supply chain disruptions mitigation strategy. *International Journal of Production Research* 56(15):5249–5271.

CDC (2020) CDC calls on americans to wear masks to prevent COVID-19 spread Accessed from `https://www.cdc.gov/media/releases/2020/p0714-americans-to-wear-masks.html?start=yes`.

Chan CW, Sarhangian V (2018) Dynamic server assignment in multiclass queues with shifts, with application to nurse staffing in emergency. Accessed from http://www.columbia.edu/~cc3179/shift_scheduling_2017.pdf.

Chan CW, Yom-Tov G, Escobar G (2014) When to use speedup: An examination of service systems with returns. *Operations Research* 62(2):462–482.

Chan FT, Chung SH (2004) A multi-criterion genetic algorithm for order distribution in a demand driven supply chain. *International Journal of Computer Integrated Manufacturing* 17(4):339–351.

Chang V, Walters RJ, Wills GB (2016) Organisational sustainability modelling – an emerging service and analytics model for evaluating cloud computing adoption with two case studies. *International Journal of Information Management* 36(1):167–179.

Chang V, Wills G, De Roure D (2010) A review of cloud business models and sustainability. *2010 IEEE 3rd International Conference on Cloud Computing*, 43–50 (IEEE).

Chapel J (2019) Cloud waste to hit over $14 billion in 2019. *DevOps.com* Accessed from https://devops.com/cloud-waste-to-hit-over-14-billion-in-2019/.

Chen F, Federgruen A, Zheng YS (2001) Coordination mechanisms for a distribution system with one supplier and multiple retailers. *Management science* 47(5):693–708.

Chen IJ, Popovich K (2003) Understanding customer relationship management (CRM) people, process and technology. *Business process management journal* 9(5):672–688.

Chen K, Xiao T (2009) Demand disruption and coordination of the supply chain with a dominant retailer. *European Journal of Operational Research* 197(1):225–234.

Chen N, Lee D, Shen H (2018) Can customer arrival rates be modelled by sine waves? *Preprint* Available from at SSRN 3125120.

Chen S, Lee H, Moinzadeh K (2019) Pricing schemes in cloud computing: Utilization-based vs. reservation-based. *Production and Operations Management* 28(1):82–102.

Chen X, Qiu Z (2020) Scenario analysis of non-pharmaceutical interventions on global COVID-19 transmissions. *Preprint* Available from arXiv: 2004.04529.

Chevalier P, Tabordon N (2003) Overflow analysis and cross-trained servers. *International Journal of Production Economics* 85(1):47–60.

Chiang YJ, Ouyang YC (2014) Profit optimization in sla-aware cloud services with a finite capacity queuing model. *Mathematical Problems in Engineering* 2014.

Chiou JS, Wu LY, Hsu JC (2002) The adoption of form postponement strategy in a global logistics system: the case of taiwanese information technology industry. *Journal of Business Logistics* 23(1):107–124.

Choi K, Narasimhan R, Kim SW (2012) Postponement strategy for international transfer of products in a global supply chain: A system dynamics examination. *Journal of operations Management* 30(3):167–179.

Cohen J (1900) Basic problems of telephone traffic the ory and the influence of repeated calls. *Pillips Telecomm. Rev.* 18(2).

Columbus L (2016) 2015 gartner CRM market share analysis shows salesforce in the lead, growing faster than market. *Forbes* Accessed from `https://www.forbes.com/sites/louiscolumbus/2016/05/28/2015-gartner-crm-market-share-analysis-shows-salesforce-in-the-lead-growing-faster-than-market/`.

Crawford G (1977) Wrsk/blss analysis, the plans-oriented requirements model. *Headquarters Pacific Air Forces/OA, March* .

Crawford GB (1981) Palm's theorem for nonstationary processes. Technical report, RAND CORP SANTA MONICA CA.

Crawley M (2020) How Ontario hospitals are preparing for a surge in COVID-19 cases Accessed from `https://www.cbc.ca/news/canada/toronto/covid-19-coronavirus-ontario-hospitals-emergency-plans-1.5504991`.

Daly R (2020) After patient volume collapsed amid the coronavirus pandemic, some see signs of recovery Accessed from `https://www.hfma.org/topics/news/2020/05/after-patient-volume-collapsed-amid-the-coronavirus--some-see-si.html`.

Dant RP, Berger PD (1996) Modelling cooperative advertising decisions in franchising. *Journal of the operational research society* 47(9):1120–1136.

Dash SR, Mishra US, Mishra P (2013) Emerging issues and opportunities in disaster response supply chain management. *International Journal of Supply Chain Management* 2(1):55–61.

Daw A, Pender J (2019) New perspectives on the erlang-a queue. *Advances in Applied Probability* 51(1):268–299.

de Assunção MD, Cardonha CH, Netto MA, Cunha RL (2016) Impact of user patience on auto-scaling resource capacity for cloud services. *Future Generation Computer Systems* 55:41–50.

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

de Véricourt F, Jennings OB (2011) Nurse staffing in medical units: A queueing perspective. *Operations Research* 59(6):1320–1331.

Dean AM (2002) Service quality in call centres: implications for customer loyalty. *Managing Service Quality: An International Journal* 12(6):414–423.

Defraeye M, Van Nieuwenhuyse I (2013) Controlling excessive waiting times in small service systems with time-varying demand: an extension of the isa algorithm. *Decision Support Systems* 54(4):1558–1567.

Defraeye M, Van Nieuwenhuyse I (2016) Staffing and scheduling under nonstationary demand for service: A literature review. *Omega* 58:4–25.

Ding S, Remerova M, van der Mei RD, Zwart B (2015) Fluid approximation of a call center model with redials and reconnects. *Performance Evaluation* 92:24–39.

Dolinskaya I, Besiou M, Guerrero-Garcia S (2018) Humanitarian medical supply chain in disaster response. *Journal of Humanitarian Logistics and Supply Chain Management* .

Dong J, Whitt W (2015a) Stochastic grey-box modeling of queueing systems: fitting birth-and-death processes to data. *Queueing Systems* 79(3-4):391–426.

Dong J, Whitt W (2015b) Using a birth-and-death process to estimate the steady-state distribution of a periodic queue. *Naval Research Logistics (NRL)* 62(8):664–685.

Dong Y, Yao Y, Cui TH (2011) When acquisition spoils retention: Direct selling vs. delegation under CRM. *Management Science* 57(7):1288–1299.

Dorsch C, Häckel B (2012) Matching economic efficiency and environmental sustainability: The potential of exchanging excess capacity in cloud service environments. *ICIS 2012* Accessed from `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.706.3555&rep=rep1&type=pdf`.

Dorsch C, Häckel B (2014) Combining models of capacity supply to handle volatile demand: The economic impact of surplus capacity in cloud service environments. *Decision Support Systems* 58:3–14.

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

Eick SG, Massey WA, Whitt W (1993b) The physics of the $M_t/G/\infty$ queue. *Operations Research* 41(4):731–742.

Evgeniou T, Fekom M, Ovchinnikov A, Porcher R, Pouchol C, Vayatis N (2020) Epidemic models for personalised COVID-19 isolation and exit policies using clinical risk predictions. *Preprint* Available at SSRN 3588401.

Falin G, Templeton JG (1997) *Retrial queues*, volume 75 (CRC Press).

FDA (2020) Personal protective equipment for infection control Accessed from `https://www.fda.gov/medical-devices/general-hospital-devices-and-supplies/personal-protective-equipment-infection-control`.

Feinberg RA, Hokama L, Kadam R, Kim I (2002) Operational determinants of caller satisfaction in the banking/financial services call center. *International Journal of Bank Marketing* 20(4):174–180.

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

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

Flaxman S, Mishra S, Gandy A, Unwin HJT, Coupland H, Mellan TA, Zhu H, Berah T, Eaton JW, Guzman PN, et al. (2020) Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in european countries: technical description update. *Preprint* Available at arXiv:2004.11342.

Fletcher R (1975) An ideal penalty function for constrained optimization. *IMA Journal of Applied Mathematics* 15(3):319–342.

FLEXERA (2019) Research report, 2019 state of the cloud report: See the latest cloud trends. *FLEXERA* Accessed from `https://info.flexera.com/SLO-CM-WP-State-of-the-Cloud-2019`.

Foley RD (1982) The non-homogeneous $M/G/\infty$ queue. *Opsearch* 19(1):40–48.

Foley RD (1986) Stationary poisson departure processes from non-stationary queues. *Journal of applied probability* 256–260.

Forbes (2015) 30% of servers are sitting ”comatose” according to research. *Forbes* Accessed from `https://www.forbes.com/sites/benkepes/2015/06/03/30-of-servers-are-sitting-comatose-according-to-research/#484c37d759c7`.

Fox ER, Birt A, James KB, Kokko H, Salverson S, Soflin DL (2009) Ashp guidelines on managing drug product shortages in hospitals and health systems. *American Journal of Health-System Pharmacy* 66(15):1399–1406.

Fruchter GE, Zhang ZJ (2004) Dynamic targeted promotions: A customer retention and acquisition perspective. *Journal of Service Research* 7(1):3–19.

Furman E, Diamant A (2020) Optimal capacity planning for cloud service providers with periodic, time-varying demand. *Preprint* Available from at SSRN 3648442.

Furman E, Diamant A, Kristal M (2019) Customer acquisition and retention: A fluid approach for staffing. *Preprint* Available from at SSRN 3422404.

Futamura K (2000) The multiple server effect: Optimal allocation of servers to stations with different service-time distributions in tandem queueing networks. *Annals of Operations Research* 93(1):71–90.

Gans N, Koole G, Mandelbaum A (2003) Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management* 5(2):79–141.

Garg SK, Buyya R (2012) Green cloud computing and environmental sustainability. *Harnessing Green IT: Principles and Practices* 2012:315–340.

Gartner (2019) Gartner forecasts worldwide public cloud revenue to grow 172020. *Gartner* Accessed from `https://www.gartner.com/en/newsroom/press-releases/2019-11-13-gartner-forecasts-worldwide-public-cloud-revenue-to-grow-17-percent-in-2020`.

Gee R, Coates G, Nicholson M (2008) Understanding and profitably managing customer loyalty. *Marketing Intelligence & Planning* 26(4):359–374.

Goldstein T, O'Donoghue B, Setzer S, Baraniuk R (2014) Fast alternating direction optimization methods. *SIAM Journal on Imaging Sciences* 7(3):1588–1623.

Gondi S, Beckman AL, Deveau N, Raja AS, Ranney ML, Popkin R, He S (2020) Personal protective equipment needs in the usa during the COVID-19 pandemic. *Lancet (London, England)* 395(10237):e90.

Gottlieb S, Shu CW (1998) Total variation diminishing runge-kutta schemes. *Mathematics of computation* 67(221):73–85.

Green L, Kolesar P (1991) The pointwise stationary approximation for queues with nonstationary arrivals. *Management Science* 37(1):84–97.

Green L, Kolesar P, Svoronos A (1991) Some effects of nonstationarity on multiserver markovian queueing systems. *Operations Research* 39(3):502–511.

Green LV, Kolesar PJ (1998) A note on approximating peak congestion in $M_t/G/\infty$ queues with sinusoidal arrivals. *Management science* 44(11-part-2):S137–S144.

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

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

Green LV, Nguyen V (2001) Strategies for cutting hospital beds: the impact on patient service. *Health services research* 36(2):421.

Griffin J (2001) Winning customers back. *Business & Economic Review* 48(1):8–8.

Hahn J (2016) Hopf bifurcations in fast/slow systems with rate-dependent tipping. *Preprint* Available from arXiv: 1610.09418.

Halfin S, Whitt W (1981) Heavy-traffic limits for queues with many exponential servers. *Operations research* 29(3):567–588.

Hampshire RC, Jennings OB, Massey WA (2009) A time-varying call center design via lagrangian mechanics. *Probability in the Engineering and Informational Sciences* 23(2):231–259.

Harapan H, Itoh N, Yufika A, Winardi W, Keam S, Te H, Megawati D, Hayati Z, Wagner AL, Mudatsir M (2020) Coronavirus disease 2019 (COVID-19): A literature review. *Journal of Infection and Public Health* .

Harrison JM, Zeevi A (2004) Dynamic scheduling of a multiclass queue in the halfin-whitt heavy traffic regime. *Operations Research* 52(2):243–257.

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

Henderson S, Mason A, Ziedins I, Thomson R (1999) A heuristic for determining efficient staffing requirements for call centres. Technical report, Technical Report, Department of Engineering Science, University of Auckland.

Hillestad RJ, Carrillo MJ (1980) Models and techniques for recoverable item stockage when demand and the repair process are nonstationary. part i. performance measurement. Technical report, RAND CORP SANTA MONICA CA.

Hillier FS, So KC (1989) The assignment of extra servers to stations in tandem queueing systems with small or no buffers. *Performance Evaluation* 10(3):219–231.

Hoyer-Leitzel A, Nadeau A, Roberts A, Steyer A (2017) Detecting transient rate-tipping using steklov averages and lyapunov vectors. *Preprint* Available from arXiv: 1702.02955.

Hu K, Allon G, Bassamboo A (2016) Understanding customers retrial in call centers: Preferences for service quality and service speed. *Preprint* Available from at SSRN 2838998.

Huang J, Carmeli B, Mandelbaum A (2015) Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Operations Research* 63(4):892–908.

Iglehart DL (1965) Limiting diffusion approximations for the many server queue and the repairman problem. *Journal of Applied Probability* 2(2):429–441.

Iglehart DL (1973a) Weak convergence in queueing theory. *Advances in Applied Probability* 5(3):570–594.

Iglehart DL (1973b) Weak convergence of compound stochastic process, I. *Stochastic Processes and their Applications* 1(1):11–31.

Jacobson EU, Argon NT, Ziya S (2012) Priority assignment in emergency response. *Operations Research* 60(4):813–832.

Jain D, Singh SS (2002) Customer lifetime value research in marketing: A review and future directions. *Journal of interactive marketing* 16(2):34–46.

Jang D, Eom J, Park MJ, Rho JJ (2016) Variability of electricity load patterns and its effect on demand response: A critical peak pricing experiment on korean commercial and industrial customers. *Energy Policy* 88:11–26.

Janssen A, van Leeuwaarden JS (2015) Staffing many-server systems with admission control and retrials. *Advances in Applied Probability* 47(2):450–475.

Janssen A, Van Leeuwaarden JS, Zwart B (2011) Refining square-root safety staffing by expanding erlang c. *Operations Research* 59(6):1512–1522.

Jiang Y, Perng Cs, Li T, Chang R (2012) Self-adaptive cloud capacity planning. *2012 IEEE Ninth International Conference on Services Computing*, 73–80 (IEEE).

Jin H, Wang X, Wu S, Di S, Shi X (2014) Towards optimized fine-grained pricing of iaas cloud platform. *IEEE Transactions on cloud Computing* 3(4):436–448.

Joshi KD, Nalwade P (2013) Modified k-means for better initial cluster centres. *International Journal of Computer Science and Mobile Computing* 2(7):219–223.

Jouini O, Koole G, Roubos A (2013) Performance indicators for call centers with impatient customers. *Iie Transactions* 45(3):341–354.

Kalange Pooja R (2013) Applications of green cloud computing in energy efficiency and environmental sustainability. *IOSR Journal of Computer Engineering (IOSR-JCE)* 25–33.

Kang SW, Park HS (2015) Emergency department visit volume variability. *Clinical and experimental emergency medicine* 2(3):150.

Kang W (2015) Fluid limits of many-server retrial queues with nonpersistent customers. *Queueing Systems* 79(2):183–219.

Kang W, Ramanan K, et al. (2010a) Fluid limits of many-server queues with reneging. *The Annals of Applied Probability* 20(6):2204–2260.

Kang W, Ramanan K, et al. (2010b) Fluid limits of many-server queues with reneging. *The Annals of Applied Probability* 20(6):2204–2260.

Karako K, Song P, Chen Y, Tang W (2020) Analysis of COVID-19 infection spread in japan based on stochastic transition model. *Bioscience trends* .

Keaveney SM (1995) Customer switching behavior in service industries: An exploratory study. *The Journal of Marketing* 71–82.

Khazaei H, Misic J, Misic VB (2011) Performance analysis of cloud computing centers using $M/G/m/m+r$ queuing systems. *IEEE Transactions on parallel and distributed systems* 23(5):936–943.

Khintchine AY (1955) Mathematical methods in the theory of queueing. *Khintchine Mathematical Methods in the Theory of Queueing* .

Khudyakov P, Feigin PD, Mandelbaum A (2010) Designing a call center with an IVR (interactive voice response). *Queueing Systems* 66(3):215–237.

King GJ, Chao X, Duenyas I (2016) Dynamic customer acquisition and retention management. *Production and Operations Management* 25(8):1332–1343.

Kitaev MY, Serfozo RF (1999) M/m/1 queues with switching costs and hysteretic optimal control. *Operations Research* 47(2):310–312.

Koole G (1997) Assigning a single server to inhomogeneous queues with switching costs. *Theoretical Computer Science* 182(1-2):203–216.

Koomey JG, et al. (2007) Estimating total power consumption by servers in the us and the world.

Krishnamoorthy A, Ushakumari P (2002) Gi/m/1/1 queue with finite retrials and finite orbits. *stochastic analysis and applications* 20(2):357–374.

Lee N, Kulkarni VG, Hirasawa Y (2014) Optimal static assignment and routing policies for service centers with correlated traffic. *Probability in the Engineering and Informational Sciences* 28(3):279–311.

Li C, Yin W, Jiang H, Zhang Y (2013) An efficient augmented lagrangian method with applications to total variation minimization. *Computational Optimization and Applications* 56(3):507–530.

Liu R, Xie X (2018) Physician staffing for emergency departments with time-varying demand. *INFORMS Journal on Computing* 30(3):588–607.

Liu Y (2018) Staffing to stabilize the tail probability of delay in service systems with time-varying demand. *Operations Research* 66(2):514–534.

Liu Y, Gayle AA, Wilder-Smith A, Rocklov J (2020) The reproductive number of COVID-19 is higher compared to sars coronavirus. *Journal of travel medicine* .

Liu Y, Whitt W (2017) Stabilizing performance in a service system with time-varying arrivals and customer feedback. *European Journal of Operational Research* 256(2):473–486.

Livingston E, Desai A, Berkwits M (2020) Sourcing personal protective equipment during the COVID-19 pandemic. *Jama* 323(19):1912–1914.

Lu Y, Musalem A, Olivares M, Schilkrut A (2013) Measuring the effect of queues on customer purchases. *Management Science* 59(8):1743–1763.

Malhotra NK (2007) Review of marketing research. *Review of Marketing Research* (Emerald Group Publishing Limited).

Mandelbaum A, Massey WA, Reiman MI (1998) Strong approximations for markovian service networks. *Queueing Systems* 30(1-2):149–201.

Massey WA, Pender J (2018) Dynamic rate erlang-a queues. *Queueing Systems* 89(1-2):127–164.

Massey WA, Whitt W (1993) Networks of infinite-server queues with nonstationary poisson input. *Queueing Systems* 13(1-3):183–250.

Mattingley J, Boyd S (2010) Real-time convex optimization in signal processing. *IEEE Signal processing magazine* 27(3):50–61.

McInnes L, Healy J, Melville J (2018) UMAP: Uniform manifold approximation and projection for dimension reduction. *Preprint* Available from arXiv: 1802.03426.

Mehrotra P, Malani P, Yadav P (2020) Personal protective equipment shortages during COVID-19 – supply chain–related causes and mitigation strategies. *JAMA Health Forum*, volume 1, e200553–e200553 (American Medical Association).

Milner JM, Rosenblatt MJ (2002) Flexible supply contracts for short life-cycle goods: The buyer's perspective. *Naval Research Logistics (NRL)* 49(1):25–45.

Milovic B (2012) Application of customer relationship management strategy (CRM) in different business areas. *Facta Universitatis Series Economics and Organization* 9(3):341–354.

Niyirora J, Pender J (2016) Optimal staffing in nonstationary service centers with constraints. *Naval Research Logistics (NRL)* 63(8):615–630.

Ontario Health (2020a) COVID-19 – what we know so far about... routes of transmission Accessed from `https://www.publichealthontario.ca/-/media/documents/ncov/wwksf-routes-transmission-mar-06-2020.pdf?la=en`.

Ontario Health (2020b) Public Health Ontario. technical brief: IPAC recommendations for use of Personal Protective Equipment for care of individuals with suspect or confirmed COVID-19. Accessed from `https://www.publichealthontario.ca/-/media/documents/ncov/updated-ipac-measures-covid-19.pdf?la=en`.

Oracle (2011) Customer experience impact report. Accessed from `http://www.oracle.com/us/products/applications/cust-exp-impact-report-epss-1560493.pdf`.

Palm C (1943) Variation in intensity in telephone conversation. *Ericsson Technics* 4:1–189.

Pan Y, Maini S, Blevis E (2010) Framing the issues of cloud computing & sustainability: A design perspective. *2010 IEEE Second International Conference on Cloud Computing Technology and Science*, 603–608 (IEEE).

Park PS, Bobrowski PM (1989) Job release and labor flexibility in a dual resource constrained job shop. *Journal of operations management* 8(3):230–249.

Parsons LJ (1975) The product life cycle and time-varying advertising elasticities. *Journal of Marketing Research* 12(4):476–480.

Pasricha SV, Jung HY, Kushnir V, Mak D, Koppula R, Guo Y, Kwan JL, Lapointe-Shaw L, Rawal S, Tang T, et al. (2020) Assessing the quality of clinical and administrative data extracted from hospitals: The General Medicine Inpatient Initiative (GEMINI) experience. *medRxiv* .

Pender J (2016) Risk measures and their application to staffing nonstationary service systems. *European Journal of Operational Research* 254(1):113–126.

Pender J, Rand RH, Wesson E (2017) Queues with choice via delay differential equations. *International Journal of Bifurcation and Chaos* 27(04):1730016.

Public Services and Procurement Canada (2020) Supplying the canadian healthcare sector in response to COVID-19 Accessed from `https://www.tpsgc-pwgsc.gc.ca/comm/aic-scr/provisions-supplies-eng.html`.

Pustova S (2010) Investigation of call centers as retrial queuing systems. *Cybernetics and Systems Analysis* 46(3):494–499.

Qi X, Bard JF, Yu G (2004) Supply chain coordination with demand disruptions. *Omega* 32(4):301–312.

Ranney ML, Griffeth V, Jha AK (2020) Critical supply shortages—the need for ventilators and personal protective equipment during the COVID-19 pandemic. *New England Journal of Medicine* 382(18):e41.

Razak F, Shin S, Pogacar F, Jung HY, Pus L, Moser A, Lapointe-Shaw L, Tang T, Kwan JL, Weinerman A, et al. (2020) Modelling resource requirements and physician staffing to provide virtual urgent medical care for residents of long-term care homes: a cross-sectional study. *CMAJ open* 8(3):E514–E521.

Reichheld FF, Sasser JW (1990) Zero defections: Quality comes to services. *Harvard business review* 68(5):105–111.

Reichheld FF, Schefter P (2000) E-loyalty: your secret weapon on the web. *Harvard business review* 78(4):105–113.

Reinartz W, Thomas JS, Kumar V (2005) Balancing acquisition and retention resources to maximize customer profitability. *Journal of marketing* 69(1):63–79.

Reinsel D, Gantz J, Rydning J (2018) The digitization of the world from edge to core. *IDC White Paper* .

Robbins TR, Harrison TP (2010) A stochastic programming model for scheduling call centers with global service level agreements. *European Journal of Operational Research* 207(3):1608–1619.

Rowan NJ, Laffey JG (2020) Challenges and solutions for addressing critical shortage of supply chain for personal and protective equipment (PPE) arising from coronavirus disease (COVID-19) pandemic–case study from the republic of ireland. *Science of The Total Environment* 138532.

Ryu D, Ostriker JP, Kang H, Cen R (1993) A cosmological hydrodynamic code based on the total variation diminishing scheme .

Sahin AR, Erdogan A, Agaoglu PM, Dineri Y, Cakirci AY, Senel ME, Okyay RA, Tasdogan AM (2020) 2019 novel coronavirus (COVID-19) outbreak: a review of the current literature. *EJMO* 4(1):1–7.

Shu CW (1988) Total-variation-diminishing time discretizations. *SIAM Journal on Scientific and Statistical Computing* 9(6):1073–1084.

Simha A, Prasad RV, Narayana S (2020) A simple stochastic SIR model for COVID-19 infection dynamics for karnataka: Learning from europe. *Preprint* Available from arXiv: 2003.11920.

Smith JM, Barnes R (2015) Optimal server allocation in closed finite queueing networks. *Flexible Services and Manufacturing Journal* 27(1):58–85.

Smith JM, Cruz F, van Woensel T (2010) Optimal server allocation in general, finite, multi-server queueing networks. *Applied Stochastic Models in Business and Industry* 26(6):705–736.

Soh SB, Gurvich I (2017) Call center staffing: Service-level constraints and index priorities. *Operations Research* 65(2):537–555.

Stecke KE, Kumar S (2009) Sources of supply chain disruptions, factors that breed vulnerability, and mitigating strategies. *Journal of Marketing Channels* 16(3):193–226.

Sung KW, Chae KC (2000) An approximate analysis of the M/M/1/1 queue with finite number of retrials. *Journal of Korean Institute of Industrial Engineers* 26(3):206–212.

Tang CS (2006) Robust strategies for mitigating supply chain disruptions. *International Journal of Logistics: Research and Applications* 9(1):33–45.

Taylor M (2018) 18 CRM statistics you need to know for 2018. Accessed from `https://www.superoffice.com/blog/crm-software-statistics/`.

Thomas JS (2001) A methodology for linking customer acquisition to customer retention. *Journal of Marketing Research* 38(2):262–268.

Thompson GL, Teng JT (1984) Optimal pricing and advertising policies for new product oligopoly models. *Marketing Science* 3(2):148–168.

Toda AA (2020) Susceptible-infected-recovered (SIR) dynamics of COVID-19 and economic impact. *Preprint* Available from arXiv: 2003.11221.

Tsai YC, Argon NT (2008) Dynamic server assignment policies for assembly-type queues with flexible servers. *Naval Research Logistics (NRL)* 55(3):234–251.

Tuite AR, Fisman DN, Greer AL (2020) Mathematical modelling of COVID-19 transmission and mitigation strategies in the population of Ontario, Canada. *CMAJ* 192(19):E497–E505.

Uchechukwu A, Li K, Shen Y (2014) Energy consumption in cloud computing data centers. *International Journal of Cloud Computing and Services Science (IJ-CLOSER)* 3(3):31–48.

UHN (2020) Service reintroduction at uhn Accessed from `https://www.uhn.ca/covid19#arriving`.

University of Wisconsin (2020) Online platform helps healthcare facilities in need of face shields meet their match Accessed from `https://www.engr.wisc.edu/news/online-platform-helps-healthcare-facilities-in-need-of-face-shields-meet-their-match/`.

Verdouw C, Beulens A, Trienekens J, Van der Vorst J (2011) A framework for modelling business processes in demand-driven supply chains. *Production Planning & Control* 22(4):365–388.

Verma AA, Guo Y, Kwan JL, Lapointe-Shaw L, Rawal S, Tang T, Weinerman A, Cram P, Dhalla IA, Hwang SW, et al. (2017) Patient characteristics, resource use and outcomes associated with general internal medicine hospital care: the General Medicine Inpatient Initiative (GEMINI) retrospective cohort study. *CMAJ open* 5(4):E842.

Verma AA, Guo Y, Kwan JL, Lapointe-Shaw L, Rawal S, Tang T, Weinerman A, Razak F (2018) Prevalence and costs of discharge diagnoses in inpatient general internal medicine: a multi-center cross-sectional study. *Journal of general internal medicine* 33(11):1899–1904.

Vilaplana J, Solsona F, Teixidó I, Mateo J, Abella F, Rius J (2014) A queuing theory model for cloud computing. *The Journal of Supercomputing* 69(1):492–507.

Viswanathan J, Grossmann IE (1990) A combined penalty function and outer-approximation method for minlp optimization. *Computers & Chemical Engineering* 14(7):769–782.

Vizarreta P, Trivedi K, Helvik B, Heegaard P, Blenk A, Kellerer W, Machuca CM (2018) Assessing the maturity of sdn controllers with software reliability growth models. *IEEE Transactions on Network and Service Management* 15(3):1090–1104.

Wang H, Jing Q, He B, Qian Z, Zhou L (2010) Distributed systems meet economics: pricing in the cloud .

Wangping J, Ke H, Yang S, Wenzhe C, Shengshu W, Shanshan Y, Jianwei W, Fuyin K, Penggang T, Jing L, et al. (2020) Extended SIR prediction of the epidemics trend of COVID-19 in italy and compared with hunan, china. *Frontiers in medicine* 7:169.

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

Whitt W (2014) The steady-state distribution of the mt/m/∞ queue with a sinusoidal arrival rate function. *Operations Research Letters* 42(5):311–318.

Whitt W (2015) Many-server limits for periodic infinite-server queues. *Columbia University* .

Whitt W (2016) Queues with time-varying arrival rates: A bibliography. Technical report, Working paper.

Whitt W (2018) Time-varying queues. *Queueing models and service management* 1(2).

Whitt W, Zhang X (2017) A data-driven model of an emergency department. *Operations Research for Health Care* 12:1–15.

WHO (2020) Coronavirus disease (COVID-19) advice for the public: When and how to use masks Accessed from https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/when-and-how-to-use-masks.

Wilkinson RI (1956) Theories for toll traffic engineering in the usa. *Bell System Technical Journal* 35(2):421–514.

Workman AD, Welling DB, Carter BS, Curry WT, Holbrook EH, Gray ST, Scangas GA, Bleier BS (2020) Endonasal instrumentation and aerosolization risk in the era of COVID-19: simulation, literature review, and proposed mitigation strategies. *International forum of allergy & rhinology* (Wiley Online Library).

Xu M, Qi X, Yu G, Zhang H, Gao C (2003) The demand disruption management problem for a supply chain system with nonlinear demand functions. *Journal of Systems Science and Systems Engineering* 12(1):82–97.

Yeniay Ö (2005) Penalty function methods for constrained optimization with genetic algorithms. *Mathematical and computational Applications* 10(1):45–56.

Yeung JHY, Selen W, Deming Z, Min Z (2007) Postponement strategy from a supply chain perspective: cases from china. *International Journal of Physical Distribution & Logistics Management* .

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

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

Zeltyn S, Marmor YN, Mandelbaum A, Carmeli B, Greenshpan O, Mesika Y, Wasserkrug S, Vortman P, Shtub A, Lauterman T, et al. (2011) Simulation-based models of emergency departments: Operational, tactical, and strategic staffing. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 21(4):1–25.

Zhang AZ, Enns E (2020) Optimal timing and effectiveness of COVID-19 outbreak responses in china: a modelling study. *Preprint* Available at SSRN.

Zhang Y, You C, Cai Z, Sun J, Hu W, Zhou XH (2020) Prediction of the COVID-19 outbreak based on a realistic stochastic model. *medRxiv* .

Zhao XL, Wang F, Ng MK (2014) A new convex optimization model for multiplicative noise and blur removal. *SIAM Journal on Imaging Sciences* 7(1):456–475.

Zhu L, Lee L, Ma Y, Ye Y, Mazzeo R, Xing L (2008) Using total-variation regularization for intensity modulated radiation therapy inverse planning with field-specific numbers of segments. *Physics in Medicine & Biology* 53(23):6653.

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

Appendix A

Chapter 1: Implementation Details and Proofs of Statements

State Space

We define a piecewise smooth system (PSS) using a finite set of ODEs as follows

$$\dot{\boldsymbol{q}} = f_i(\boldsymbol{q}, t), \quad \boldsymbol{q}_0 = (q_a, q_b, q_c)^T, \quad \boldsymbol{q} \in \mathcal{S}_i \subset \mathbb{R}^3_{\geq 0}, \tag{A.1}$$

where $\mathcal{S}_i$, $i = 1, 2, 3, 4$, are open non-overlapping regions separated by two-dimensional boundaries (planes). Planes defined by $q_a(t) = s_a$ and $q_b(t) = s_b$ partition $\mathbb{R}^3_{\geq 0}$ into four subsets with the values of right-hand side (RHS) $f_i(\boldsymbol{q}, t)$, changing in each $\mathcal{S}_i$. Define $\mathcal{S}_i$

$\mathcal{S}_1 = \{q_a(t), q_b(t), q_c(t) \in \mathbb{R}_{\geq 0} : q_a(t) > s_a, q_b(t) > s_b\}$, $\mathcal{S}_2 = \{q_a(t), q_b(t), q_c(t) \in \mathbb{R}_{\geq 0} : q_a(t) < s_a, q_b(t) > s_b\}$,

$\mathcal{S}_3 = \{q_a(t), q_b(t), q_c(t) \in \mathbb{R}_{\geq 0} : q_a(t) < s_a, q_b(t) < s_b\}$, $\mathcal{S}_4 = \{q_a(t), q_b(t), q_c(t) \in \mathbb{R}_{\geq 0} : q_a(t) > s_a, q_b(t) < s_b\}$.

Observe that $\bigcup_{i=1}^{4} \mathcal{S}_i$ does not include the boundaries among $\mathcal{S}_i$. Hence, the subset of boundaries, $\Sigma = \bigcup_{j,k=1,j\neq k}^{4} \Sigma_{jk}$, can be defined as (with $\bar{\mathcal{S}}_i$ denoting the closure of $\mathcal{S}_i$)

$$\Sigma_{12} = \left(\bar{\mathcal{S}}_1 \cap \bar{\mathcal{S}}_2\right) \setminus \Sigma_{13} = \{q_a(t) = s_a, q_b(t) > s_b\}, \quad \Sigma_{13} = \bar{\mathcal{S}}_1 \cap \bar{\mathcal{S}}_3 = \{q_a(t) = s_a, q_b(t) = s_b\},$$

$$\Sigma_{14} = \left(\bar{\mathcal{S}}_1 \cap \bar{\mathcal{S}}_4\right) \setminus \Sigma_{13} = \{q_a(t) > s_a, q_b(t) = s_b\}, \quad \Sigma_{23} = \left(\bar{\mathcal{S}}_2 \cap \bar{\mathcal{S}}_3\right) \setminus \Sigma_{13} = \{q_a(t) < s_a, q_b(t) = s_b\},$$

$$\Sigma_{24} = \bar{\mathcal{S}}_2 \cap \bar{\mathcal{S}}_4 = \bar{\mathcal{S}}_1 \cap \bar{\mathcal{S}}_3 = \Sigma_{13}, \quad \Sigma_{34} = \left(\bar{\mathcal{S}}_3 \cap \bar{\mathcal{S}}_4\right) \setminus \Sigma_{13} = \{q_a(t) = s_a, q_b(t) < s_b\}.$$

Simulation Parameters

Table A.1: Queueing Parameters

	$\tau_a = \tau_b$	$\mu_a = \mu_b$	r
minimum (low)	0.5	1.2	0.5
average	2.5	2	2.9
maximum (high)	7	4.2	8

Table A.2: Advertising Campaign

	λ_0	λ_1	δ
fast	3	22	0.6
fast	3	22	0.4
intermediate	3	22	0.2
slow	3	22	0.1
slow	3	22	0.01

Table A.1 describes the range of queueing parameters (minimum, average, and maximum) that we use for the simulations of the advertising campaign and clinical setting scenarios. Table A.2 presents the arrival function parameters associated with (2.10) used in the advertising campaign experiments. We consider cases where the arrival function approaches its limit quickly in (A.2), with an intermediate speed in (A.3) and slowly in (A.4). That is,

$$\lambda_{1,exp}(t) = 22e^{-0.6t} + 3, \quad \lambda_{2,exp}(t) = 22e^{-0.4t} + 3, \tag{A.2}$$

$$\lambda_{3,exp}(t) = 22e^{-0.2t} + 3, \tag{A.3}$$

$$\lambda_{4,exp}(t) = 22e^{-0.1t} + 3, \quad \lambda_{5,exp}(t) = 22e^{-0.01t} + 3. \tag{A.4}$$

Table A.3: Clinical Setting

	Fast						Intermediate						Slow							
$\bar{\lambda}$	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
σ	0.9	0.7	0.5	0.9	0.7	0.9	0.3	0.5	0.7	0.5	0.9	0.7	0.1	0.3	0.1	0.3	0.1	0.5	0.3	0.1
Period	2	2	2	4	4	6	2	4	6	6	8	8	2	4	4	6	6	8	8	8

Table A.3 presents the arrival function parameters associated with (3.12) used in the clinical setting experiments. The arrival function approaches an asymptotic periodic curve quickly (6 scenarios), with an intermediate speed (6 scenarios), and slowly (8 scenarios). For example, in the first column of Table A.3, the average of $\lambda_{1,sine}(t)$ is equal to 7, its relative amplitude is 0.9, and its period is 2. Thus, the arrival function oscillates quickly with amplitude equal to 90% of its average value.

For EXP and the corresponding benchmarks SRS(0) and SRS(1), we conduct 90 simulations (3 staffing policies; low and high abandonment, service time and the arrival of base clients; 5 variants of the arrival function, i.e., $3 \cdot (2 + 2 + 2) \cdot 5 = 90$). We repeat the simulations for short and long time horizons, $t = 10$ and $t = 25$, respectively, which results in 180 simulations in total. Similarly, for the staffing policy in SINE, we conduct 360 simulations ($3 \cdot (2 + 2 + 2) \cdot 20 = 360$). We repeat the simulations for $t = 10$ and $t = 25$, which results in 720 simulations in total. Note that additional 1440 simulations were performed to investigate managerially relevant parameter settings; for simplicity, we have omitted the results and only show the most important regimes.

Summary of Results for EXP, $s = 20$

Each row of Tables 4-7 displays the performance of EXP, $GSRS_0$, $GSRS_t$ as compared to the stochastic optimal policy (OPT). The results are presented in groups of four: for $\delta = 0.6$ (Column 2-5), $\delta = 0.2$ (Column 6-9) and $\delta = 0.01$ (Column 10-13). Table 4 and 6 display results for the Erlang-C version of $GSRS_0$ and $GSRS_t$. Conversely, Table 5 and 7 display results for the Erlang-A version of $GSRS_0$ and $GSRS_t$.

Table 8 displays how many candidate solutions were considered in order to obtain the optimal solution to EXP with the length of the time horizon and the value of δ fixed, Row 1 and 2, respectively.

OPT is constructed by simulating all valid staffing levels in the stochastic regime and choosing the best performing policy according to the metric of interest (total served, total abandonment, etc).

Bolded text indicates which staffing policy is closest to OPT.

Table A.4 Total Served, Erlang-C version of $GSRS_0$ and $GSRS_t$, Time Horizon = 25

δ	0.6				0.2				0.01			
Policy	EXP	$GSRS_0$	$GSRS_t$	OPT	EXP	$GSRS_0$	$GSRS_t$	OPT	EXP	$GSRS_0$	$GSRS_t$	OPT
Fast servers	**284.8**	196.6	187.1	289.8	**425.2**	362.4	253.4	425.2	**733.8**	714.9	702.2	735.6
Slow servers	**235.3**	231.1	191.2	257.1	**351.8**	345.5	250.4	352.3	490.1	**514.8**	507.9	526.9
Impatient clients	**260**	174.8	190.7	272	347.7	**350.4**	236.3	380	616.5	619.1	**621.4**	621
Patient clients	**286.2**	220.8	229.5	301	365.1	**392.6**	282.4	406.2	**664.3**	661.3	657.9	664.3
High demand of base clients	**323.9**	296.1	289.1	345.4	374	**414**	351.4	414.3	619.6	619.6	**632.5**	633.7
Low demand of base clients	**145.8**	97.2	104	152.5	**264.1**	195.8	142.2	273.9	**615.4**	613.4	613.5	621.5
average	**256.0000**	202.7667	198.6000	269.6333	**354.6500**	343.4500	252.6833	375.3167	623.2833	**623.8500**	622.5667	633.8333

Table A.5 Total Served, Erlang-A version of $GSRS_0$ and $GSRS_t$, Time Horizon = 25

δ	0.6				0.2				0.01			
Policy	EXP	$GSRS_0$	$GSRS_t$	OPT	EXP	$GSRS_0$	$GSRS_t$	OPT	EXP	$GSRS_0$	$GSRS_t$	OPT
Fast servers	**284.8**	196.6	181.4	289.8	**425.2**	240.7	263.7	425.2	**733.8**	580.2	596	735.6
Slow servers	**235.3**	191.2	152.6	257.1	**351.8**	335.2	250.3	352.3	490.1	490.1	**491.9**	526.9
Impatient clients	**260**	115	131.6	272	**347.7**	137.4	140.3	380	**616.5**	385.1	376.6	621
Patient clients	**286.2**	220.8	236.7	301	365.1	**392.6**	283.6	406.2	**664.3**	661.3	651.7	664.3
High demand of base clients	**323.9**	296.1	187.7	345.4	374	**394.9**	342.4	414.3	**619.6**	568.7	565.4	633.7
Low demand of base clients	**145.8**	97.2	64.7	152.5	**264.1**	157.3	122.1	273.9	**615.4**	487.5	474.4	621.5
average	**256.0000**	186.1500	159.1167	269.6333	**354.6500**	276.3500	233.7333	375.3167	**623.2833**	528.8167	526.0000	633.8333

Table A.6 Abandonment Ratio, Erlang-C version of $GSRS_0$ and $GSRS_t$, Time Horizon = 25

δ	0.6				0.2				0.01			
Policy	EXP	$GSRS_0$	$GSRS_t$	OPT	EXP	$GSRS_0$	$GSRS_t$	OPT	EXP	$GSRS_0$	$GSRS_t$	OPT
Fast servers	**0.0475**	0.2053	0.2142	0.0453	**0.1552**	0.1734	0.3138	0.1384	**0.3602**	0.3628	0.3662	0.3602
Slow servers	0.1330	**0.1179**	0.1915	0.0772	**0.1940**	0.2067	0.3236	0.1940	**0.4787**	0.5113	0.5044	0.4770
Impatient clients	**0.0934**	0.2360	0.2031	0.0770	**0.2545**	0.2029	0.3465	0.1810	0.4361	0.4348	**0.4299**	0.4348
Patient clients	**0.0536**	0.1345	0.1197	0.0402	0.2331	**0.1487**	0.2650	0.1484	0.3786	**0.3720**	0.3801	0.3715
High demand of base clients	0.1706	**0.1380**	0.1406	0.1169	0.3482	**0.2620**	0.2731	0.2615	0.5146	0.5146	**0.5059**	0.4946
Low demand of base clients	**0.0402**	0.3116	0.2930	0.0060	**0.0344**	0.2530	0.4314	0.0235	**0.1516**	0.1544	0.1565	0.1503
average	**0.0897**	0.1906	0.1937	0.0604	**0.2032**	0.2078	0.3256	0.1578	**0.3867**	0.3917	0.3905	0.3814

Table A.7 Abandonment Ratio, Erlang-A version of $GSRS_0$ and $GSRS_t$, Time Horizon = 25

δ	0.6				0.2				0.01			
Policy	EXP	$GSRS_0$	$GSRS_t$	OPT	EXP	$GSRS_0$	$GSRS_t$	OPT	EXP	$GSRS_0$	$GSRS_t$	OPT
Fast servers	**0.0475**	0.2053	0.2254	0.0453	**0.1552**	0.3301	0.3013	0.1384	**0.3602**	0.3945	0.3899	0.3602
Slow servers	**0.1330**	0.1960	0.2869	0.0772	**0.1940**	0.2113	0.3086	0.1940	**0.4787**	**0.4787**	0.4800	0.4770
Impatient clients	**0.0934**	0.4029	0.3627	0.0770	**0.2545**	0.5350	0.5267	0.1810	**0.4361**	0.5171	0.5230	0.4348
Patient clients	**0.0536**	0.1345	0.1214	0.0402	0.2331	**0.1487**	0.2638	0.1484	0.3786	**0.3720**	0.3834	0.3715
High demand of base clients	0.1706	**0.1380**	0.2850	0.1169	0.3482	**0.2615**	0.2746	0.2615	0.5146	**0.4963**	0.5029	0.4946
Low demand of base clients	**0.0402**	0.3116	0.5204	0.0060	**0.0344**	0.3650	0.4872	0.0235	**0.1516**	0.2849	0.2996	0.1503
average	**0.0897**	0.2314	0.3003	0.0604	**0.2032**	0.3086	0.3603	0.1578	**0.3867**	0.4239	0.4298	0.3814

Table A.8 Number of iterations to find optimal staffing level.

	Time Horizon = 10					Time Horizon = 25				
δ	0.6	0.4	0.2	0.1	0.001	0.6	0.4	0.2	0.1	0.001
Fast servers	1	1	1	1	1	1	1	1	1	1
Slow servers	1	6	7	4	1	1	6	7	7	1
Impatient clients	1	1	1	6	2	1	3	1	6	3
Patient clients	1	1	5	6	2	1	1	1	6	3
High demand of base clients	4	6	7	7	1	3	6	7	8	2
Low demand of base clients	1	1	1	1	1	1	1	1	1	1

Summary of Results for SINE, $T = 2$, $s = 20$

Each row of Tables 9-12 displays the performance of SINE, $GSRS_0$, $GSRS_t$ as compared to the stochastic optimal policy (OPT). The results are presented in groups of four: for $\delta = 0.6$ (Column 2-5), $\delta = 0.2$ (Column 6-9) and $\delta = 0.01$ (Column 10-13). Table 9 and 11 display results for the Erlang-C version of $GSRS_0$ and $GSRS_t$. Conversely, Table 10 and 12 display results for the Erlang-A version of $GSRS_0$ and $GSRS_t$.

Table 13 displays how many candidate solutions were considered in order to obtain the optimal solution to SINE with the length of the time horizon and the value of δ fixed, Row 1 and 2, respectively.

OPT is constructed by simulating all valid staffing levels in the stochastic regime and choosing the best performing policy according to the metric of interest (total served, total abandonment, etc).

Bolded text indicates which staffing policy is closest to OPT.

Table A.9 Total Served, Erlang-C version of $GSRS_0$ and $GSRS_t$, Time Horizon = 25

σ	0.9				0.5				0.1			
Policy	SINE	$GSRS_0$	$GSRS_t$	OPT	SINE	$GSRS_0$	$GSRS_t$	OPT	SINE	$GSRS_0$	$GSRS_t$	OPT
Fast servers	**610.1**	461.9	508.2	738.6	**499.4**	**499.4**	415.4	726	**496.2**	**496.2**	483.8	708.5
Slow servers	**572**	472.8	538.4	631.8	**525.8**	468.3	529.1	615.5	**467.1**	**467.1**	458.3	609.8
Impatient clients	**517.2**	457.4	441.7	670.3	460.7	460.7	**468.2**	672.7	**451.9**	**451.9**	438.9	643
Patient clients	**640.8**	547.2	619.3	731	556.8	556.8	**563.4**	712.5	**533.1**	**533.1**	516.8	684.1
High demand of base clients	**847**	747.7	756.5	847	**794.6**	**794.6**	745.4	868.6	**749.7**	**749.7**	733.3	845.7
Low demand of base clients	**273.1**	237.1	252.3	373.4	**251.5**	**251.5**	243.6	359.7	**238.5**	**238.5**	218.1	348
average	**576.7**	487.35	519.4	665.35	**514.8**	505.2167	494.1833	659.1667	**489.4167**	**489.4167**	474.8667	639.85

Table A.10 Total Served, Erlang-A version of $GSRS_0$ and $GSRS_t$, Time Horizon = 25

σ	0.9				0.5				0.1			
Policy	SINE	$GSRS_0$	$GSRS_t$	OPT	SINE	$GSRS_0$	$GSRS_t$	OPT	SINE	$GSRS_0$	$GSRS_t$	OPT
Fast servers	**610.1**	283.6	393.5	738.6	**499.4**	296.4	429.9	726	**496.2**	283.9	305.9	708.5
Slow servers	**572**	398.5	383.1	631.8	**525.8**	423.9	426.3	615.5	**467.1**	337.9	421.7	609.8
Impatient clients	**517.2**	146.5	204.5	670.3	**460.7**	142.7	251.3	672.7	**451.9**	146.8	152.9	643
Patient clients	**640.8**	547.2	610.6	731	556.8	556.8	**581**	712.5	533.1	533.1	**541.1**	684.1
High demand of base clients	**847**	623	608.7	847	**794.6**	625.6	566.9	868.6	**749.7**	616.4	662.6	845.7
Low demand of base clients	**273.1**	192.7	193.4	373.4	**251.5**	205.1	178	359.7	**238.5**	196	197.9	348
average	**576.7**	365.25	398.9667	665.35	**514.8**	375.0833	405.5667	659.1667	**489.4167**	352.35	380.35	639.85

Table A.11 Abandonment Ratio, Erlang-C version of $GSRS_0$ and $GSRS_t$, Time Horizon = 25

σ	0.9				0.5				0.1			
Policy	SINE	$GSRS_0$	$GSRS_t$	OPT	SINE	$GSRS_0$	$GSRS_t$	OPT	SINE	$GSRS_0$	$GSRS_t$	OPT
Fast servers	**0.1009**	0.2006	0.1755	0.0229	**0.1547**	**0.1547**	0.2315	0.0193	**0.1448**	**0.1448**	0.1543	0.0118
Slow servers	**0.1119**	0.1869	0.1970	0.0961	**0.1261**	0.1716	0.1649	0.0852	**0.1653**	**0.1653**	0.1784	0.0663
Impatient clients	**0.1586**	0.2134	0.2318	0.0694	**0.1926**	**0.1926**	0.1994	0.0511	**0.1858**	**0.1858**	0.2016	0.0506
Patient clients	**0.0622**	0.1136	0.1331	0.0239	**0.0929**	**0.0929**	0.1339	0.0169	**0.0910**	**0.0910**	0.1273	0.0153
High demand of base clients	**0.1139**	0.1343	0.1647	0.1139	**0.1085**	**0.1085**	0.1350	0.1021	**0.1100**	**0.1100**	0.1246	0.1005
Low demand of base clients	**0.2184**	0.3026	0.3288	0.0046	**0.2466**	**0.2466**	0.3087	0.0014	**0.2577**	**0.2577**	0.3159	0.0023
average	**0.1277**	0.1919	0.2052	0.0551	**0.1536**	0.1611	0.1956	0.0460	**0.1591**	**0.1591**	0.1837	0.0411

Table A.12 Abandonment Ratio, Erlang-A version of $GSRS_0$ and $GSRS_t$, Time Horizon = 25

σ	0.9				0.5				0.1			
Policy	SINE	$GSRS_0$	$GSRS_t$	OPT	SINE	$GSRS_0$	$GSRS_t$	OPT	SINE	$GSRS_0$	$GSRS_t$	OPT
Fast servers	**0.1009**	0.3773	0.2591	0.0229	**0.1547**	0.3451	0.2155	0.0193	**0.1448**	0.3496	0.3231	0.0118
Slow servers	**0.1119**	0.2428	0.2892	0.0961	**0.1261**	0.2096	0.2329	0.0852	**0.1653**	0.2823	0.2162	0.0663
Impatient clients	**0.1586**	0.6035	0.5034	0.0694	**0.1926**	0.5977	0.4259	0.0511	**0.1858**	0.5832	0.5698	0.0506
Patient clients	**0.0622**	0.1136	0.1193	0.0239	**0.0929**	**0.0929**	0.1074	0.0169	**0.0910**	**0.0910**	0.1118	0.0153
High demand of base clients	**0.1139**	0.1819	0.2025	0.1139	**0.1085**	0.1629	0.2038	0.1021	**0.1100**	0.1595	0.1429	0.1005
Low demand of base clients	**0.2184**	0.4107	0.4350	0.0046	**0.2466**	0.3597	0.4425	0.0014	**0.2577**	0.3591	0.3602	0.0023
average	**0.1277**	0.3216	0.3014	0.0551	**0.1536**	0.2946	0.2713	0.0460	**0.1591**	0.3041	0.2873	0.0411

Table A.13 Number of iterations to find optimal staffing level.

	Time Horizon = 10					Time Horizon = 25				
σ	0.9	0.7	0.5	0.3	0.1	0.9	0.7	0.5	0.3	0.1
Fast servers	1	1	1	1	1	1	1	1	1	1
Slow servers	1	1	1	1	1	1	1	1	1	1
Impatient clients	1	1	1	1	1	1	1	1	1	1
Patient clients	1	1	1	1	1	1	1	1	1	1
High demand of base clients	1	1	1	1	1	1	1	1	1	1
Low demand of base clients	1	1	1	1	1	1	1	1	1	1

Summary of Results for SINE, $T = 8$, $s = 20$

Each row of Tables 14-17 displays the performance of SINE, $GSRS_0$, $GSRS_t$ as compared to the stochastic optimal policy (OPT). The results are presented in groups of four: for δ = 0.6 (Column 2-5), δ = 0.2 (Column 6-9) and δ = 0.01 (Column 10-13). Table 14 and 16 display results for the Erlang-C version of $GSRS_0$ and $GSRS_t$. Conversely, Table 15 and 17 display results for the Erlang-A version of $GSRS_0$ and $GSRS_t$.

Table 18 displays how many candidate solutions were considered in order to obtain the optimal solution to SINE with the length of the time horizon and the value of δ fixed, Row 1 and 2, respectively.

OPT is constructed by simulating all valid staffing levels in the stochastic regime and choosing the best performing policy according to the metric of interest (total served, total abandonment, etc).

Bolded text indicates which staffing policy is closest to OPT.

Table A.14 Total Served, Erlang-C version of $GSRS_0$ and $GSRS_t$, Time Horizon = 25

σ	0.9				0.5				0.1			
Policy	SINE	$GSRS_0$	$GSRS_t$	OPT	SINE	$GSRS_0$	$GSRS_t$	OPT	SINE	$GSRS_0$	$GSRS_t$	OPT
Fast servers	**572.3**	478.3	310.8	750.2	**492.1**	**492.1**	316.6	757.1	485.5	485.5	**507.8**	693.2
Slow servers	**551.4**	438.7	224.8	578.9	**526.7**	509.1	398.5	633.4	**457.9**	**457.9**	427.5	595.2
Impatient clients	**568.6**	425.8	183.3	636.3	**559.4**	520	314.7	688.3	**472.2**	**472.2**	403.4	648.6
Patient clients	**648**	521.5	234.8	724.6	**645.4**	572.4	359.5	746.3	**535.8**	**535.8**	447.8	681.1
High demand of base clients	**793**	705.2	321.5	795.1	**858.5**	826.3	519.1	867	**798.4**	**798.4**	644.4	835.9
Low demand of base clients	**290.5**	228.8	98.5	399.7	**298.9**	260.7	171.5	386.6	**247.7**	**247.7**	202.8	338
average	**570.6333**	466.3833	228.95	647.4667	**563.5**	530.1	346.65	679.7833	**499.5833**	**499.5833**	438.95	632

Table A.15 Total Served, Erlang-A version of $GSRS_0$ and $GSRS_t$, Time Horizon = 25

σ	0.9				0.5				0.1			
Policy	SINE	$GSRS_0$	$GSRS_t$	OPT	SINE	$GSRS_0$	$GSRS_t$	OPT	SINE	$GSRS_0$	$GSRS_t$	OPT
Fast servers	**572.3**	260.6	290.3	750.2	**492.1**	287.3	339.4	757.1	**485.5**	289	297.3	693.2
Slow servers	**551.4**	374.5	224	578.9	**526.7**	372.4	304	633.4	**457.9**	346.2	361.5	595.2
Impatient clients	**568.6**	146.9	156.7	636.3	**559.4**	151	165.9	688.3	**472.2**	138.4	157.4	648.6
Patient clients	**648**	521.5	365.7	724.6	**645.4**	572.4	461.6	746.3	**535.8**	**535.8**	442.2	681.1
High demand of base clients	**793**	617.9	323.3	795.1	**858.5**	652.7	508.1	867	**798.4**	635.6	650.2	835.9
Low demand of base clients	**290.5**	179	87.3	399.7	**298.9**	203.4	152	386.6	**247.7**	197.4	199.6	338
average	**570.6333**	350.0667	241.2167	647.4667	**563.5**	373.2	321.8333	679.7833	**499.5833**	357.0667	351.3667	632

Table A.16 Abandonment Ratio, Erlang-C version of $GSRS_0$ and $GSRS_t$, Time Horizon = 25

σ	0.9				0.5				0.1			
Policy	SINE	$GSRS_0$	$GSRS_t$	OPT	SINE	$GSRS_0$	$GSRS_t$	OPT	SINE	$GSRS_0$	$GSRS_t$	OPT
Fast servers	**0.1381**	0.2171	0.3716	0.0559	**0.1817**	**0.1817**	0.3445	0.0228	0.1427	0.1427	**0.1357**	0.0099
Slow servers	**0.1553**	0.2358	0.4876	0.1448	**0.1538**	0.1695	0.2661	0.0919	**0.1595**	**0.1595**	0.1851	0.0619
Impatient clients	**0.1569**	0.2652	0.5582	0.1220	**0.1467**	0.1819	0.3624	0.0692	**0.1613**	**0.1613**	0.2262	0.0381
Patient clients	**0.0864**	0.1591	0.4484	0.0627	**0.0679**	0.1164	0.2809	0.0248	**0.0844**	**0.0844**	0.1600	0.0132
High demand of base clients	**0.1535**	0.1661	0.4117	0.1535	**0.1168**	0.1204	0.2465	0.1168	**0.0959**	**0.0959**	0.1437	0.0959
Low demand of base clients	**0.2272**	0.3669	0.6926	0.0179	**0.1904**	0.2730	0.4847	0.0018	**0.2327**	**0.2327**	0.3309	0.0003
average	**0.1529**	0.2351	0.4950	0.0928	**0.1429**	0.1738	0.3308	0.0545	**0.1461**	**0.1461**	0.1969	0.0366

Table A.17 Abandonment Ratio, Erlang-A version of $GSRS_0$ and $GSRS_t$, Time Horizon = 25

σ	0.9				0.5				0.1			
Policy	SINE	$GSRS_0$	$GSRS_t$	OPT	SINE	$GSRS_0$	$GSRS_t$	OPT	SINE	$GSRS_0$	$GSRS_t$	OPT
Fast servers	**0.1381**	0.4321	0.3927	0.0559	**0.1817**	0.3811	0.3308	0.0228	**0.1427**	0.3344	0.3228	0.0099
Slow servers	**0.1553**	0.3030	0.4896	0.1448	**0.1538**	0.2913	0.3673	0.0919	**0.1595**	0.2756	0.2546	0.0619
Impatient clients	**0.1569**	0.6250	0.6096	0.1220	**0.1467**	0.6052	0.5758	0.0692	**0.1613**	0.5894	0.5562	0.0381
Patient clients	**0.0864**	0.1591	0.2958	0.0627	**0.0679**	0.1164	0.1899	0.0248	**0.0844**	**0.0844**	0.1543	0.0132
High demand of base clients	**0.1535**	0.2067	0.4090	0.1535	**0.1168**	0.1812	0.2559	0.1168	**0.0959**	0.1426	0.1398	0.0959
Low demand of base clients	**0.2272**	0.4757	0.7214	0.0179	**0.1904**	0.3968	0.5272	0.0018	**0.2327**	0.3474	0.3475	0.0003
average	**0.1529**	0.3669	0.4863	0.0928	**0.1429**	0.3287	0.3745	0.0545	**0.1461**	0.2956	0.2959	0.0366

Table A.18 Number of iterations to find optimal staffing level.

	Time Horizon = 10					Time Horizon = 25				
σ	0.9	0.7	0.5	0.3	0.1	0.9	0.7	0.5	0.3	0.1
Fast servers	1	1	1	1	1	1	1	1	1	1
Slow servers	4	4	1	1	1	4	4	3	1	1
Impatient clients	1	1	1	1	1	1	1	1	1	1
Patient clients	1	1	1	1	1	1	1	1	1	1
High demand of base clients	3	1	1	1	1	3	2	2	1	1
Low demand of base clients	1	1	1	1	1	1	1	1	1	1

Proofs of the results

Proof of Lemma 2

Proof. Consider IVP (2.8) where $f_i(\boldsymbol{q}, t) \in C^1_{\geq 0}$ for each region $\mathcal{S}_i$, $i \in \{1, 2, 3, 4\}$ (see Appendix A "State Space"). According to the classical theory of linear ODEs, the solution to each of the equations $\dot{\boldsymbol{q}} = f_i(\boldsymbol{q}, t)$, $\boldsymbol{q}_0 = (q_a, q_b, q_c)^T$ exists and is unique for all i with a maximal interval of existence J_i. The difficulty arises if $\boldsymbol{q}(t) \in \Sigma$ for some time t. We show that the number of points where the solution belongs to the boundary Σ is finite, i.e., such points form a set with finite cardinality and are included in the maximal interval of existence of (2.8).

Let $H(\boldsymbol{q})$ be an orthogonal vector to a boundary Σ_{ij} $(i \neq j)$ at point $\boldsymbol{q}(t)$, so that $\boldsymbol{q}(t) \in \Sigma_{ij}$, i.e., the solution $\boldsymbol{q}(t)$ hits the boundary at some time. Then, define $\sigma(\boldsymbol{q}) = \langle H(\boldsymbol{q}), f_i(\boldsymbol{q}, t) \rangle \langle H(\boldsymbol{q}), f_j(\boldsymbol{q}, t) \rangle$, where $\langle \cdot, \cdot \rangle$ denotes the scalar product. Let $\mathcal{T}_\Sigma = \{t : \sigma(\boldsymbol{q}) > 0\}$ be a set of time points where $\boldsymbol{q}(t)$ approaches Σ transversally and let $\mathcal{E}_\Sigma = \{t : \sigma(\boldsymbol{q}) = 0\}$ be a set of time points where $\boldsymbol{q}(t)$ is tangent to Σ. Because for any $\boldsymbol{q}(t) \in \Sigma_{ij}$ separating two adjacent regions $\mathcal{S}_i$ and $\mathcal{S}_j$, $f_i(\boldsymbol{q}, t) = f_j(\boldsymbol{q}, t)$, $i, j \in \{1, 2, 3, 4\}$ and $i \neq j$, (A.1) is piecewise smooth continuous. In such systems, orbits in region $\mathcal{S}_i$ approaching the boundary Σ_{ij} transversally, cross it, and enter into the adjacent region $\mathcal{S}_j$. Then, $\boldsymbol{q}(t)$ achieves a local extremum at $t \in \mathcal{E}_\Sigma$. Hence, $\mathcal{E}_\Sigma \subset \cup_{i=1}^4 J_i$ and the cardinality of $\mathcal{T}_\Sigma$ is finite. Let $I_0 = \cup_{i=1}^4 J_i \cup \mathcal{E}_\Sigma \setminus \mathcal{T}_\Sigma = \cup_{i=1}^4 J_i \setminus \mathcal{T}_\Sigma$, then because I_0 is the union of open intervals, over which the solution to (2.8) exists and is unique, we can define values of $\boldsymbol{q}(t)$ as the finite collection of points excluded from I_0 as initial conditions. As a result, the solution to (2.8) exists over $I_0 \cup \mathcal{T}_\Sigma = \mathbb{R}_{\geq 0}$ and is unique; it is also piecewise smooth. By specifying that $I_0 \cup \mathcal{T}_\Sigma$ is the maximal interval of existence of (2.8), the proof is complete. Note that a similar result for a system with unspecialized servers is presented in Mandelbaum et al. (1998). The difference is that in our system, the servers are specialized. □

Proof of Lemma 3

Proof. A process $\boldsymbol{x}(t)$ is stable if $\lim_{t\to+\infty} \boldsymbol{x}(t) < \infty$. Thus, we must show that $\lim_{t\to+\infty} \boldsymbol{q}(t) < \infty$. We first show that $\lim_{t\to+\infty} q_a(t) < 0$. The general solution to the ODE for $q_a(t)$ is

$$q_a(t) = \frac{1}{u(t)} \left(\int u(t) b(t) dt + C \right), \tag{A.5}$$

where $u(t) = e^{\int a dt}$ is an integration factor, C is a real number, and $b(t)$ is a forcing term. Because $b(t)$ is a sum of $\lambda(t)$ and a scalar, and $\lim_{t\to+\infty} \lambda(t) < \infty$, $\lim_{t\to\infty} b(t)$ exists and $\lim_{t\to+\infty} q_a(t) < \infty$.

Separating $q_a(t)$ from $q_b(t)$ and $q_c(t)$, we rewrite the non-homogeneous system of equations (2.8) with respect to $\boldsymbol{q}_1(t) = (q_b(t), q_c(t)^T)$ in matrix form

$$\dot{\boldsymbol{q}}_1(t) = \boldsymbol{A}\boldsymbol{q}_1(t) + \boldsymbol{b}(t), \tag{A.6}$$

where $\boldsymbol{A}$ is a non-singular $n \times n$ matrix of coefficients of a homogeneous part of (A.6) and $\boldsymbol{b}(t)$ is a continuous vector-valued function. By Proposition 2, the solution to (2.8) exists and is unique. Denote a fundamental matrix solution to $\dot{\boldsymbol{q}}_1(t) = \boldsymbol{A}\boldsymbol{q}_1(t)$ by $\boldsymbol{\Phi}(t)$. From the classical theory of linear ODEs, the solution to (A.6) is

$$\boldsymbol{q}_1(t) = \boldsymbol{\Phi}(t)\boldsymbol{\Phi}^{-1}(0)\boldsymbol{q}_0 + \int_0^t \boldsymbol{\Phi}(t)\boldsymbol{\Phi}^{-1}(z)\boldsymbol{b}(z)dz, \tag{A.7}$$

where $\boldsymbol{q}_0$ is a vector of initial conditions. We must show that $\lim_{t\to+\infty} \boldsymbol{q}_1(t) < \infty$. Recall that $\boldsymbol{\Phi}(t) = e^{\boldsymbol{A}t}$, such that $\boldsymbol{A} = \begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \end{bmatrix}$, where

$$a_1 = \begin{cases} -\tau_b, & \textit{if } q_b > s_b, \\ -\mu_b, & \textit{if } q_b \le s_b, \end{cases} \qquad a_2 = \begin{cases} \tau_b\theta_c, & \textit{if } q_b > s_b, \\ \mu_b\theta_{bc}, & \textit{if } q_b \le s_b, \end{cases} \qquad b_1 = r, \quad b_2 = -(\zeta + r). \tag{A.8}$$

Consider the characteristic equation $\psi^2 - T\psi + D = 0$, where T and D are the trace and the determinant of $\boldsymbol{A}$ respectively. For a set of positive queueing parameters $\sqrt{T^2 - 4D} > 0$, and because $T < 0$ and $D > 0$ for $\boldsymbol{q}(t) \in \cup_{i=1}^4 \mathcal{S}_i$, there are two distinct eigenvalues of $\boldsymbol{A}$, ψ_1 and ψ_2, that are negative and real. Because elements of $e^{\boldsymbol{A}t}$ are linear combinations of $e^{\psi_1 t}$ and $e^{\psi_2 t}$, $e^{\boldsymbol{A}t} \to \boldsymbol{0}$ as $t \to +\infty$. Therefore, $\lim_{t\to+\infty} \boldsymbol{\Phi}(t)\boldsymbol{\Phi}^{-1}(0)\boldsymbol{q}_0 = \lim_{t\to+\infty} e^{\boldsymbol{A}t}\boldsymbol{\Phi}^{-1}(0)\boldsymbol{q}_0 = \boldsymbol{0}$. The limit of the second term in (A.7) exists when the improper integral converges. Because $\boldsymbol{b}(t)$ is a vector whose elements are the sum of $\lambda(t)$, a scalar, and $q_a(t)$ multiplied by a scalar, $\lim_{t\to+\infty} \lambda(t) < \infty$. Thus, $\lim_{t\to+\infty} \boldsymbol{b}(t) < \infty$ and further, there exists vector $\boldsymbol{M}$, such that $|\boldsymbol{b}(t)| \le \boldsymbol{M}$. By the squeeze theorem:

$$\lim_{t\to+\infty}\left(\int_0^t \boldsymbol{\Phi}(t)\boldsymbol{\Phi}^{-1}(z)\boldsymbol{b}(z)dz\right) \leq \lim_{t\to+\infty}\left(\int_0^t \boldsymbol{\Phi}(t)\boldsymbol{\Phi}^{-1}(z)\boldsymbol{M}dz\right) = \lim_{t\to+\infty}\left(\boldsymbol{M}e^{\boldsymbol{A}t}\int_0^t e^{-\boldsymbol{A}z}dz\right)$$
$$= \lim_{t\to+\infty} -\boldsymbol{M}\boldsymbol{A}^{-1}e^{\boldsymbol{A}t}\left(e^{-\boldsymbol{A}t}-\boldsymbol{I}\right) = \lim_{t\to+\infty}\left(-\boldsymbol{M}\boldsymbol{A}^{-1}+\boldsymbol{M}\boldsymbol{A}^{-1}e^{\boldsymbol{A}t}\right) < \infty.$$

We have shown that the limit of (A.7) exists if $\lim_{t\to+\infty}\lambda(t) < \infty$ which completes the proof. □

Proof of Lemma 4

Proof. We consider the general solution to (A.6) in (A.7). From Lemma 3, $\boldsymbol{q}_a(t)$ is bounded. Further, there exists a vector $\boldsymbol{M}_1$, such that $|\boldsymbol{\Phi}(t)\boldsymbol{\Phi}^{-1}(0)\boldsymbol{q}_0| < \boldsymbol{M}_1$ for any time $t \in \mathbb{R}_{\geq 0}$. Suppose there exists a scalar K' such that $\lambda(t) < K'$. Then, as in Lemma 3, there exists a vector $\boldsymbol{K}$ such that $|\boldsymbol{b}(t)| \leq \boldsymbol{K}$. Further, there exists a vector $\boldsymbol{M}_2$ such that

$$\lim_{t\to+\infty}\int_0^t \boldsymbol{\Phi}(t)\boldsymbol{\Phi}^{-1}(z)\boldsymbol{b}(z)dz \leq \lim_{t\to+\infty}\int_0^t \boldsymbol{\Phi}(t)\boldsymbol{\Phi}^{-1}(z)\boldsymbol{K}dz \leq \boldsymbol{M}_2 < \infty.$$

Thus, $\boldsymbol{q}_1(t) < \boldsymbol{M}$ where $\boldsymbol{M} = \boldsymbol{M}_1 + \boldsymbol{M}_2$. Hence, $\boldsymbol{q}(t)$ is bounded which concludes the proof. □

Proof of Proposition 1

Proof. We evaluate the asymptotic behaviour of $\boldsymbol{q}(t)$ in the launch and loyalty regions. The launch phase includes two scenarios: (I) $\boldsymbol{q}_a(t) \in \mathcal{S}_1$ and (II) $\boldsymbol{q}_a(t) \in \mathcal{S}_4$. We consider them separately.

I. We rewrite (2.11) with subscript "1" in state functions corresponding to the mode of operation:

$$\begin{pmatrix}\dot{q}_{a_1}(t)\\ \dot{q}_{b_1}(t)\\ \dot{q}_{c_1}(t)\end{pmatrix} = \begin{pmatrix}\lambda_1 e^{-\delta t}+\lambda_0-\mu_a s_a-\tau_a(q_{a_1}(t)-s_a)\\ rq_{c_1}(t)-\mu_b s_b-\tau_b(q_{b_1}(t)-s_b)\\ \theta_{ac}s_a\mu_a+\theta_{bc}s_b\mu_b+\theta_c\tau_b(q_{b_1}(t)-s_b)-(r+\zeta)q_{c_1}(t)\end{pmatrix},$$
$$\boldsymbol{q}_{0_1} = (q_{a_1}(t_0), q_{b_1}(t_0), q_{c_1}(t_0))^T.$$

The non-homogeneous linear ODE for $q_a(t)$ is solved in closed-form by standard methods:

$$q_{a_1}(t) = \left[q_{a_1}(t_0) - \frac{\lambda_1}{\tau_a-\delta}e^{-\delta t_0} - \frac{\lambda_0+s_a(\tau_a-\mu_a)}{\tau_a}\right]e^{-\tau_a(t-t_0)} + \frac{\lambda_1}{\tau_a-\delta}e^{-\delta t} + q^*_{a_1}, \tag{A.9}$$

where $q^*_{a_1} = \lim_{t\to+\infty} q_{a_1}(t) = \frac{\lambda_0+s_a(\tau_a-\mu_a)}{\tau_a}$. The expression for the equilibrium point $(q^*_{b_1}, q^*_{c_1})$ is obtained by solving the linear system $(\dot{q}_{b_1}(t), \dot{q}_{c_1}(t))^T = \mathbf{0}$. Then, by Cramer's rule:

$$(q^*_{b_1}, q^*_{c_1}) = \left(\frac{r\theta_{ac}s_a\mu_a + s_b(\tau_b-\mu_b)(r+\zeta) - s_b r(\theta_{\tau c}\tau_b - \theta_{bc}\mu_b)}{\tau_b(r(1-\theta_{\tau c})+\zeta)}, \frac{\theta_{ac}s_a\mu_a + s_b\mu_b(\theta_{bc}-\theta_c)}{r(1-\theta_c)+\zeta}\right).$$

II. We rewrite (2.11) accordingly

$$\begin{pmatrix} \dot{q}_{a_1}(t) \\ \dot{q}_{b_1}(t) \\ \dot{q}_{c_1}(t) \end{pmatrix} = \begin{pmatrix} \lambda_1 e^{-\delta t} + \lambda_0 - \mu_a s_a - \tau_a(q_{a_1}(t) - s_a) \\ r q_{c_1}(t) - \mu_b q_{b_1}(t) \\ \theta_{ac}s_a\mu_a + \theta_{bc}q_{b_1}(t)\mu_b - (r+\zeta)q_{c_1}(t) \end{pmatrix},$$
$$\boldsymbol{q}_{0_1} = (q_{a_1}(t_0), q_{b_1}(t_0), q_{c_1}(t_0))^T.$$

The solution to $q_{a_1}(t)$ remains as in Scenario I and the equilibrium point $(q^*_{b_1}, q^*_{c_1})$ is obtained:

$$(q^*_{a_1}, q^*_{b_1}, q^*_{c_1})^T = \left(\frac{\lambda_0 + s_a(\tau_a-\mu_a)}{\tau_a}, \frac{r\theta_{ac}s_a\mu_a}{\mu_b(r(1-\theta_{bc})+\zeta)}, \frac{\theta_{ac}s_a\mu_a}{r(1-\theta_{bc})+\zeta}\right)^T. \tag{A.10}$$

For the loyalty phase, we write down (2.11) for $q_a(t) < s_a$ and $q_b(t) \geq s_b$

$$\begin{pmatrix} \dot{q}_{a_2}(t) \\ \dot{q}_{b_2}(t) \\ \dot{q}_{c_2}(t) \end{pmatrix} = \begin{pmatrix} \lambda_1 e^{-\delta t} + \lambda_0 - \mu_a q_{a_2}(t) \\ r q_{c_2}(t) - \mu_b s_b - \tau_b(q_{b_2}(t) - s_b) \\ \theta_{ac}q_{a_2}(t)\mu_a + \theta_{bc}s_b\mu_b + \theta_c\tau_b(q_{b_2}(t) - s_b) - (r+\zeta)q_{c_2}(t) \end{pmatrix},$$
$$\boldsymbol{q}_{0_2} = (q_{a_2}(t_0), q_{b_2}(t_0), q_{c_2}(t_0))^T.$$

The equation for $q_{a_2}(t)$ is solved in closed-form

$$q_{a_2}(t) = \left[q_{a_2}(t_0) - \frac{\lambda_0}{\mu_a} - \frac{\lambda_1}{\mu_a-\delta}e^{-\delta t_0}\right]e^{-\mu_a(t-t_0)} + \frac{\lambda_1}{\mu_a-\delta}e^{-\delta t} + \frac{\lambda_0}{\mu_a} \tag{A.11}$$

with $q^*_{a_2} = \frac{\lambda_0}{\mu_a}$. We evaluate the asymptotic behaviour of $(q_{b_2}(t), q_{c_2}(t))^T$ by considering the system

$$\begin{pmatrix} \dot{q}_{b_2}(t) \\ \dot{q}_{c_2}(t) \end{pmatrix} = \begin{pmatrix} r q_{c_2}(t) - \mu_b s_b - \tau_b(q_{b_2}(t) - s_b) \\ \theta_{ac}q_{a_2}(t)\mu_a + \theta_{bc}s_b\mu_b + \theta_c\tau_b(q_{b_2}(t) - s_b) - (r+\zeta)q_{c_2}(t) \end{pmatrix}, \tag{A.12}$$
$$\boldsymbol{q}_{0_2} = (q_{a_2}(t_0), q_{b_2}(t_0), q_{c_2}(t_0))^T.$$

The dynamical system (A.12) is non-autonomous, and the solution to $(\dot{q}_{b_2}(t), \dot{q}_{c_2}(t))^T = \mathbf{0}$ is a vector valued function of time. To overcome this difficulty, we construct a quasi-static equilibrium (QSE) of (A.12) (see, e.g, Ding et al. 2015, Hahn 2016, Hoyer-Leitzel et al. 2017). A QSE of $\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t), q(t))$ is a set of equilibrium solutions of the corresponding autonomous systems with $q(t)$ equal to some constant, i.e., any point in the range of $q(t)$. Thus, the QSE is a set of equilibria parameterized by $p \in \mathbb{R}_{\geq 0}$ which gives a sequence of autonomous systems $\{(q_{b_2}(t), q_{c_2}(t))^T\}_p$. Their solutions track the QSE in that for any time t, there exists a ball with finite radius $\boldsymbol{R}$, such that $\left|(q_{b_2}(t), q_{c_2}(t))^T - (\bar{q}^*_{b_2}, \bar{q}^*_{c_2})^T\right| < \boldsymbol{R}$, where $(\bar{q}^*_{b_2}, \bar{q}^*_{c_2})^T$ is an equilibrium point of $\{(q_{b_2}(t), q_{c_2}(t))^T\}_p$ (it follows from Lemma 4). The solution of (A.12) moves in the direction of the equilibrium points $\{(q_{b_2}(t), q_{c_2}(t))^T\}_{p\to 0}$ and $\{(q_{b_2}(t), q_{c_2}(t))^T\}_{p\to +\infty}$ when $t \to 0$ and $t \to +\infty$, respectively. Because we are interested in the behaviour of $\boldsymbol{q}(t)$ at early times, we use the equilibrium point $\{(q_{b_2}(t), q_{c_2}(t))^T\}_{p\to 0}$, i.e., $(q^*_{b_2}, q^*_{c_2})$. Thus, the asymptotic dynamics of $\boldsymbol{q}(t)$ in the loyalty phase is

$$\begin{pmatrix} q^*_{a_2} \\ q^*_{b_2} \\ q^*_{c_2} \end{pmatrix} = \left(\frac{\lambda_0}{\mu_a}, \frac{r\theta_{ac} q_{a_2}(t_0)\mu_a + s_b(\tau_b - \mu_b)(r+\zeta) - s_b r(\theta_{\tau c}\tau_b - \theta_{bc}\mu_b)}{\tau_b(r(1-\theta_{\tau c}) + \zeta)}, \frac{\theta_{ac} q_{a_2}(t_0)\mu_a + s_b\mu_b(\theta_{bc} - \theta_c)}{r(1-\theta_c)+\zeta} \right)^T .$$

□

Proof of Lemma 5

Proof. Let $\mathcal{QN}_1$, $\mathcal{QN}_2$, $\mathcal{QN}_3$ be queueing systems defined by (2.11) and originating in the launch, loyalty, and lessening phases with initial conditions $\boldsymbol{q}_{0_1} \in \mathcal{S}_1 \cup \mathcal{S}_4$, $\boldsymbol{q}_{0_2} \in \mathcal{S}_2$ and $\boldsymbol{q}_{0_3} \in \mathcal{S}_3$, respectively. Also, let neither system transitions to a higher mode of operations as t gets larger. The asymptotic behaviour of $\mathcal{QN}_1$ and $\mathcal{QN}_2$ is characterized by the limiting points in Proposition 1. The limiting behaviour of $\mathcal{QN}_3$ is defined by the dynamics of $q_{a_2}(t)$ in (A.11) and the solution to $(\dot{q}_{b_3}(t), \dot{q}_{c_3}(t))^T = \mathbf{0}$. As in Proposition 1, we construct a sequence $\{(q_{b_3}(t), q_{c_3}(t))^T\}_p$ indexed by p and derive an expression for the equilibrium point of the term with $p \to 0$.

$$\left(q^*_{b_3}(t), q^*_{c_3}(t)\right)^T|_{p=0} = \left(q^*_{b_3}, q^*_{c_3}\right)^T = \left(\frac{r\theta_{ac} q_{a_3}(t_0)\mu_a}{\mu_b(r(1-\theta_{bc})+\zeta)}, \frac{\theta_{ac} q_{a_3}(t_0)\mu_a}{r(1-\theta_{bc})+\zeta} \right)^T . \tag{A.13}$$

Because $q_{a_3}(t_0) < q_{a_2}(t_0)$, and if $s_a \gg q_{a_3}(t_0)$, and $\theta_{bc} > \theta_c$ the following holds

$$\frac{\theta_{ac}q_{a_3}(t_0)\mu_a}{r(1-\theta_{bc})+\zeta} < \frac{\theta_{ac}q_{a_2}(t_0)\mu_a + s_b\mu_b(\theta_{bc}-\theta_c)}{r(1-\theta_{\tau c})+\zeta} < \frac{\theta_{ac}s_a\mu_a + s_b\mu_b(\theta_{bc}-\theta_c)}{r(1-\theta_c)+\zeta}. \qquad \square$$

Proof of Proposition 2

Proof. Let $\mathcal{QN}_1$ and $\mathcal{QN}_2$ be fluid systems defined by (2.11) and originating in mode $k \in \{1,2\}$, respectively. The homogeneous systems of the ODEs with respect to $(q^h_{b_k}, q^h_{c_k})^T$ are as follows:

$$\begin{pmatrix} \dot{q}^h_{b_1}(t) \\ \dot{q}^h_{c_1}(t) \end{pmatrix} = \begin{pmatrix} rq^h_{c_1}(t) - \tau_b q^h_{b_1}(t) \\ \theta_c\tau_b q^h_{b_1}(t) - (r+\zeta)q^h_{c_1}(t) \end{pmatrix}, \quad \begin{pmatrix} \dot{q}^h_{b_2}(t) \\ \dot{q}^h_{c_2}(t) \end{pmatrix} = \begin{pmatrix} rq^h_{c_2}(t) - \mu_b q^h_{b_2}(t) \\ \theta_{bc}\mu_b q^h_{b_2}(t) - (r+\zeta)q^h_{c_2}(t) \end{pmatrix}, \tag{A.14}$$
$$\boldsymbol{q}_{0_1} = (q_{b_1}(t_0), q_{c_1}(t_0))^T. \qquad\qquad \boldsymbol{q}_{0_2} = (q_{b_2}(t_0), q_{c_2}(t_0))^T,$$

where the first system corresponds to mode 1 with $\boldsymbol{q}(t) \in \mathcal{S}_1$ and mode 2, and the second system corresponds to mode 1 with $\boldsymbol{q}(t) \in \mathcal{S}_4$. Let ψ_k and χ_k be the smallest and largest eigenvalues of $\boldsymbol{A}_k$ corresponding to mode k, respectively. Matrix $\boldsymbol{A}_k$ is defined in Lemma 3, where we also show that it has negative real and distinct eigenvalues. Let $\boldsymbol{v}(\psi_k) = (v_1(\psi_k), v_2(\psi_k))^T$ and $\boldsymbol{v}(\chi_k) = (v_1(\chi_k), v_2(\chi_k))^T$ be the eigenvectors corresponding to eigenvalues ψ_k and χ_k, respectively. Then,

$$\begin{pmatrix} q^h_{b_k}(t) \\ q^h_{c_k}(t) \end{pmatrix} = \begin{pmatrix} u_k e^{\psi_k t} v_1(\psi_k) + \omega_k e^{\chi_k t} v_1(\chi_k) \\ u_k e^{\psi_k t} v_2(\psi_k) + \omega_k e^{\chi_k t} v_2(\chi_k) \end{pmatrix}, \tag{A.15}$$

where ω_k and u_k are scalars. Because $\psi_k < \chi_k < 0$, $(q^h_{b_k}(t), q^h_{c_k}(t))^T$ approaches $\mathbf{0}$ as $t \to \infty$, $(q_{b_k}(t), q_{c_k}(t))^T$ approaches its equilibrium solution when $\omega_k e^{\chi_k t} \to 0$.

For a fixed $\boldsymbol{q}_{0_k} = (q_{b_k}(t_0), q_{c_k}(t_0))^T$, the expression for ω_k is obtained using Cramer's rule. That is, denoting the first coordinate of the eigenvectors corresponding to the eigenvalues ψ_k and χ_k by $v(\psi_k)$ and $v(\chi_k)$, respectively, we have that

$$w_k = \frac{v(\psi_k)[q_c(t_0) - q^*_{c_k}(t)] - q_b(t_0) + q^*_{b_k}(t)}{e^{\chi_k t_0}[v(\psi_k) - v(\chi_k)]} \text{ for } k \in \{1,2,3\}. \qquad \square$$

Proof of Corollary 1

Proof. Let $q_{a_1}(t)$ approach $q^*_{a_1}$ at time t^* in the launch phase. Then, as per (A.9),

$$\lim_{t \to t^*} \left[q_{a_1}(t_0) - \frac{\lambda_1}{\tau_a - \delta} e^{-\delta t_0} - \frac{\lambda_0 + s_a(\tau_a - \mu_a)}{\tau_a} \right] e^{-\tau_a (t - t_0)} = 0, \quad \lim_{t \to t^*} \frac{\lambda_1}{\tau_a - \delta} e^{-\delta t} = 0.$$

The desired result holds as $t \to t^*$:

$$\omega_0 := q_{a_1}(t_0) - \frac{\lambda_1}{\tau_a - \delta} e^{-\delta t_0} - \frac{\lambda_0 + s_a(\tau_a - \mu_a)}{\tau_a}, \quad |\omega_0| e^{-\tau_a t} \to 0 \qquad \square$$

Proof of Lemma 6

Proof. Let $\min q_a(t) \geq s_a$ (i.e., $q_a(t) \geq s_a$, $\forall t$). Then, the solution to $\dot{q}_a(t) = \bar{\lambda} + \bar{\lambda}\sigma \sin(\omega t) - \mu_a s_a - \tau_a(q_a(t) - s_a)$ approaches $q^*_a(t) = \alpha_a + \beta_a \sin(\omega t) + \gamma_a \cos(\omega t)$ as $t \to +\infty$, where the constants

$$\alpha_a = \frac{\bar{\lambda} + s_a(\tau_a - \mu_a)}{\tau_a}, \quad \beta_a = \frac{\tau_a \bar{\lambda}\sigma}{\omega^2 + \tau_a^2}, \quad \gamma_a = -\frac{\omega \bar{\lambda}\sigma}{\omega^2 + \tau_a^2}$$

are derived by the method of undetermined coefficients. Because $q_a(t)$ is overloaded, the asymptotic behaviour of the system of ODEs with respect to $(q_b(t), q_c(t))^T$ remains as in Proposition 1 part I.

Let $\max q_a(t) \leq s_a$ (i.e., $q_a(t) \leq s_a$, $\forall t$). Then, $q_a(t) = \omega_0 e^{-\mu_a t} + \alpha_a + \beta_a \sin(\omega t) + \gamma_a \cos(\omega t)$ where

$$\alpha_a = \frac{\bar{\lambda}}{\mu_a}, \quad \beta_a = \frac{\mu_a \bar{\lambda}\sigma}{\omega^2 + \mu_a^2}, \quad \gamma_a = -\frac{\omega \bar{\lambda}\sigma}{\omega^2 + \mu_a^2}.$$

The constants can again by derived by the method of undetermined coefficients and as a result, $q_a(t)$ approaches the periodic orbit $q^*_a(t) = \alpha_a + \beta_a \sin(\omega t) + \gamma_a \cos(\omega t)$ as $t \to +\infty$.

To show that $q_b(t)$ and $q_c(t)$ approach periodic orbits, we need to consider two systems of ODEs corresponding to cases where $q_b(t) \geq s_b$ and $q_b(t) < s_b$, respectively. We write them in general form:

$$\begin{pmatrix} \dot{q}_b(t) \\ \dot{q}_c(t) \end{pmatrix} = \begin{pmatrix} a_1 q_b(t) + b_1 q_c(t) + d_1 \\ a_2 q_b(t) + b_2 q_c(t) + d_2 + \mu_a \theta_{ac} \omega_0 e^{-\mu_a t} + \mu_a \theta_{ac} \phi(t) \end{pmatrix}, \tag{A.16}$$

where

$$a_1 = \begin{cases} -\tau_b, \ \textit{if } q_b \geq s_b, \\ -\mu_b, \ \textit{if } q_b < s_b, \end{cases} \qquad a_2 = \begin{cases} \tau_b\theta_c, \ \textit{if } q_b \geq s_b, \\ \mu_b\theta_{bc}, \ \textit{if } q_b < s_b, \end{cases} \qquad b_1 = r, \quad b_2 = -(\zeta + r), \tag{A.17}$$

$$d_1 = \begin{cases} s_b(\tau_b - \mu_b), \ \textit{if } q_b \geq s_b, \\ 0, \ \textit{if } q_b < s_b, \end{cases} \qquad d_2 = \begin{cases} \theta_{ac}\bar{\lambda} + s_b(\theta_{bc}\mu_b - \theta_c\tau_b), \ \textit{if } q_b \geq s_b, \\ \theta_{ac}\bar{\lambda}, \ \textit{if } q_b < s_b, \end{cases} \tag{A.18}$$

$$\phi(t) = \bar{\phi}\big(\mu_a \sin(\omega t) - \omega\cos(\omega t)\big), \quad \bar{\phi} = \frac{\bar{\lambda}\sigma}{\mu_a^2 + \omega^2}. \tag{A.19}$$

The general solution of (A.16) is obtained by standard ODE solutions methods:

$$\begin{pmatrix} q_b(t) \\ q_c(t) \end{pmatrix} = \begin{pmatrix} \omega_1 e^{\psi_1 t} \frac{-b_2+\psi_1}{a_1} + \omega_2 e^{\psi_2 t} \frac{-b_2+\psi_2}{a_1} + \alpha_b + B_1 e^{-\mu_a t} + \beta_b \sin(\omega t) + \gamma_b \cos(\omega t) \\ \omega_1 e^{\psi_1 t} + \omega_2 e^{\psi_2 t} + \alpha_c + B_2 e^{-\mu_a t} + \beta_c \sin(\omega t) + \gamma_c \cos(\omega t) \end{pmatrix}, \tag{A.20}$$

where

$$\alpha_b = -\frac{d_1}{a_1} - \frac{b_1}{a_1}\left(\frac{a_2 d_1 - d_2 a_1}{a_1 b_2 - a_2 b_1}\right), \quad \alpha_c = \frac{a_2 d_1 - d_2 a_1}{a_1 b_2 - a_2 b_1},$$

$$B_1 = \frac{b_1 \mu_a \theta_{ac}\omega_0(a_1 + \mu_a)}{(a_1+\mu_a)^2(\mu_a + b_2) - (a_1 + \mu_a)a_2 b_1}, \quad B_2 = -\frac{\mu_a\theta_{ac}\omega_0(a_1+\mu_a)}{(a_1+\mu_a)(\mu_a+b_2) - a_2 b_1}.$$

And β_c, γ_c, β_b and γ_b are computed from the following matrix equation

$$\begin{bmatrix} \beta_c \\ \gamma_c \\ \beta_b \\ \gamma_b \end{bmatrix} = \begin{bmatrix} b_2 & w & a_2 & 0 \\ -w & b_2 & 0 & a_2 \\ b_1 & 0 & a_1 & w \\ 0 & b_1 & w & a_1 \end{bmatrix}^{-1} \begin{bmatrix} -\mu_a^2\theta_{ac}\bar{\phi} \\ \omega\mu_a\theta_{ac}\bar{\phi} \\ 0 \\ 0 \end{bmatrix}.$$

As $t \to +\infty$, $q_b(t)$ and $q_c(t)$ approach periodic orbits $q_b^*(t)$ and $q_c^*(t)$ as stated. □

Proof of Lemma 7

Proof. We first derive the expression for the lower bound of $q_a^*(t)$. Because $q_a^*(t) \in \mathbb{R}_{\geq 0}$, its lower bound exists: let $\underline{q}_a^* \in \mathbb{R}_{\geq 0}$ be a lower bound of $q_a^*(t)$. To derive it, we compute $\min q_a^*(t)$. For simplicity, let $g(t) = q_a^*(t) = \alpha_a + \beta_a \sin(\omega t) + \gamma_a \cos(\omega t)$, then $\dot{g}(t) = \omega\beta_a cos(\omega t) - \omega\gamma_a sin(\omega t)$.

By solving $\dot{g}(t) = 0$, we find the critical points t^*: $\omega t^* = \arctan\left(\frac{\beta_a}{\gamma_a}\right) + n\pi$, where $n \in \mathbb{Z}$. Now, consider $\ddot{g}(t) = -\omega^2\beta_a \sin(\omega t) - \omega^2\gamma_a \cos(\omega t)$ noting $cos(\omega t^*) \neq 0$ and $sin(\omega t^*) \neq 0$ for $\omega > 0$. Observe that $\beta_a > 0$ and $\gamma_a < 0$ so $\frac{\beta_a}{\gamma_a} < 0$ as per Lemma 6. Because $\arctan(-\theta) \in (-\frac{\pi}{2}, 0)$ for $\theta > 0$, for $n = 0$, $\omega t^* \in (-\frac{\pi}{2}, 0)$ so that $-\omega^2\beta_a \sin(\omega t^*) > 0$, $-\omega^2\gamma_a \cos(\omega t^*) > 0$, and $\ddot{g}(t^*) > 0$. Thus, $\min g(t) = \alpha_a + \beta_a \sin(\omega t^*) + \gamma_a \cos(\omega t^*)$ where $\omega t^* = arctan\left(-\frac{\mu_a}{\omega}\right) + n\pi$ for $n = 0, 2, 4...$ which gives:

$$\min g(t) = \underline{q}_a^* = \frac{\bar{\lambda}}{\mu_a} - \frac{\bar{\lambda}\sigma}{\mu_a^2 + \omega^2}\left(\mu_a sin\left[arctan\left(\frac{\mu_a}{\omega}\right)\right] + \omega cos\left[arctan\left(\frac{\mu_a}{\omega}\right)\right]\right). \tag{A.21}$$

Similarly, because $q_b^*(t) \in \mathbb{R}_{\geq 0}$, its lower bound exists. In a similar fashion,

$$\underline{q}_b^* = \alpha_b + \beta_b sin\left[arctan\left(\frac{\beta_b}{\gamma_b}\right)\right] + \gamma_b cos\left[arctan\left(\frac{\beta_b}{\gamma_b}\right)\right],$$

with α_b, β_b, and γ_b derived in Lemma 6.

The upper bound of $q_c^*(t)$ equals to q_c^* as in Proposition 1 part I. We use a balance of flow argument to show this result. Customers arrive to station 3 from station 1 and 2. Hence, if $q_a(t)$ and $q_b(t)$ increase, $q_c(t)$ also weakly increases. As a result, $q_c(t)$ should obtain its largest value when $q_a(t) \geq s_a$ and $q_b(t) \geq s_b$. Thus, for any $q_a(t) \geq s_a$ and $q_b(t) \geq s_b$, the asymptotic behaviour of $q_c(t)$ is described by the equilibrium point from Proposition 1, i.e., $q_c^* = \frac{\theta_{ac}s_a\mu_a + s_b\mu_b(\theta_{bc} - \theta_c)}{r(1-\theta_c)+\zeta}$. This holds for the asymptotic curves $q_a^*(t)$, $q_b^*(t)$, and $q_c^*(t)$: for any value of $q_a^*(t)$ and $q_b^*(t)$ less than s_a and s_b, respectively, $q_c^*(t) \leq q_c^*$, $\forall t$. Thus, $\bar{q}_c^*$ is the upper bound of $q_c^*(t)$ as desired. □

Proof of Lemma 8

Proof. We find roots of the equation $q_a^*(t) = s_a$. First note the trigonometric identities

$$sin(\omega t) = 2sin\left(\frac{\omega t}{2}\right)cos\left(\frac{\omega t}{2}\right) = 2tan\left(\frac{\omega t}{2}\right)cos^2\left(\frac{\omega t}{2}\right) = \frac{2tan\left(\frac{\omega t}{2}\right)}{sec^2\left(\frac{\omega t}{2}\right)}, \tag{A.22}$$

$$sec^2\left(\frac{\omega t}{2}\right) = \frac{1}{cos^2\left(\frac{\omega t}{2}\right)} = \frac{sin^2\left(\frac{\omega t}{2}\right) + cos^2\left(\frac{\omega t}{2}\right)}{cos^2\left(\frac{\omega t}{2}\right)} = tan^2\left(\frac{\omega t}{2}\right) + 1. \tag{A.23}$$

Substituting $z = tan\left(\frac{\omega t}{2}\right)$ and (A.23) into (A.22) we obtain $sin(\omega t) = \frac{2z}{1+z^2}$. Similarly, rewriting

$$cos(\omega t) = 1 - 2sin^2\left(\frac{\omega t}{2}\right) = 1 - 2tan^2\left(\frac{\omega t}{2}\right)cos^2\left(\frac{\omega t}{2}\right) = 1 - \frac{2tan^2\left(\frac{\omega t}{2}\right)}{sec^2\left(\frac{\omega t}{2}\right)},$$

and substituting $z = tan\left(\frac{\omega t}{2}\right)$ gives $cos(\omega t) = 1 - \frac{2z^2}{1+z^2} = \frac{1-z^2}{1+z^2}$. As a result, $q_a^*(t) = s_a = \alpha_a + \beta_a\frac{2z}{1+z^2} + \gamma_a\frac{1-z^2}{1+z^2}$ which can be solved for z. By substituting $tan\left(\frac{\omega t}{2}\right) = z$ in the quadratic formula, we get

$$t = \frac{2\arctan\left(\frac{\beta_a \pm \sqrt{\beta_a^2 - (s_a - \alpha_a + \gamma_a)(s_a - \alpha_a - \gamma_a)}}{s_a - \alpha_a + \gamma_a}\right)}{\omega} + \frac{2\pi n}{\omega},$$

where $n \in \mathbb{Z}$, such that α_a, β_a, and γ_a are defined in Lemma 6. □

Proof of Proposition 3

Proof. 1. Consider EXP and denote its optimal solution (s_a^*, s_b^*). We now derive the bounds for s_a^*. There are two scenarios to consider: $q_a(t_0)$ is overloaded and $q_a(t_0)$ is underloaded.

If $q_a(t_0)$ is overloaded, then $q_a(t_0) > s_a^*$ and $q_a(t_0)$ is an upper bound. Alternatively, if $q_a(t_0)$ is underloaded, then $q_a(t)$ must switch to an overloaded regime at some time t because s_a^* is optimal and, thus, $\bar{w}_0(s_a^*) \geq 0$. Hence, the maximum of $q_{a_2}(t)$ is an upper bound of s_a^*.

The equation for $q_{a_2}(t)$ is given by (A.11). Because both exponential terms are decreasing in t, $q_{a_2}(t)$ is either decreasing or has a global maximum. The extremum of $q_{a_2}(t)$ is given by taking its first derivative, equating it to zero, and solving for t. This gives

$$t^* = \frac{log\left(-\mu_a\left[q_{a_2}(t_0) - \frac{\lambda_0}{\mu_a} - \frac{\lambda_1}{\mu_a - \delta}e^{-\delta t_0}\right]e^{\mu_a t_0}\right) - log\left(\frac{\delta\lambda_1}{\mu_a - \delta}\right)}{\mu_a - \delta}.$$

Thus, $s_a^* \leq q_a(t_0) \vee q_{a_2}(t^*)$. For the lower bound, because $q_a(t)$ cannot be overloaded indefinitely, it eventually switches to the underloaded regime. The minimum of $q_{a_2}(t)$ coincides with this equilibrium point, $\frac{\lambda_0}{\mu_a}$. As a result, we have that $\frac{\lambda_0}{\mu_a} \leq s_a^* \leq q_a(t_0) \vee q_{a_2}(t^*)$.

2. If $\mathcal{S}_{exp}$ denotes the search region of EXP, the corresponding solution range S_{exp} becomes:

$$S_{exp} = \left[q_{a_2}(t_0) - \frac{\lambda_0}{\mu_a} - \frac{\lambda_1}{\mu_a - \delta}e^{-\delta t_0}\right]e^{-\mu_a(t^* - t_0)} + \frac{\lambda_1}{\mu_a - \delta}e^{-\delta t^*}. \tag{A.24}$$

We now show that S_{exp} does not become unbounded as the values of the parameters increase. Clearly, S_{exp} can get large with λ_0 and/or λ_1. Thus, we consider a limit of S_{exp} as $\lambda_1 \to \infty$. For this analysis, we let $t_0 = 0$ as we determine the bounds of the solution space at the beginning of the time horizon, and assume $q_{a_2}(t_0) < s_a^*$ is a constant. Because λ_1 and λ_0 may change together, without loss of generality, we set $\lambda_0 = h(\lambda_1) \in \mathbb{C}^1$ which represents any continuous, smooth non-linear function of λ_1. In order to evaluate the limit of S_{exp}, we first determine what happens with t^* as λ_1 increases, i.e.,

$$\lim_{\lambda_1 \to \infty} t^* = \lim_{\lambda_1 \to \infty} \left[\frac{log\left(-\mu_a \left[q_{a_2}(0) - \frac{\lambda_0}{\mu_a} - \frac{\lambda_1}{\mu_a - \delta}\right]\right) - log\left(\frac{\delta\lambda_1}{\mu_a - \delta}\right)}{\mu_a - \delta} \right].$$

This is equivalent to:

$$\lim_{\lambda_1 \to \infty} t^* = \lim_{\lambda_1 \to \infty} \frac{log\left[\left(-\mu_a q_{a_2}(0) + \lambda_0 + \frac{\mu_a \lambda_1}{\mu_a - \delta}\right) \frac{\mu_a - \delta}{\delta \lambda_1}\right]}{\mu_a - \delta}.$$

By L'Hospital's Rule, the expression under the natural logarithm goes to $L := \frac{\dot{h}(\lambda_1)(\mu_a - \delta) + \mu_a}{\delta}$ as $\lambda_1 \to \infty$. Observe that because $\lim_{\lambda_1 \to \infty} t^* \geq 0$ and $\mu_a > \delta$, without loss of generality, $\lim_{\lambda_1 \to \infty} \dot{h}(\lambda_1) \geq 0$. Thus, λ_0 is increasing in λ_1 as $\lambda_1 \to \infty$. Further, $L = \infty$ because the first derivative of $h(\lambda_1)$ is a function of λ_1 and it is asymptotically positive. When $L = \infty$, $t^* \to \infty$ as $\lambda_1 \to \infty$. Thus, the limit of (A.24) as $\lambda_1 \to \infty$ is 0: $\lim_{\lambda_1 \to \infty} \frac{\lambda_1}{\mu_a - \delta} e^{-\delta t^*} = 0$ and $\lim_{\lambda_1 \to \infty} \frac{\lambda_1}{\mu_a - \delta} e^{-\mu_a t^*} = 0$ because an exponent reaches infinity faster than a polynomial; $\lim_{\lambda_1 \to \infty} \frac{\lambda_0}{\mu_a} e^{-\mu_a (t^*)} = 0$ because S_{exp} is non-negative, which implies that λ_0 cannot reach infinity faster than $\mu_a e^{\mu_a (t^*)}$. We conclude that if $\lambda_0 \to \infty$ as $\lambda_1 \to \infty$, then $|S_{exp}| \to 0$. □

Proof of Proposition 4

Proof. 1. Consider SINE and denote its optimal solution by (s_a^*, s_b^*). We now determine the bounds of s_a^*. Contrary to Proposition 3, we only require that the overloaded and underloaded regimes are reached periodically as time gets large. Thus, the maximum and minimum of $q_a^*(t)$ in the underloaded regime are the upper and lower bound of s_a^*, respectively. Denote the underloaded version of $q_a^*(t)$ by $\underline{q}_a^*(t)$. The expression for its minimum has been derived in (A.21). As per Lemma 7, the maximum of $\underline{q}_a^*(t)$ is reached at time points

$$\omega t^* = \arctan\left(\frac{\beta_a}{\gamma_a}\right) + n\pi,$$

where $n \in \{1, 3, 5, \ldots\}$, so that

$$\max \underline{q}_a^*(t) = \frac{\bar{\lambda}\sigma}{\mu_a^2 + \omega^2}\left(\omega cos\left[arctan\left(\frac{\mu_a}{\omega}\right)\right] + \mu_a sin\left[arctan\left(\frac{\mu_a}{\omega}\right)\right]\right) + \frac{\bar{\lambda}}{\mu_a} \tag{A.25}$$

and

$$\min \underline{q}_a^*(t) \leq s_a^* \leq \max \underline{q}_a^*(t).$$

2. If $\mathcal{S}_{sine}$ denotes the search region of SINE, then the corresponding solution range S_{sine} becomes:

$$S_{sine} = \frac{2\bar{\lambda}\sigma}{\mu_a^2 + \omega^2}\left(\omega cos\left[arctan\left(\frac{\mu_a}{\omega}\right)\right] + \mu_a sin\left[arctan\left(\frac{\mu_a}{\omega}\right)\right]\right).$$

We show that S_{sine} does not become unbounded as values of the parameters increase. Because $\bar{\lambda}$ is the only parameter that can cause the solution range to become unbounded, we evaluate the limit of S_{sine} as $\bar{\lambda} \to \infty$. It clearly follows that if $\sigma \to 0$ as fast as $\bar{\lambda} \to \infty$, $\lim_{\bar{\lambda}\to\infty} S_{sine} = 0$. Note that we assume $\infty \cdot 0 := 0$. □

Appendix B

Chapter 2: Implementation Details and Proofs of Statements

Case Study: Arrival Function and Asymptotic OLF

Our data set represents arrival counts per second over a typical 24-hour period, i.e., 86400 data points in total. We transform this data to average instantaneous counts at the beginning of each one second period. This can be done by converting counts per second to counts per millisecond. We then fit the resulting data of instantaneous arrival counts in a curve fitting toolbox of Matlab R2015a. Our best fitting function $\lambda(t) \in \mathbb{C}^1$ is an eighth order polynomial. Consequently, the asymptotic OLF, a particular solution to (3.4), is also a polynomial function of eighth order such that

$$\lambda(t) = \sum_{k=0}^{8} \lambda_k t^k, \quad \upsilon(t) = \sum_{k=0}^{8} \upsilon_k t^k,$$

where $\lambda_k \in \mathbb{R}$ and $\upsilon_k \in \mathbb{R}$ for $k \in \{0, 1, \ldots, 8\}$ are polynomial coefficients. Because λ_k and υ_k have many digits after the decimal point, they quickly accumulate a rounding error, and can only be reported in a .mat file attachment to ensure reproducibility. We make a mat file containing these coefficients available in a shared Google Drive folder.

k	λ_k	υ_k
0	23.3155319631730	0.582912116248628
1	-0.0381052644866565	-0.000952686772086839
2	$4.41275772705046e-05$	$1.10319840853815e-06$
3	$-4.78759197337980e-09$	$-1.19690340518401e-10$
4	$2.16472930776578e-13$	$5.41183906337450e-15$
5	$-5.05405757472458e-18$	$-1.26351680268878e-19$
6	$6.42401295446502e-23$	$1.60600509026714e-24$
7	$-4.23234261180524e-28$	$-1.05808621985254e-29$
8	$1.13380246953280e-33$	$2.83450617383199e-35$

Case Study: Tables

Tables B.1 - B.3 compliment **Figures** 3.3b,3.4a,3.4b, respectively. Note: bolded rows match groups of bars in a corresponding figure from left to right.

$\boldsymbol{\alpha}$, $\boldsymbol{\beta}$: values of the SLA constants α and β.

$\boldsymbol{s^*}$: the optimal capacity level obtained by solving GVAR.

$\hat{\mathbb{E}}[\boldsymbol{W}]$: the expected waiting time (in seconds) before accessing service per Definition 2.

$\hat{\mathbb{P}}(\boldsymbol{W} > \boldsymbol{0})$: the long-term probability of waiting, i.e., accessing service after at least 2 attempts.

$\boldsymbol{C}$: the annual operational cost of capacity s^*.

Table B.1: $\beta = 0.75$, complimentary to Figure 3.3b

α	β	s^*	$\hat{\mathbb{E}}[W]$	$\hat{\mathbb{P}}(W > 0)$	C
0.0100	**0.7500**	**35.2508**	**0.0000**	**0.0000**	**9263.8984**
0.0123	0.7500	35.2193	0.0000	0.0000	9255.6309
0.0473	0.7500	33.9180	0.0000	0.0000	8913.6599
0.0823	**0.7500**	**31.5919**	**0.0000**	**0.0000**	**8302.3398**
0.1173	0.7500	31.3327	0.0000	0.0000	8234.2274
0.1522	0.7500	30.7493	0.0000	0.0000	8080.9041
0.1872	**0.7500**	**29.9313**	**0.0000**	**0.0000**	**7865.9509**
0.2222	0.7500	28.9246	0.0000	0.0000	7601.3931
0.2572	0.7500	27.7635	0.0000	0.0000	7296.2498
0.2922	**0.7500**	**26.4854**	**0.0000**	**0.0000**	**6960.3631**
0.3272	0.7500	25.1209	871.8314	0.1010	6601.7856
0.3622	**0.7500**	**23.6962**	**1466.1979**	**0.1448**	**6227.3614**

Table B.2: $\beta = 0.80$, complimentary to Figure 3.4a

α	β	s^*	$\hat{\mathbb{E}}[W]$	$\hat{\mathbb{P}}(W > 0)$	C
0.0100	**0.8000**	**33.0486**	**0.0000**	**0.0000**	**8685.1782**
0.0423	0.8000	32.0648	0.0000	0.0000	8426.6331
0.0824	**0.8000**	**29.6173**	**0.0000**	**0.0000**	**7783.4175**
0.1225	0.8000	29.3079	0.0000	0.0000	7702.1055
0.1627	**0.8000**	**28.6215**	**0.0000**	**0.0000**	**7521.7252**
0.2028	0.8000	27.6577	0.0000	0.0000	7268.4411
0.2430	**0.8000**	**26.4854**	**0.0000**	**0.0000**	**6960.3631**
0.2831	0.8000	25.1495	793.0319	0.1000	6609.3014
0.3232	**0.8000**	**23.6962**	**1377.9107**	**0.1448**	**6227.3614**

Table B.3: $\beta = 0.85$, complimentary to Figure 3.4b

α	β	s^*	$\hat{\mathbb{E}}[W]$	$\hat{\mathbb{P}}(W > 0)$	C
0.0100	**0.8500**	**31.1039**	**0.0000**	**0.0000**	**8174.1006**
0.0362	0.8500	30.4347	0.0000	0.0000	7998.2298
0.0855	**0.8500**	**27.8693**	**0.0000**	**0.0000**	**7324.0584**
0.1347	0.8500	27.4203	0.0000	0.0000	7206.0596
0.1839	**0.8500**	**26.4854**	**0.0000**	**0.0000**	**6960.3631**
0.2331	0.8500	25.2125	698.5645	0.0976	6625.8363
0.2823	**0.8500**	**23.6962**	**1287.2927**	**0.1448**	**6227.3614**

Proofs of the results

Proof of Lemma 9

Proof. (i) Fixing the phase ϕ to 0, we plug (3.12) into (3.4) to obtain $\dot{u}(t) = \bar{\lambda} + \bar{\lambda}\sigma \sin(\omega t) - \mu u(t)$. This is an ordinary linear nonhomogeneous differential equation. We use the method of undetermined coefficients to obtain a closed-form expression for $u(t)$

$$u(t) = u_0 e^{-\mu t} + \upsilon_1 + \upsilon_2 \cos(\omega t) + \upsilon_3 \sin(\omega t)$$

where

$$u_0 \in \mathbb{R}, \quad \upsilon_1 = \frac{\bar{\lambda}}{\mu}, \quad \upsilon_2 = -\frac{\omega \bar{\lambda}\sigma}{\omega^2 + \mu_a^2}, \quad \upsilon_3 = \frac{\mu \bar{\lambda}\sigma}{\omega^2 + \mu^2}.$$

(ii) As $t \to \infty$, $u_0 e^{-\mu t}$ approaches 0. Thus, $\upsilon(t) = \upsilon_1 + \upsilon_2 \cos(\omega t) + \upsilon_3 \sin(\omega t)$.

(iii) This follows from Lemma 8 in Furman et al. (2019). □

Proof of Proposition 5

Proof. Define a piecewise smooth function $z_j(t, \boldsymbol{r}_j, s)$ for $j \in \{1, 2, \ldots, J-1\}$ to be the modified OLF of a system with j orbits as in (3.9)-(3.10). We need to show that $y(t; \boldsymbol{r}, s) = z_{J-1}(t; \boldsymbol{r}_{J-1}, s) - \bar{\upsilon}(t; s)$. Observe that $\bar{\upsilon}(t; s)$ represents the workload that accesses service immediately upon arrival, i.e., attempts service one time in total. Also, $(\upsilon(t - r_1^{-1}) - s)^+$ is a horizontal shift r_1^{-1} of the workload that requires at least two service attempts and equals $z_1(t; \boldsymbol{r}_1, s) - \bar{\upsilon}(t; s)$. By induction, $z_j(t; \boldsymbol{r}_j, s) - \bar{\upsilon}(t; s)$ describes the aggregate workload that requires at least $j+1$ attempts to get serviced and re-enters the system after at most j sequential horizontal shifts $r_1^{-1}, r_2^{-1}, \ldots, r_j^{-1}$. Therefore, $z_{J-1}(t; \boldsymbol{r}_{J-1}, s) - \bar{\upsilon}(t; s) = y(t; \boldsymbol{r}, s)$. By (3.8), we also have $z_{J-1}(t; \boldsymbol{r}_{J-1}, s) = z(t; \boldsymbol{r}, s)$. □

Proof of Lemma 10

Proof. We prove this result by contradiction. Let there exist such s and r_1 that constraint (3.17) is satisfied. Then, $(\upsilon(t - r_1^{-1}) - s)^+ = 0, \forall t \in \{\upsilon(t) \geq \beta s\}$ and $(\upsilon(t - r_1^{-1}) - s)^+ \geq 0, \forall t \in \{\upsilon(t) < \beta s\}$. That is, fluid does not re-enter the system when $\upsilon(t) \geq \beta s$. Assume that $e_\Omega(s) > d_\Omega(\beta s)$, then, the cavity of $\upsilon(t)$ under βs level cannot contain the entire excess fluid and $\exists t \in \{\upsilon(t) \geq \beta s\}$ such that $(\upsilon(t - r_1^{-1}) - s)^+ \neq 0$. Thus, we acheive a contradiction and $e_\Omega(s) \leq d_\Omega(\beta s)$ must hold. □

Proof of Proposition 6

Proof. Let $\upsilon(t)$ be defined as in Lemma 9. We first show that there exists a unique value of service capacity s and retrial rate r_1 that optimally solves $\min\limits_{s \in \mathbb{R}_{>0}, r_j \in \mathbb{R}_{\geq 0}} V(z(t; \boldsymbol{r}, s))$ and also maximizes throughput. Let $z_j(t; \boldsymbol{r}_j, s)$ be defined as in Proposition 5. Because $\upsilon(t)$ has a reference line at υ_1 level and $\exists \delta > 0$, such that $\upsilon(t) - \upsilon_1 = \upsilon_1 - \upsilon(t + \delta^{-1})$ for any $t \in \left[0, \frac{\Omega}{2}\right]$, $V(z_1(t; \delta, \upsilon_1)) = 0$. Thus, $z_1(t; \delta, \upsilon_1) \wedge \upsilon_1 = z_1(t; \delta, \upsilon_1)$ and $(z_1(t; \delta, \upsilon_1) - \upsilon_1)^+ = 0$ for $\forall t$. We have that $z_2(t; \delta, r_2, \upsilon_1) = \ldots = z_{J-1}(t; \delta, r_2, \ldots, r_{J-1}, \upsilon_1) = z_1(t; \delta, \upsilon_1)$ for $r_j \geq 0$ and $j \in \{2, 3, \ldots, J-1\}$.

Second, we show that having one orbit in the system is sufficient to serve all clients and to satisfy (3.16). Observe that s^*, an optimal solution to VAR, is always greater than υ_1. By definition, s^* satisfies (3.16), i.e., there exist such time t and optimal retrial rates $\boldsymbol{r}^*$ that $z(t; \boldsymbol{r}^*, s^*) < \beta s$. However, $z(t; \delta, \upsilon_1) > \beta s$ for any t. Thus, to satisfy (3.16) more capacity than υ_1 is required.

Because capacity υ_1 is sufficient to serve the entire workload after at most two service attempts, any $s > \upsilon_1$ serves all jobs by employing a single orbit, i.e., there exists such δ^* that $z_2(t; \delta^*, r_2, s_\alpha) = \ldots = z_{J-1}(t; \delta^*, r_2, \ldots, r_{J-1}, s_\alpha) = z_1(t; \delta^*, s_\alpha)$ for $r_j \geq 0$ and $j \in \{2, 3, \ldots, J-1\}$.

Third, we show that $V(z(t; \boldsymbol{r}, s))$ is convex in s. For any $s \in [0, \upsilon_1]$, $\max(z(t; \boldsymbol{r}, s))$ decreases in s while $\min(z(t; \boldsymbol{r}, s))$ increases in s. Then, $V(z(t; \boldsymbol{r}, s))$ is decreasing. Similarly, for any $s \in [\upsilon_1, \max(\upsilon(t))]$, $\max(z(t; \boldsymbol{r}, s))$ increases in s while $\min(z(t; \boldsymbol{r}, s))$ decreases in s. Then, $V(z(t; \boldsymbol{r}, s))$ is increasing. Thus, $V(z(t; \boldsymbol{r}, s))$ is convex in s over $[0, \max(\upsilon(t))]$.

Fourth, due to the aforementioned convexity of $V(z(t; \boldsymbol{r}, s))$, s^* is the minimum capacity that satisfies (3.16) and (3.17) at the same time. By Lemma 10, (3.17) implies $e_\Omega(s) \leq d_\Omega(\beta s)$. Further, because $s^* \geq \upsilon_1$, $V(z(t; \boldsymbol{r}, s))$ is increasing in $s \geq s^*$ and $s \leq \max\{\upsilon(t)\}$. Equivalently, s^* is the minimum capacity ensuring that at least one of the constraints (3.16) and (3.18) is binding. Thus, s_1, the solution to equation $\frac{t_2(n,\beta s) - t_1(n,\beta s) + |\mathcal{Z}(n;\beta s)|}{\Omega} = \alpha$ is the optimal solution to VAR with (3.16) only while s_2, the solution to equation $e_\Omega(s) = d_\Omega(\beta s)$, is the optimal solution to VAR with (3.17) only. Thus, $s^* = \max\{s_1, s_2\}$ can be computed by solving a single non-linear equation and solves VAR optimally.

Finally, we select such retrial rate r_1^* that satisfies (3.17). Let $t_2(n, s) > t_1(n, s) > t_0(n, s) > 0$ be consecutive solutions to the equation $\upsilon(t) = s$ for $n \in \mathbb{Z}$ where $\dot{\upsilon}(t_0(n, s)) < 0$, $\dot{\upsilon}(t_1(n, s)) > 0$ and $\dot{\upsilon}(t_2(n, s)) < 0$, respectively. For any capacity $s \geq \upsilon_1$, to ensure that (3.17) is satisfied, the following constraints hold by construction

$$(t_1(n, \beta s) - t_0(n, s))^{-1} \leq r_1 \leq (t_2(n, \beta s) - t_1(n, s))^{-1}. \tag{B.1}$$

Writing (B.1) for s^* results in (3.19) while $r_j^* \geq 0$ for $j \in \{2, 3, \ldots, J-1\}$. □

Corollary 4 (Solution to VAR)**.** *Let s^* defined as in Proposition 6 be the optimal solution to VAR. Also, let capacity level υ_1 be the optimal solution to the unconstrained VAR. Then, there exists retrial rate δ^* that solves VAR optimally and it equals $t_2(n, l) - t_1(n, s^*)$, where $l = \upsilon_1 - (s^* - \upsilon_1)$.*

Proof of Corollary 4

Proof. Let capacity level s^* solve VAR optimally. Then, for some $\delta^* > 0$, similar to Proposition 6, because $s^* \geq \upsilon_1$, there exists a unique level $l \leq \beta s^*$ such that $z_1(t; \delta^*, s^*)$ fills the cavity below βs^*

with fluid up to this level exactly. Thus, $\min(z_1(t;\delta^*,s^*)) = l$ and $\min(z_1(t;r_1,s^*)) < l$ for any $r_1 \neq \delta^*$. Hence, given that $\max(z_1(t;\boldsymbol{r}_1,s^*)) = s^*$ remains fixed, $V(z_1(t;\delta^*,s^*))) < V(z_1(t;\boldsymbol{r}_1,s^*)))$ by definition of the total variation of a function. Hence, s^* and δ^* are jointly optimal. Let $t_2(n,l)$ be a solution to $v(t) = l$, such that $\dot{v}(t_2(n,l)) < 0$. Notice that because $v(t)$ has point symmetry, $l = v_1 - (s^* - v_1)$. Then, we have that $\delta_1^* = t_2(n,l) - t_1(n,s^*)$ by construction. □

Proof of Proposition 7

Proof. By construction, $x(t) = z(t;\boldsymbol{r}^*,s^*)$, and thus (3.16) and (3.5) are equivalent. We also have that (3.17) and (3.6) are equivalent for $j = 1$ and (3.7) always hold for $j \in \{2,3,\ldots,J-1\}$ as these orbits are empty. The value of r_1 that solves VAR optimally is unique and it satisfies (3.19) given capacity s^* which is implied by (3.17). Therefore, because s^* is the minimal capacity that satisfies (3.5)-(3.6) and serves all jobs at the same time, the optimal solution to VAR is the optimal solution to OPT. □

Proof of Lemma 11

Proof. (i) Let $\lambda(t)$ admit a polynomial representation and $\lambda_\Omega(t)$ be defined as in (3.20). Then, the solution to $\dot{u}(t) = \lambda_\Omega(t) - \mu u(t)$ over period Ω becomes $u(t) = u_0 e^{-\mu t} + v(t)$, where $u_0 \in \mathbb{R}$ and $v(t)$ also admits a polynomial representation. Because the term $\lambda_\Omega(t)$ is periodic, $v(t)$ inherits this property. Thus, for any time $t \geq 0$,

$$u(t) = u_0 e^{-\mu t} + v(t \mod \Omega).$$

Without loss of generality we denote $u(t)$ by $u_\Omega(t)$.

(ii) As $t \to \infty$, $u_0 e^{-\mu t}$ approaches 0. Thus, $u_\Omega(t)$ approaches a periodic orbit asymptotically, i.e., $v_\Omega(t) = v(t \mod \Omega)$. □

Proof of Corollary 2

Proof. Let s_α be the optimal solution to GVAR with (3.22) only. We first observe that because each cavity below level βs is large enough to accommodate the entire workload, i.e., $|\mathcal{Z}_m(n;\beta s)| = 0$, $\exists s$

and r_1 such that $v(t) - s \leq s - v\left(t + \frac{1}{r_1}\right)$ holds for all n, m and $t \in \bigcup_{m=1}^{M} \mathcal{A}_m(n, s)$. Therefore, all jobs get serviced after at most two attempts and the left-hand side of (3.22) is a function of $v(t)$ whose closed-form expression is available so that constraint (3.25) is non-binding. Because $|\mathcal{Z}_m(n; \beta s)| = 0$, we have that constraints (3.23)-(3.24) are also non-binding. Further, because the left-hand side of (3.22) decreases in s while $V(z(t; \boldsymbol{r}, s))$ increases in s, by the convexity argument in Proposition 6, constraint (3.22) is always binding. Then, $s^* = s_\alpha$ is the optimal solution to GVAR, i.e., it is the minimum capacity that satisfies constraint (3.22).

Then, we select such r_1^* that also satisfies (3.23). Observe that, for any capacity s, to ensure that (3.23) is satisfied, the following constraints hold by construction

$$\frac{1}{r_1} \geq \max\{|\mathcal{A}_m(n, \beta s) \cap \mathcal{B}_m(n, s)| + |\mathcal{A}_m(n, s)|\}_{m=1}^{M}, \tag{B.2}$$

$$\frac{1}{r_1} \leq \min\{|\mathcal{A}_m(n, \beta s) \cap \mathcal{B}_m(n, s)| + |\mathcal{B}_m(n, \beta s)|\}_{m=1}^{M}. \tag{B.3}$$

Constraints (3.24) are always satisfied as these orbits are empty under the assumption of Corollary 2. Writing (B.2) - (B.3) for s^* results in (3.26) - (3.27) while $r_j^* \geq 0$ for $j \in \{2, 3, \ldots, J-1\}$. □

Proof of Corollary 3

Proof. By construction, $x(t) = z(t; \boldsymbol{r}^*, s^*)$ and, thus, (3.22) and (3.5) are equivalent. We also have that (3.23)-(3.24) and (3.6)-(3.7) are equivalent. Further, given the penalty term of the objective function in OPT, the optimal capacity s^* serves the entire demand. This property is ensured by constraint (3.25). By the convexity of $V(t; \boldsymbol{r}, s)$ in s, the optimal value of s that solves GVAR is either the minimal capacity that satisfies (3.22)-(3.24) or the minimal capacity that satisfies (3.25), i.e., it equals s^*. Similar to Corollary 2, because the unique value of r_1 that solves GVAR optimally is an optimal solution to OPT, the optimal solution to GVAR is the optimal solution to OPT. □

Appendix C

Chapter 3: Implementation Details and Proofs of Statements

Appendices and Proofs of Statements

Proof of Lemma 12

For $\sigma_i > 0$ and $T > \sigma_i$ for all i, we apply the conditional expectation operator to (4.3). Then, given (4.2), and by the linearity of expectation, (4.6) holds. $\square$

Clustering Procedure

We cluster patients based on 16 variables which include a patient's length-of-stay in the hospital (days) and their average daily count of 15 medical interactions as per Column 1 in Table C.2. For example, a patient with a length-of-stay equal to 3.5 days and 12 vital sign measurements will be assigned an average daily count of $12/3.5 = 3.4$ measurements of vital signs.

To improve the quality of our clustering procedure, we employ Uniform Manifold Approximation and Projection technique (UMAP) (McInnes et al. 2018) as a pre-processing step. Contrary to other non-linear projection methods (t-SNE or Isomap, for instance), it does not favor the preservation of local distances over global distance. That is, using UMAP as a pre-processing step for clustering preserves both the local (dissimilarities within clusters) and global (dissimilarities between clusters) structure of the data set. Further, the algorithm is less computationally intensive than t-SNE, for instance, and in contrast to linear projection techniques such as Principal Component Analysis (PCA), does not attempt to construct multidimensional vectors to recreate the location of each data point. Thus, it is not vulnerable to 20% - 30% loss in representative accuracy.

Because we do not aim to predict cluster membership for future patients and would like to cluster all existing patients without exceptions, we choose a standard k-means approach. To ensure the stability of the clusters, we initialize the procedure with 25 random starting partitions (see nstart option for the kmeans function in the R documentation). Contrary to our needs, density based techniques, (HDBSCAN, for instance) may consider some of the data points as noise.

Table C.1: Within cluster variation as a function of the number of clusters used.

Number of Clusters, k	Total Squared Error Within Clusters
1	290,250
2	140,491.9
3	86,060.95
4	56,175.68
5	42,133.06
6	33,761.11
7	25535.31
8	20880.25
9	17563.34
10	15222.13

Table C.2: Average PPE usage per patient-practitioner interaction.

Interaction Types, j	Gowns, $u_{1,j}$	Gloves, $u_{2,j}$	Surgical Masks, $u_{3,j}$	N95 Masks, $u_{4,j}$	Shields, $u_{5,j}$	Bouffants, $u_{6,j}$	Boot Covers, $u_{7,j}$
Vital signs measurement	0	1	0	0	0	0	0
Medication administration	0	1	0	0	0	0	0
Lab Test Collection	0	1	0	0	0	0	0
X-ray	0	2	0	0	0	0	0
CT	0	2	0	0	0	0	0
MRI	0	2	0	0	0	0	0
Ultrasound	0	2	0	0	0	0	0
Nuclear Medicine	0	1	0	0	0	0	0
Interventional Radiology	3.5	3.5	0	3.5	0	3.5	3.5
Transthoracic Echocardiography (TTE)	0	1	0	0	0	0	0
Transesophageal Echocardiography (TEE)	3	3	3	3	0	3	3
Bronchoscopy	4	4	4	4	0	4	4
Dialysis	0	1	0	0	0	0	0
Surgical Procedure	5.5	5.5	4	2	0	5.5	5.5
Room Transfer	0	1.5	0	0	0	0	0

We use the total squared error within clusters as a single aggregate measure of similarity amongst patients. This is because a sample variance estimate is dependent on the size of the cluster. Table C.1 presents the within cluster variation as a function of the number of clusters that are used. Due to the multi-dimensional nature of the data, the clustering technique aims to reduce the variation amongst all variables at the same time rather than focusing on one of them specifically.

Estimation of the Number of Medical Interactions Per Day as well as PPE Usage

The first column in Table C.2 represents the most common types of medical interactions between patients and practitioners for individuals admitted to the GIM service at St. Micheal's hospital. In columns 2-8, we display the count of type $n \in \{1, 2, \ldots, 7\}$ PPE used during each type of interaction. While the values in column 1 are obtained by analyzing the types of interactions in the data set, the values in columns 2-8 were obtained by conducting semi-structured interviews with various medical practitioners in each sub-speciality of our partner hospital and summarizing their responses.

More specifically, we engaged key stakeholders from clinical departments throughout St. Michael's Hospital. They included a nurse manager for the general internal medicine (GIM) department, a medical imaging manager, an echocardiography team leader and a cardiac sonographer, a dialysis charge nurse, a gastroenterologist, a respirologist, and an anesthesiologist. The semi-structured interviews were conducted with each individual and process-mapping techniques were used to understand the workflow, number of patient interactions, personnel needs, and the PPE usage per episode of patient care. For elective and non-elective surgeries, we used common surgical procedures conducted on GIM patients, including laparoscopic intra-abdominal surgeries or vascular procedures such as amputation, to inform the model. These interviews provided pragmatic estimates of PPE usage and helped to estimate the number of patient interactions on a daily basis.