Dynamic Server Assignment in Multiclass Queues with Shifts, with Application to Nurse Staffing in Emergency Departments

Carri W. Chan
Division of Decision, Risk, and Operations, Columbia Business School,
New York, NY 10027, cwchan@columbia.edu

Michael Huang
Marshall School of Business, University of Southern California,
Los Angeles, CA 90089, huan076@usc.edu

Vahid Sarhangian
Department of Mechanical and Industrial Engineering, University of Toronto,
Toronto, ON M5S 3G8, sarhangian@mie.utoronto.ca

Many service systems are staffed by workers who work in *shifts*. In this work, we study the dynamic assignment of servers to different areas of a service system at the beginning of discrete time-intervals, i.e., shifts. The ability to reassign servers at discrete intervals, rather than continuously, introduces a *partial flexibility* that provides an opportunity for reducing the expected waiting time of customers. The problem is primarily motivated by an application to nurse staffing in emergency departments (EDs) where nurses can work in different areas of the ED, but their assignment can only be changed at the beginning of their shifts (typically 8-12 hours). To investigate the reassignment decision and its potential benefits, we consider a multiclass queueing system, where customers of each class differ in terms of their average service requirements and the holding cost incurred as they wait in the queues. We study a discrete-time fluid control problem to minimize transient holding costs over a finite horizon and show that an appropriate "translation" of the solution to the fluid control problem is asymptotically optimal for the original stochastic system. Through analysis of the fluid control problem we further obtain insights on the structure of "good" policies in presence of the shift constraint. Leveraging these insights, we develop heuristic policies and use simulation to demonstrate their effectiveness in systems with dynamics often observed in EDs. We find that in a parameter regime relevant to our motivating application, the partial flexibility introduced by reassigning servers at the beginning of shifts can substantially reduce the expected cost of the system – by 10–50% in some parameter regimes – compared to the status-quo, dedicated staffing.

Key words: Stochastic Control; Multiclass Queues; Fluid Models; Partial Flexibility; Nurse Staffing

1. Introduction

In a number of service settings, servers (staff) are assigned to specific tasks (or classes of tasks) for fixed, finite intervals – shifts. Much of the prior literature has focused on optimizing the number and type of workers allocated to each shift as well as optimizing the timing of the shifts (e.g., Brusco et al. 1995, Ingolfsson et al. 2002, Atlason et al. 2008). In contrast, we assume that the schedule and number of servers per shift is given and focus on the dynamic assignment of servers to different areas in a service system at the beginning of each shift.

In the simplest setup, which we refer to as *dedicated staffing*, the system manager assigns a fixed number of servers to each area which remains static across all shifts. Due to fluctuations in customer arrivals and service times, there are likely to be times when one area is congested while there are servers idling in other areas. Ideally, one could assign the servers dynamically as soon as imbalances in the congestion warrant it. Unfortunately, this *full flexibility* (in terms of reassignment of servers to areas) is not always possible. This may be because the tasks are highly specialized and/or the areas are far apart, making the switch from one area (or task) to another during a shift impractical. However, because staff (servers) tend to rotate in and out at the beginning of each shift, it may be possible to reassign servers at these discrete intervals. In this work, we investigate the potential benefits of a *partially flexible* design that allows for adjusting the server assignments at the beginning of shifts. In this setting, the total number of servers per shift is known in advance, but the number assigned to each area may change when the servers arrive for their shift.

Our main motivation comes from Emergency Department (ED) nurse staffing, and was initiated as part of a collaboration with a large academic hospital in New York City. Overcrowding and long delays for patients is prevalent in EDs with potentially serious ramifications in quality of care. As pointed out by Green (2010), nurses are the primary managers and caretakers of patients in the ED. Hence, their unavailability can be a major contributor to delays experienced by patients. With nursing costs that comprise a significant fraction of hospital budgets, it is of growing interest to explore innovative ways to effectively utilize nursing resources with the goal of reducing patient waiting times.

In our partner hospital, the ED is partitioned into 4 areas that are physically separated from each other. New patients are assigned to different areas primarily based on their medical needs and Emergency Severity Index (ESI) level, which is a coarse assessment of their severity determined at triage. Nurses work in 12-hour shifts and are assigned to a certain area weeks in advance. Additionally, the number of nurses assigned to different areas remains fixed throughout time. This structure is convenient for nurses as they know in advance which shift and area they are assigned to. Although nurses can work in any of the areas (and are actively rotated across different shifts), they typically cannot be reassigned mid-shift since they have to setup their workspace and due to concerns with continuity of care. Nevertheless, from time-to-time when imbalances among areas occur, the nursing director may reassign a nurse from a less congested area to a heavily loaded one at the beginning of a new shift. This happens very infrequently. In this work, we investigate this reassignment decision and its potential benefits. We note that although our partner ED maybe among the few with 4 areas, many EDs (e.g., that in Song et al. 2015) have 2 areas, with one area dedicated to lower acuity patients often referred to as the "Fast Track"[1].

Although our work is primarily motivated by the nurse staffing application described above, our model and analysis can provide insights for other service systems with a similar structure, i.e., services that are provided in separate locations (most likely in close proximity) by cross-trained staff that work in shifts which only allow for periodic reassignments. Examples include checkout-lines in large department stores and security checkpoints at airports. See Martonosi (2011) for a discussion of an application in airport security check-lines. Additionally, the partially flexible design could also be beneficial for services where there are diseconomies associated with full flexibility (pooling) of servers. For instance, Jouini et al. (2008) illustrates how migrating from a pooled system to a dedicated system where a team of servers in a call center is assigned to each customer class could increase throughput by allowing better workforce management and enabling efficiency gains due to the specialization and "focus" of servers (e.g. Staats and Gino 2012). In such cases, the partially flexible design allows for servers to focus on one customer class or task during each shift, while still potentially achieving some of the benefits of a pooled system through adjusting the assignment of servers to different classes at the beginning of each shift.

A queueing system with partial flexibility: To investigate the benefits of the partially flexible design, we consider a multiclass queueing system with multiple server pools. Each customer class is associated with a queue and server pool. Customer classes differ in terms of their average service requirement and the waiting cost incurred as they wait in the queues. The objective is to dynamically assign servers to the server pools in order to minimize the expected waiting cost of the system over a finite horizon, but under the restriction that the assignments can only be changed at the beginning of discrete intervals; i.e., at the start of each shift. We refer to this restriction as the *shift constraint*. The transient control of capacity assignment is of particular interest in EDs where surges in demand (e.g. due to flu outbreaks) can move the system far from its "normal" operations.

The underlying model is fairly standard in the queueing literature and can be viewed as a parallel server system (see, e.g., Dai and Tezcan 2011) with multiple server pools that are each capable of serving a subset of customer classes. Accordingly, the dynamic scheduling decision is to choose the next customer class to serve, each time a server becomes idle. The key feature differentiating our work from the literature is the presence of the shift constraint. We assume that servers are cross-trained and are capable of serving all customer classes, but can only serve a single customer class during each shift. We note that the discrete-time structure of control in our work bears similarities to the literature on discrete-review scheduling policies for single-server queues (e.g., Harrison 1998, Maglaras 2000) where the scheduling decision is made at the beginning of discrete intervals of chosen length due to a design choice selected by the system controller. In contrast, the restriction in our problem is due to the underlying primitive constraints of the service system and the length of the review period (shift length in our problem) cannot be "fine-tuned". As a result, our goal is to find policies that perform well for a given shift length, and to understand the benefits of dynamic control for different parameter regimes and shift lengths.

The shift constraint introduces a *partial* flexibility into our system, which is in contrast to two well known designs: dedicated and fully flexible staffing. Under dedicated staffing, the number of servers in each pool remains fixed in time. Hence, the (static) staffing problem is to find the number of servers for each pool that minimizes the total cost of the system. Under full flexibility, the assignment of servers can be adjusted at any point in time. In this case, the server assignment problem reduces to a well-known scheduling problem that is (assuming non-preemption) equivalent to choosing the next customer to serve from the set of customers waiting, each time a server becomes available. In this case, the myopic $c\mu$ policy that chooses the class with the highest $c\mu-$index (i.e., holding cost rate × service rate) is known to be optimal even under more general conditions and different cost-criteria (e.g. Van Mieghem 1995 and Mandelbaum and Stolyar 2004).

The control problem under the partially flexible design introduces new complexities. Even assuming a Markovian setting, the typical approach of exploiting a Markov Decision Process (MDP) formulation is both computationally and analytically hard. This is in large part due to more complex transition probabilities compared to the fully flexible design. Unlike the control problem under full flexibility where one can apply uniformization to obtain a discrete-time MDP with simple transition probabilities, under the partially flexible design the transition probabilities are dictated by the transient distributions of independent multiserver queues which are not available in closed form.

In addition to the above technical complexities, we note that the simple myopic $c\mu$ policy under the fully flexible design does not easily "translate" into useful policies for the problem under the partially flexible design. We find that myopic assignments could lead to "excessive" idleness due to the inability to reallocate over the duration of a shift, ultimately resulting in worse performance compared to dedicated staffing. However, when the tradeoff between myopic cost reduction and mid-shift idleness is properly taken into account, partial flexibility could result in significant reduction of expected waiting costs in practically relevant parameter regimes. Our main contributions can be summarized as follows:

- **Queueing model with shift constraint.** We propose a multiclass and multiserver queueing model which allows for dynamic server assignment at discrete intervals. The key component we consider is the dynamic assignment of servers under a new staffing constraint – the shift constraint – that arises in a number of service settings. Our goal is to develop an understanding of which features drive the reassignment decision as well as the potential benefits associated with the introduction of partial flexibility.

- **Fluid analysis and asymptotic optimality.** We propose an associated discrete-time fluid control problem and exploit a dynamic programming formulation to show structural properties of the optimal policy that provides *insights on the structure of "good" policies under the partially flexible design*. We propose a "translation" of the solution to the fluid control problem for the original stochastic system and prove its fluid-scale asymptotic optimality. Using simulation, we observe that *the benefit of partial flexibility* can be substantial compared to dedicated staffing when starting from large and/or imbalanced initial states. In addition, as long as the system does not quickly reach its steady-state during the shifts (i.e., the shift length is not "too long" compared to the time between service completions) there can also be significant benefits with respect to the long-run average cost of the system.

- **Model extensions and heuristics.** We consider extensions to our model to include time-varying arrivals and customer abandonment. We then leverage insights from the analysis of the fluid control problem to develop simple heuristic policies and demonstrate their effectiveness and investigate the impact of customer abandonment using simulation experiments. Further, and as a "proof of concept", we apply our heuristics to a simulation model of the system calibrated using data from our partner ED. These results suggest that the insights generated from our original model are robust and can be useful when considering more complex system dynamics.

1.1. Related Work

There is a large literature on shift scheduling (or rostering) for service systems; see Van den Bergh et al. (2013) for a recent review and Burke et al. (2004) for a review devoted to the nurse scheduling problem. This body of literature typically does not explicitly consider queueing dynamics. Accounting for queueing dynamics introduces complex stochastic interdependencies across shifts. Kolesar et al. (1975) and Green and Kolesar 1989 derive shift schedules for police patrol cars by assuming the system reaches stationarity in each shift. For the same application, Ingolfsson et al. (2002) directly take into account time-varying queueing dynamics in scheduling a workforce by using a genetic algorithm to search over potential schedules. Batt et al. (2019) empirically studies the effect of discrete work shifts on service rates and investigates the impact of hand-offs (from one provider to another) at the end of shifts. While we do not incorporate the end of shift effects in our model, we do consider how staffing decisions in one shift can impact the dynamics in future shifts. This is also done in Anderson and Gamarnik (2015) which uses a fluid approximation of a single-class queueing model to compare two intern scheduling policies in order to determine the length of the shifts and number of interns that work during each shift. In contrast to these prior works, we assume that the shifts and the number of staff available for each shift is given and use a queueing model to investigate their dynamic assignment to different areas of the service system at the beginning of each shift. Additionally, we take a different methodological approach in order to derive asymptotically optimal policies.

The queueing model considered in this paper consists of parallel server pools and multiple classes of customers. Parallel server queueing systems have been extensively used in the literature to model service processing networks such as call centers (e.g., Bassamboo et al. 2006) or healthcare systems (e.g., Dai and Shi 2019). When it comes to dynamic control, previous literature has focused on continuous-time routing (sending arrivals to queues) or scheduling (determining the next customer to be served by an idle server) decisions. In our work, since the control is restricted to the beginning of discrete shifts, the scheduling problem turns into a server assignment problem where servers in each pool can only serve customers in the corresponding queue but the number of servers assigned to the pools can be adjusted at the beginning of discrete time-intervals. There appear to be only a few examples in the literature considering discrete-time control of queueing systems. One example is Tirdad et al. (2016) which considers a hysteretic policy in discrete-time for controlling service capacity of a single-class time-varying multiserver loss system. Another example is a closely related study by Martonosi (2011) which considers a similar model to ours with two customer classes and identical holding costs, including a constant switching time for moving servers. The problem there is motivated by the possibility of moving servers among airport security check lines to reduce overall congestion. The author exploits an MDP formulation and proposes heuristic policies, but finds that using intuitive heuristics could hurt the performance of the system. In contrast, we study the problem using an asymptotic approach that allows us to design effective and robust policies, leads to insights on the structure of good policies, and identifies the parameter regimes where the dynamic server assignment leads to significant savings.

Due to the complexity of the control problems for parallel server systems, many papers have exploited asymptotic parameter regimes (e.g., Harrison 1998 or Harrison and Zeevi 2004, Gurvich and Whitt 2010). In this work we exploit a fluid asymptotic regime that is suitable for our problem as the resulting fluid models (approximately) capture the transient dynamics of the system during the shifts. The asymptotic regime is closely related to one commonly used in the study of networks of multiclass single-server queues (e.g., Meyn 1997, Chen and Meyn 1999, Maglaras 2000 and Bäuerle 2000). The closest to our methodological approach is Maglaras (2000) which proposes a general discrete-review policy for dynamic scheduling in a network of single-server queues and uses a Functional Strong Law of Large Numbers (FSLL) type scaling to establish asymptotic optimality in the fluid asymptotic regime formalized in Dai (1995). Under the discrete-review policy, the controller reviews the state of the system at the beginning of review periods of chosen length and prescribes a scheduling plan (number of customers in each class to be served) for that review period.

To adapt the approach of Maglaras (2000) to our multiserver setting, we use a many-server fluid regime similar to e.g., Mandelbaum et al. (1998) and Atar et al. (2010) that involves uniform acceleration of the arrival rates and the number of servers. The key feature in our asymptotic regime is that we leave the shift length unscaled. This stands in contrast to Maglaras (2000) and other discrete-review policies in the literature (e.g., Harrison 1996, 1998, Bassamboo et al. 2006) where the length of the review period is chosen by the controller and becomes negligible in the limit. By keeping the shift length unscaled, we preserve the discrete-time structure in the limit and obtain a discrete-time fluid control problem. This leads to a tighter asymptotic lower bound for our problem (compared to the case where the shift length vanishes in the limit). Further, by analyzing the limiting fluid control problem, we gain insights into the structure of good policies under discrete-time control. As a comparison, we also use the solution of the continuous-time control problem to develop a tracking discrete-review policy similar to those in the literature. We observe that although a naive implementation of this policy could perform poorly, a refined version which requires careful tuning and uses the insights obtained from the analysis of the discrete-time control problem leads to robust performance in the relevant parameter regime even under time-varying arrivals and inclusion of customer abandonment (see Section 6.3.2).

Another related (fluid) asymptotic regime is the one proposed by Bassamboo et al. (2006) (see also Harrison and Zeevi 2005 and Bassamboo et al. 2005). The asymptotic regime there is characterized by a large number of arrivals and servers but short service times and impatient customers such that in the limit the system "equilibrates instantly". That regime, which is related to the concept of pointwise stationary approximations (Green and Kolesar 1991, Whitt 1991), is not relevant for our dynamic server assignment problem as it corresponds to a setting where the transient effects are negligible. In our setting, the service times are on the same order of magnitude of the shift lengths (e.g. 3-4 hours versus 8-12 hours), so the transient dynamics play an important role in the optimization problem. As we discuss and show numerically, the benefits of a partially flexible design are small when the queue length processes quickly reach their steady-state distribution during the shifts.

Finally, our study relates to the growing literature on queueing models of patient flow in hospitals; see Armony et al. (2015) for a recent literature review. Closer to our work are models in support of staffing (e.g., Green et al. 2006, Yom-Tov and Mandelbaum 2014, Yankovic and Green 2011, Véricourt and Jennings 2011) which focus on determining the required number of providers. In contrast, we assume a fixed number of providers (nurses) and consider their dynamic assignment to different areas under the shift constraint. Other papers in the literature have studied dynamic scheduling or control of patient flows in the ED (e.g., Saghafian et al. 2012, Huang et al. 2015). To the best of our knowledge, dynamic assignment of providers, the main focus of our work, has not been previously considered in the literature.

1.2. Notation

We denote the nonnegative real line $[0,\infty)$ by $\mathbb{R}_+$, the set of nonnegative integers by $\mathbb{Z}_+$, and the $I-$dimensional Euclidean space by $\mathbb{R}^I$. For $x\in\mathbb{R}^I$ we denote the L_1 norm by $\|x\|=\sum_{i\in\mathcal{I}}|x_i|$. For two vectors a and b in $\mathbb{R}^I$, $a\geq b$ means that $a-b$ belongs to the nonnegative orthant $\mathbb{R}^I_+$. In particular, $a\geq b$ iff $a_i\geq b_i$ for all i. A vector valued function $f:\mathbb{R}^I_+\to\mathbb{R}^I_+$ is monotone iff each entry f_i is monotonic. Similarly, f is convex iff each real-valued entry f_i is convex.

All stochastic processes in this paper are defined on a common probability space $(\Omega,\mathcal{F},\mathbb{P})$ with expectation with respect to $\mathbb{P}$ denoted by $\mathbb{E}$. Further, all processes have sample paths in the $d-$dimensional Skorohod space $\mathbb{D}^d[0,\infty)$ for some positive integer $d\geq 1$. For $f:\mathbb{R}_+\to\mathbb{R}$ let $|f|_T=\sup_{0\leq t\leq T}|f(t)|$ and for $f:\mathbb{R}_+\to\mathbb{R}^I$ define $\|f\|_T=\sup_{0\leq t\leq T}\|f(t)\|$. A sequence of functions $\{f^n\}$ in $\mathbb{D}^d[0,\infty)$ converges uniformly on compact sets, denoted $f^n(t)\to f(t)$ u.o.c., if $\|f^n-f\|_T\to 0$ as $n\to\infty$ for any $T>0$. RCLL stands for right-continuous with left limits, a.e. for almost everywhere, and a.s. for almost surely.

All proofs are provided in the E-Companion.

2. Model Description and Formulation of the Control Problem

To study the impact of the shift constraint, we begin with a parsimonious queueing system with I customer classes, I parallel queues, and I server pools. Class $i\in\mathcal{I}\equiv\{1,\dots,I\}$ customers arrive to the system according to a Poisson process with intensity λ_i and have service requirements that are exponentially distributed with rate μ_i. There are a total of n servers available. Servers are cross-trained and capable of serving all customer classes. However, a server can only attend to a customer class if it is *assigned* to the server pool associated with that class. Under the fully flexible design, the assignment of servers to pools can be adjusted at any point in time. However, under the partially flexible design, the assignments can be adjusted only at the beginning of shifts of constant length $\tau>0$. Customers who find all servers assigned to their pool busy upon arrival are placed in an infinite capacity queue dedicated to their class. Each server can only be assigned to a single pool at any point in time. Servers of each pool serve customers of their class in a first-come, first-served and work conserving manner. That is, an assigned server to pool i is idle if and only if queue i is empty. We assume a preemptive service discipline so that if the assignment of a server is changed while the server is busy, the service can be interrupted, with the customer whose service is interrupted returning to the queue. This initial model will allow us to isolate the impact of the shift constraint. We will leverage the insights from its analysis to understand and study models with some relaxations of the modeling assumptions in Section 6.

A dynamic control policy in our formulation takes the form of a stochastic *assignment* process $U(t) = (U_1(t), \ldots, U_I(t))$ taking values in $\mathbb{Z}_+^I$ and with RCLL sample paths. Denote by $A_i(t)$ and $S_i(t), i \in \mathcal{I}$ a set of $2I$ mutually independent standard (unit rate) Poisson processes and let $X(0) \in \mathbb{Z}_+^I$ denote a random variable representing the initial number of customers in system (at time $t = 0$). An assignment process is said to be admissible under the fully flexible design if there exist (unique up to a set with measure zero) processes $X(t) = (X_1(t), \ldots, X_I(t))$, $Q(t) = (Q_1(t), \ldots, Q_I(t))$ and $Z(t) = (Z_1(t), \ldots, Z_I(t))$ taking values in $\mathbb{Z}_+^I$ and satisfying the following for $t \geq 0$:

$$X_i(t) = X_i(0) + A_i(\lambda_i t) - S_i\left(\mu_i \int_0^t Z_i(s)ds\right) \geq 0, \quad \text{for all } i \in \mathcal{I}, \tag{1}$$

$$Q(t) = X(t) - Z(t) \geq 0, \tag{2}$$

$$U(t) - Z(t) \geq 0 \tag{3}$$

$$Q_i(t)(U_i(t) - Z_i(t)) = 0, \quad \text{for all } i \in \mathcal{I}, \tag{4}$$

$$U(t) \cdot \mathrm{e} \leq n, \tag{5}$$

$$U(t) \geq 0. \tag{6}$$

Processes X, Q and Z are interpreted as the headcount, queue-length and the number of busy servers, respectively. That is, $X_i(t)$ is the total number of class i customers in the system; $Q_i(t)$ is the number of class i customers waiting in queue i; and $Z_i(t)$ is the number of busy servers in pool i (serving class i customers) at time $t \geq 0$. Equations (1)–(6) describe the system dynamics for any admissible control policy U. The first two equations describe the dynamics of the headcount and the queue-length processes. Equation (3) requires the number of busy servers in each pool to be less than or equal to the number of servers assigned to that pool. Equation (4) is the work-conserving condition. Note that although the policies are required to be work-conserving within each class, they are not necessarily work conserving with respect to the total work in system, i.e., there may be idle servers assigned to class $i \in \mathcal{I}$, while there are customers waiting in the queue of another class $i' \in \mathcal{I}$. In addition, it is easy to verify that (2)–(4) imply,

$$Z_i(t) = X_i(t) \wedge U_i(t), \quad \text{for all } i \in \mathcal{I}. \tag{7}$$

Intuitively, (4) ensures that servers are always busy serving customers of the class they are assigned to. Hence, the number of busy servers in each pool must be the minimum of customers in the corresponding class and the number of servers assigned to it. Finally, (5) ensures that the total number of assigned servers is bounded by n and (6) requires that the number of assigned servers to be nonnegative. We assume that at time $t=0$ the servers are unassigned and, given $X(0)$, the policy in effect determines the value of $Z(0)$ according to (7) and the value of $Q(0)$ according to (3).

Under the partially flexible design, we need to further limit the set of admissible policies to those that remain unchanged during shifts of length $\tau > 0$. Let t_k denote the starting time of shift $k+1$ with $t_{k+1} - t_k = \tau$, $k \in \mathbb{N}$ and $t_0 = 0$. Then a dynamic control policy U is said to be admissible under the partially flexible design, or *discrete-review,* if for all sample paths and $k \in \mathbb{N}$,

$$U(t) = U(t_k) \equiv U[k], \quad t \in [t_k, t_{k+1}). \tag{8}$$

That is, a discrete-review policy U has piecewise-constant sample paths that remain constant during intervals of length τ. This requirement captures the shift constraint, distinguishing our formulation from that of the more commonly studied continuous-time control problems. When U is not limited to the set of discrete-review policies, its control is equivalent to controlling the number of busy servers in each class subject to the queue-length being nonnegative (such as that in Bassamboo et al. 2006 and Atar et al. 2010). For discrete-review policies, U provides an *indirect* control on Z.

Finally, we express the objective of the system manager. Let $h = (h_1, \ldots, h_I)$ denote the holding cost rate vector where h_i denotes the holding cost associated with class i customers waiting in the queue. The objective is to choose an admissible discrete-review policy that minimizes the expected total holding cost of customers in queue over a finite horizon of length $T > 0$, i.e.,

$$\mathbb{E}\left[\int_0^T h \cdot Q(s) ds\right], \tag{9}$$

starting from the initial state $X(0)$. We are particularly interested in the setting where the initial state is large and imbalanced, and aim to find a cost effective policy to reduce the queue lengths to a desirable state over a sufficiently long, but finite horizon.

3. Fluid Control Problems and Asymptotic Lower Bounds

3.1. Fluid Control Problems

We start by proposing two deterministic fluid control problems associated with the original problems described in Section 2. In the fluid counterparts, the arrival rate is denoted by the vector $\bar{\lambda} \in \mathbb{R}_+^I$ and the total capacity is normalized to 1. The objective is to find an admissible capacity allocation function $\{u(t) \in \mathbb{R}_+^I; 0 \le t \le T\}$ that minimizes the total transient cost of the system over a finite horizon T starting from the initial condition $x^0 \in \mathbb{R}_+^I$.

Problem 1. The first problem is a *continuous-time control* problem stated as follows.

$$J_T(x^0) \equiv \min_{u(\cdot)} \int_0^T h\cdot q(s)ds \tag{10}$$

$$\text{s.t.} \quad q(t) = (x(t)-u(t))^+ \quad \forall t\in[0,T], \tag{11}$$

$$\dot{x}(t) = \bar{\lambda} - \hat{\mu}\cdot(x(t)\wedge u(t)), \quad \forall t\in[0,T], \tag{12}$$

$$x(0) = x^0, \tag{13}$$

$$u(t)\cdot \mathrm{e} \le 1, \quad \forall t\in[0,T], \tag{14}$$

$$u(t)\ge 0, \quad \forall t\in[0,T], \tag{15}$$

where $\hat{\mu}\equiv \mathrm{diag}(\mu)$ and $(x(t),u(t))\in\mathbb{R}_+^I\times\mathbb{R}_+^I$ is admissible if $x(t)$ is absolutely continuous and they jointly satisfy (12)–(15). We denote the optimal value of the above problem, which allows for continuous reassignment, over all admissible policies by $J_T(x_0)$. We also refer to the headcount trajectory $x(t)$ under an optimal control as the optimal trajectory.

Next, we define a discrete-time optimal control problem by limiting the set of admissible policies $u(t)$ to piecewise-constant (right-continuous) functions. For simplicity assume that $T = N\tau$ for some $N\in\mathbb{N}$. For all $t\in[0,T]$ we require that $u(t)=u(\tau k)\equiv u[k]$ for $k=\lfloor t/\tau\rfloor$. The problem is then to find the sequence of controls $\{u[k];k\in\mathcal{K}\equiv\{0,\ldots,N-1\}\}$ that minimizes the corresponding cost function in (10).

We use $x[k]\equiv x(\tau k)$ for $k\in\mathcal{K}$ to denote the state of the system at the beginning of stage $k+1$. Note that under a piecewise-constant control the trajectory satisfies the ODE in (12) with a constant allocation $u(t)=u[k]$ during stage (shift) $k+1$. Since the RHS of (12) is Lipschitz in $x(t)$ and continuous in the allocation $u[k]$, there exists a unique solution that is continuous in t, allocation $u[k]$ and the initial condition $x[k]$. Denote this solution by $f^k(x[k],u[k],t)\in\mathbb{R}_+^I, t\in[k\tau,(k+1)\tau)$. Then, during each stage, the trajectory $x(t)$ evolves according to $f^k(x[k],u[k],\cdot)$. Note that by continuity of the trajectory $x(t)$ we have $x[k+1]=f^k(x[k],u[k],(k+1)\tau)$. Finally, the stage cost function for period $k+1$ under control u and starting from x can be written as

$$g^k(x,u) = \int_{\tau k}^{\tau(k+1)} h\cdot q(s)ds = \int_{\tau k}^{\tau(k+1)} h\cdot\left(f^k(x,u,s)-u\right)^+ ds. \tag{16}$$

Problem 2. The *discrete-time control* problem can be stated as,

$$J_{T,\tau}(x^0) \equiv \min_{\{u[k];k\in\mathcal{K}\}} \sum_{k=0}^{N-1} g^k(x[k],u[k]) \tag{17}$$

$$\text{s.t.} \quad x[k+1] = f^k(x[k],u[k],(k+1)\tau), \quad \forall k\in\mathcal{K}, \tag{18}$$

$$x[0] = x^0, \tag{19}$$

$$u[k]\cdot\mathrm{e}\le 1, \quad \forall k\in\mathcal{K}, \tag{20}$$

$$u[k]\ge 0, \quad \forall k\in\mathcal{K}. \tag{21}$$

We denote the optimal value of the above problem over all admissible policies by $J_{T,\tau}(x^0)$.

3.2. Asymptotic Lower Bounds

Using fairly standard techniques, one can show that the dynamics used in the description of the fluid control problems are justified through a FSLLN for a properly scaled sequence of the associated processes in the original stochastic system operating under any admissible control policy (see Proposition EC.1). More specifically, we consider a sequence of systems as described in Section 2, indexed by n. The nth system has n servers, arrival rate $\lambda^n = (\lambda_1^n, \ldots, \lambda_I^n)$, service rates $\mu^n = (\mu_1^n, \ldots, \mu_I^n)$, shift-length τ^n, and is initiated with $X^n(0)$ customers in the system.

We assume that as $n \to \infty$ the arrival rates and initial conditions scale up linearly while the service rates and shift length remain fixed, i.e, $\mu^n = (\mu_1, \ldots, \mu_I)$ and $\tau^n = \tau$ for all $n \in \mathbb{N}$. Formally, we make the following assumptions regarding the sequence of arrival rates and initial number of customers in the system.

Assumption 1. *The sequence of random variables $\{n^{-1}X^n(0)\}$ is uniformly bounded, that is, for some constant $M > 0$ and for all n we have $\|n^{-1}X^n(0)\| \leq M$. Furthermore, as $n \to \infty$,*

$$n^{-1}X^n(0) \to \bar{X}(0), \tag{22}$$

a.s. for some random variable $\bar{X}(0) \in \mathbb{R}_+^I$.

Assumption 2. *For each $i \in \mathcal{I}$ there exists a $\bar{\lambda}_i > 0$ such that as $n \to \infty$,*

$$\bar{\lambda}_i^n \equiv n^{-1}\lambda_i^n \to \bar{\lambda}_i. \tag{23}$$

Denote the stochastic processes associated with the nth system with X^n, Q^n, Z^n and the control process by U^n. A sequence of control policies $\{U^n\}$ is said to be admissible if U^n is an admissible policy for the nth system for all n. We denote the set of all admissible policies under the fully flexible design by Π^n, and the set of discrete-review policies with shift length τ by Π_τ^n. Clearly, $\Pi_\tau^n \subset \Pi^n$.

Associated with each sequence of dynamic control policies $\{U^n\}$ and a finite horizon $T > 0$ is a sequence of total costs denoted by $\{C_T^n(U^n)\}$, where the total cost of the nth system is given by,

$$C_T^n(U^n) = \int_0^T h \cdot Q^n(s) ds. \tag{24}$$

Our goal is to find a sequence of discrete-review controls that minimizes the expected fluid-scaled total cost of the system for n sufficiently large. We formalize the notation of asymptotic optimality for our problem in the following definition.

Definition 1 (Fluid-Scale Asymptotic Optimality). We say a sequence of discrete-review policies $\{U_*^n\}$ is asymptotically optimal if,

$$\limsup_{n\to\infty} \mathbb{E}[n^{-1}C_T(U_*^n)] \leq \liminf_{n\to\infty} \mathbb{E}[n^{-1}C_T(U^n)], \tag{25}$$

for any sequence of admissible policies $\{U^n\}$ such that $U^n \in \Pi_\tau^n$ for all n.

The following theorem presents two asymptotic lower bounds for the expected (scaled) cost of the system; one using the solution to Problem 1 for all admissible policies, and a *tighter* bound for discrete-review policies using the solution to Problem 2.

Theorem 1. *For any sequence of admissible control policies $\{U^n\}$, i.e., with $U^n \in \Pi^n$ for all n, we have*

$$\liminf_{n\to\infty} \mathbb{E}[n^{-1} C_T^n(U^n)] \geq \mathbb{E}[J_T(\bar{X}(0))]. \tag{26}$$

Furthermore, for any sequence of discrete-review control policies, i.e., with $U^n \in \Pi_\tau^n$ for all n, we have

$$\liminf_{n\to\infty} \mathbb{E}[n^{-1} C_T^n(U^n)] \geq \mathbb{E}[J_{T,\tau}(\bar{X}(0))], \tag{27}$$

and

$$\mathbb{E}[J_{T,\tau}(\bar{X}(0))] \geq \mathbb{E}[J_T(\bar{X}(0))]. \tag{28}$$

4. Analysis of the Fluid Control Problems

Before presenting the analysis of the fluid control problems, we first discuss a simple feasible solution for them, namely, the *dedicated allocation.* Dedicated allocation is a fixed allocation vector u^d that is feasible and remains unchanged in time, i.e., $u(t) = u^d \geq 0$ for all $t \geq 0$ and $u^d \cdot \mathrm{e} \leq 1$. Under dedicated allocation, each class $i \in \mathcal{I}$ behaves as an independent multiserver fluid queue. Define the vector $\bar{\rho} \in \mathbb{R}_+^I$ with $\bar{\rho}_i = \bar{\lambda}_i/\mu_i, i \in \mathcal{I}$. It is well-know that for each class $i \in \mathcal{I}$ with any initial state x_i^0 if $\bar{\rho}_i < u_i^d$ then $x_i(t) \to \bar{\rho}_i$ as $t \to \infty$. Therefore, assuming $\bar{\rho}_i < u_i^d$ for all $i \in \mathcal{I}$, if T is large enough the queues eventually vanish (since $u_i^d > \bar{\rho}_i$) and, hence, the total cost of the system remains finite as $T \to \infty$. In this case, we say that a *stable dedicated allocation* exists.

4.1. Continuous–Time Control (Problem 1)

When the allocation can be controlled continuously in time, the problem reduces to a standard scheduling problem. In this case, a $c\mu-$type policy is known to be optimal at least in the single server setting (see, e.g., Avram et al. 1995). Here we discuss the policy and present some numerical examples to contrast its structure to that of Problem 2.

The $c\mu$ policy for the fluid model myopically allocates capacity in order of the $c\mu-$index which only depends on the service and holding cost rates. Without loss of generality, assume that the I classes are ordered such that $h_1\mu_1 \geq \ldots \geq h_I\mu_I$. The allocation vector under the $c\mu$ policy, denoted by $\hat{u}(t)$, is given by

$$\hat{u}_i(t) = x_i(t) \wedge \left[1 - \sum_{j=1}^{i-1} x_i(t)\right]^+, \quad i \in \mathcal{I},\ t \in [0,T], \tag{29}$$

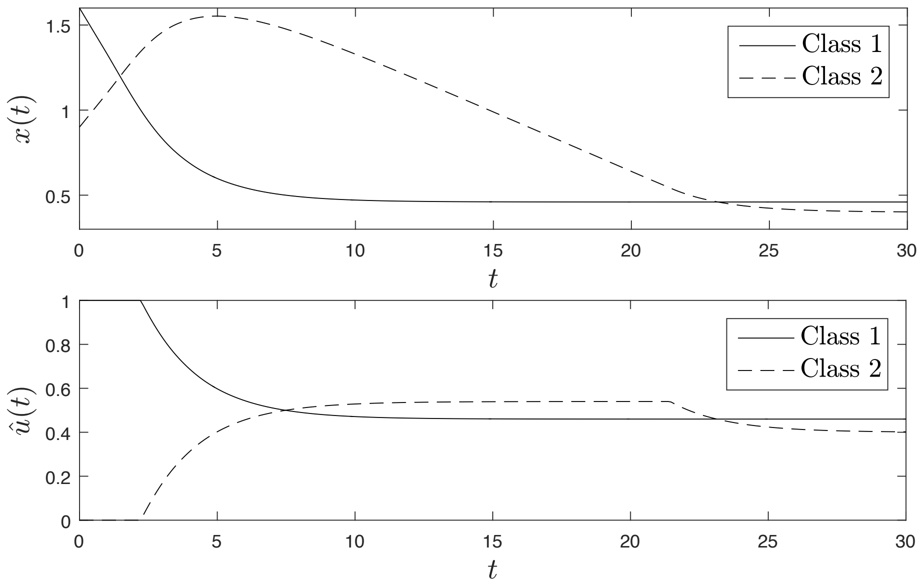

Figure 1 $c\mu$ trajectories (top) and allocations (bottom) for a $I = 2$ class system with $\bar{\lambda} = (0.23, 0.20)$, $\mu = (0.5, 0.5)$, and initial conditions $x^0 = (1.6, 0.9)$.

with the optimal trajectory satisfying the system of ODEs in (12). If there is enough capacity in the system, i.e., $\sum_{i \in \mathcal{I}} \bar{\rho}_i < 1$, the trajectory $x_i(t)$ converges to $\bar{\rho}_i$ for each class $i \in \mathcal{I}$ as $t \to \infty$, regardless of the initial condition x^0.

The system of ODEs can be solved numerically. Figure 1 presents the $c\mu$ trajectories and allocations for a 2-class system with parameters $\bar{\lambda} = (0.23, 0.20)$, $\mu = (0.5, 0.5), h = (4, 2)$, and initial condition $x^0 = (1.6, 0.9)$. Observe that, at any point in time, the allocation for class 1 (with the higher $c\mu$ index) is the minimum of class 1 fluid and the available capacity, while for class 2 it is the minimum of class 2 fluid and the left over capacity from class 1. That is, at any point in time, the allocation for class 1 is the maximum capacity such that no idleness is incurred in that class.

4.2. Discrete-Time Control (Problem 2)

4.2.1. Preliminaries Before studying Problem 2 in detail, we explicitly characterize the dynamics and the cost incurred over each stage, both of which can be obtained using $f^k(x, u, t)$, i.e., the solution to the Initial Value Problem (IVP),

$$\dot{x}_i(t) = \bar{\lambda}_i - \mu_i \left(x_i(t) \wedge u_i \right), \quad i \in \mathcal{I} \text{ and } t \geq k\tau, \tag{30}$$

given the vector of initial condition $x = x(k\tau)$ and allocation $u = u(k\tau)$. Without loss of generality, we consider the evolution over a single stage for $t = k\tau = 0$. Denote the dynamics over the stage by $f(x, u, t)$, $t \in [0, \tau)$, then we have $f^k(x, u, t) = f(x, u, t - \tau k)$, $t \in [k\tau, (k+1)\tau)$ for all $k \in \mathcal{K}$. Note that since each class evolves independently, it suffices to characterize the solution for a given class

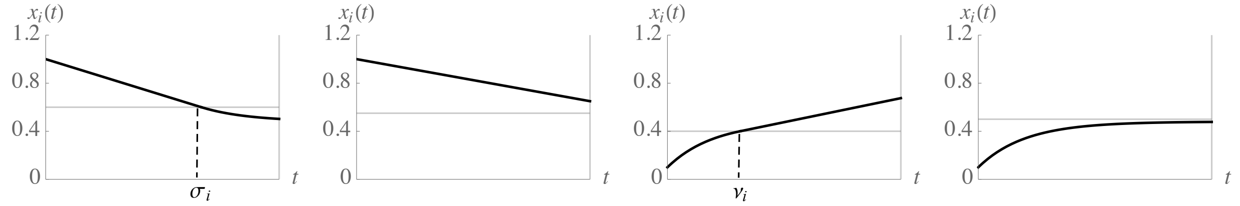

Figure 2 Possible trajectories of $x_i(t)$ during a stage under different allocations and initial conditions. (The horizontal lines correspond to the allocations.)

$i \in \mathcal{I}$. The solution is found by solving the ODE in (30) under the two possible cases $x_i(t) > u_i$ and $x_i(t) \le u_i$, and using the continuity of the solution in t to obtain,

$$f_i(x,u,t) = \begin{cases} x_i + (\bar{\lambda}_i - \mu_i u_i)t, & u_i < x_i \text{ and } u_i \le \hat{u}_i(x,t), \quad \text{(Case 1)} \\ \bar{\rho}_i + e^{-\mu_i(t-\sigma_i)}(u_i - \bar{\rho}_i), & u_i < x_i \text{ and } u_i > \hat{u}_i(x,t), \quad \text{(Case 2)} \\ u_i + (\bar{\lambda}_i - u_i\mu_i)(t - \nu_i), & u_i \ge x_i \text{ and } u_i < \check{u}_i(x,t), \quad \text{(Case 3)} \\ \bar{\rho}_i + e^{-\mu_i t}(x_i - \bar{\rho}_i), & u_i \ge x_i \text{ and } u_i \ge \check{u}_i(x,t), \quad \text{(Case 4)} \end{cases} \tag{31}$$

where

$$\hat{u}_i(x,t) = \frac{x_i + t\bar{\lambda}_i}{1 + t\mu_i}, \quad \check{u}_i(x,t) = \bar{\rho}_i + e^{-\mu_i t}(x_i - \bar{\rho}_i), \tag{32}$$

and

$$\sigma_i(x_i,u_i) = \frac{1}{\mu_i}\frac{x_i - u_i}{u_i - \bar{\rho}_i}, \quad \nu_i(x_i,u_i) = \frac{1}{\mu_i}\log\left(\frac{\bar{\rho}_i - x_i}{\bar{\rho}_i - u_i}\right). \tag{33}$$

Note that for ease of notation we have suppressed the dependence of σ_i and ν_i on x_i and u_i in (31); we will do the same in the rest of the paper unless the dependence needs to be emphasized. We also refer to the four cases in (31) as Case 1, 2, 3, and 4, respectively.

Note that $\hat{u}_i(x,t)$ and $\sigma_i < t$ are, respectively, the minimum allocation required to empty the class i queue before time t starting from $x_i > u_i$, and the time (with respect to the beginning of the stage) when the queue empties. Further, $\check{u}_i(x,t)$ and $\nu_i < t$ are, respectively, the maximum allocation such that the queue starts building up before time t and the time at which the queue starts building up. Then, given a fixed allocation u and starting from x, $x_i(t)$ has monotone trajectories in time during each stage that take one of the following forms illustrated in Figure 2. When $x_i > u_i$ we start with a *fluid queue* of size $x_i - u_i$ that drains linearly and empties during the stage at time $\sigma_i < \tau$ if $u_i > \hat{u}_i(x,\tau)$ and stays positive otherwise. When $x_i \le u_i$ we start with an empty queue that could start building up during the period at time $\nu_i < \tau$ if $u_i < \check{u}_i(x,\tau)$, and otherwise stays empty throughout the period.

Next, using (16) we can obtain the stage cost function $g^k(x,u)$ as the weighted sum of costs incurred for all classes during the stage. Since the stage cost is also independent of the stage k, we drop the superscript k and write

$$g(x,u) = h \cdot c(x,u), \tag{34}$$

where $c_i(x,u)$ is the class i stage cost given initial condition x and allocation u, and can be computed using (31) and (16):

$$c_i(x,u) = \begin{cases} (x_i - u_i)\tau - (\mu_i u_i - \lambda_i)\tau^2/2, & u_i < x_i \text{ and } u_i \le \hat{u}_i(x,\tau), \\ (x_i - u_i)^2/(2(\mu_i u_i - \lambda_i)), & u_i < x_i \text{ and } u_i > \hat{u}_i(x,\tau), \\ (\lambda_i - u_i\mu_i)(\tau - \nu_i)^2/2 & u_i \ge x_i \text{ and } u_i < \check{u}_i(x,\tau), \\ 0, & u_i \ge x_i \text{ and } u_i \ge \check{u}_i(x,\tau). \end{cases} \tag{35}$$

Note that the four cases in (35) are the same as those in (31) with fixed $t = \tau$. Further, it is easy to verify using (35) that c_i and hence g are continuous in (x,u).

4.2.2. Dynamic Programming (DP) Formulation Problem 2 can be formulated as a DP. Denote the value function by $V^k : \mathbb{R}_+^I \to \mathbb{R}_+$ for each $k \in \mathcal{K} \equiv \{0,\dots,N-1\}$, then $V^k(x)$ is the minimum cost-to-go starting from x at stage $k+1$ (time $k\tau$) and the optimal cost is given by $V^0(x^0)$. The value function satisfies the Bellman equation,

$$V^k(x) = \min_{u\in\mathcal{U}} \left[g(x,u) + V^{k+1}(f(x,u,\tau)) \right], \tag{36}$$

with $V^N \equiv 0$ and $\mathcal{U} \equiv \{u; u \ge 0, u\cdot \mathrm{e} \le 1\}$ which allows us to recursively compute the minimum cost. An optimal policy is then characterized by

$$\phi^k(x) = \{u \in \mathcal{U}; V^k(x) = \tilde{V}^k(x,u)\}, \quad k \in \mathcal{K}, \tag{37}$$

where,

$$\tilde{V}^k(x,u) \equiv g(x,u) + V^{k+1}(f(x,u,\tau)). \tag{38}$$

The following result establishes the continuity of the value function $V^k(\cdot)$ and the existence of an optimal (possibly non-unique) solution $\phi^k(\cdot)$ to Problem 2. The upper hemicontinuity of the correspondence $\phi^k(\cdot)$ will be key in establishing the asymptotic optimality of the proposed policy in Section 5.

PROPOSITION 1. *The value function $V^k(\cdot)$ is continuous and $\phi^k(\cdot)$ is a non-empty, compact-valued, and upper hemicontinuous correspondence for all $k \in \mathcal{K}$.*

In the next result, we establish the monotonicity and convexity of the value function V^k. In particular, we show that $\tilde{V}^k(x,u)$ defined in (38) is convex in (x,u) and hence evaluating the optimal policy using (36) involves solving N convex optimization problems.

THEOREM 2. *Let $V^k(x)$ denote the minimum cost starting from x at time $k\tau$. Then, $V^k(\cdot)$ is convex and nondecreasing for all $k \in \mathcal{K}$.*

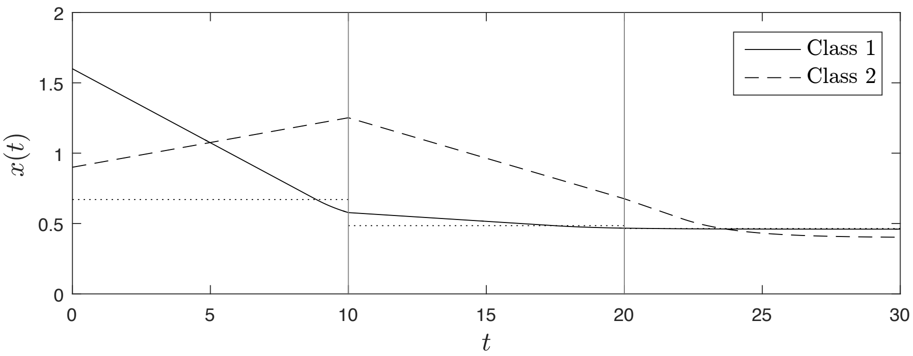

Figure 3 Optimal class 1 (solid) and class 2 (dashed) trajectories for Problem 2 with $\bar{\lambda} = (0.23, 0.20)$, $\mu = (0.5, 0.5)$, $x^0 = (1.6, 0.9)$, $h = (4, 2)$ and $\tau = 10$. The dotted horizontal lines correspond to optimal class 1 allocations for each stage.

In Figure 3, we plot the optimal trajectories for a system with the same parameters as the experiments in Figure 1, but now with control restricted to the beginning of stages of length $\tau = 10$. Observe that the optimal policy initially allocates more capacity to class 1, but not all capacity as in the case of continuous-time control, since that would result in "excessive" idleness during the stage. The optimal amount of idleness depends on the cost and system parameters and involves a tradeoff between cost reduction in the high-priority class, and building a low-priority queue that needs to be emptied in future stages. In the absence of idleness, as in the case of continuous-time control, the value of allocating more capacity to the high-priority class is the highest. However, the value diminishes as idleness is incurred during the stage, making the optimal allocation take the form of a non-trivial solution of a convex optimization problem. The intuition developed in this numerical example is formalized next.

4.2.3. Structural Results and Insights

In this section we discuss an important property of the optimal solutions to Problem 2. The main result provides insights into the structure of the optimal policy in the case of discrete-time control, and in particular on how it differs from continuous-time control. We will later leverage this result in design of the asymptotically optimal policy, as well as high-performing heuristics for an extension of the problem with time-varying arrivals and customer abandonment.

Before presenting the main result, we first present an intermediary result that establishes the existence of a *non-wasteful* optimal policy that allocates all capacity in all stages. Note that this does not imply that the policy is non-idling. In fact, we will find that the amount of idleness incurred during a stage is a critical component in the performance of an optimal policy.

PROPOSITION 2. *For any initial condition x^0, there exists an optimal policy $\{u[k]; k \in \mathcal{K}\}$ such that $u[k] \cdot e = 1$ for all $k \in \mathcal{K}$.*

We now provide a characterization of the optimal solution to Problem 2 as follows:

THEOREM 3. *Consider a two-class system ($I = 2$) and assume that the initial condition x^0 satisfies (i) $x^0 > \bar{\rho}$, and the cost parameters and service rates satisfy (ii) $h_1\mu_1 \geq h_2\mu_2$, and (iii) $h_1\mu_1\tau^2/2 + h_1\tau \geq h_2\mu_2\tau^2/2 + h_2\tau$. Then, there exists an optimal policy for Problem 2 that for all $k \in \mathcal{K}$ uses an allocation $\hat{\phi}^k(x[k]) \in \phi^k(x[k])$ satisfying $\hat{\phi}_1^k(x[k]) \geq \hat{u}_1(x[k], \tau) \wedge 1$.*

Theorem 3 provides a (state-dependent) lower-bound on the amount of capacity allocated to the "high-priority" class in each stage under assumptions $(i) - (iii)$. This lower bound is the minimum of total available capacity and the amount required to empty the queue exactly τ time units into the stage, i.e., at the end of the stage. The result implies that there exists an optimal policy that empties the class 1 queue in each stage if possible, and otherwise allocates all capacity to it.

This result highlights the impact of the shift constraint. The optimal policy is myopic, but only as long as no idleness is incurred in the high-priority class during a shift. Due to the inability to reallocate capacity mid-shift, in the presence of idleness, the optimal allocation requires carefully balancing the myopic cost-reduction in the current shift, with the future cost of emptying the low-priority queue.

To further illustrate the implications of the result, we now examine its conditions more closely. Condition (ii) is the same $c\mu-$index order assumed in the continuous-time problem. To understand condition (iii) note that $h_i\mu_i\tau^2/2 + h_i\tau$ is the marginal rate of decrease for the class i stage-cost, when allocating more capacity to it in the absence of idleness (see (EC.44) in the proof) and is comprised of two parts. The first part ($h_i\mu_i\tau^2/2$) corresponds to the cost reduction during the stage due to serving class i fluid at rate μ_i, whereas the second part corresponds to the cost reduction gained by reducing the queue for the full duration of the period by putting work into service at the beginning of the stage. Condition (iii) ensures that for the single-stage problem, *as long as no idleness is incurred in class 1*, the marginal value of allocating capacity to class 1 is greater than or equal to that of allocating to class 2. In fact, the proof relies on showing that this key property continues to hold for the multi-stage problem. For the single-stage problem, the property holds regardless of the initial condition x^0. In order to prove the extension of the property to the multi-stage problem, we require the assumption in condition (i) that the initial condition of both classes are above the stationary point $\bar{\rho}$. Nevertheless, we conjecture that the property holds for the multi-stage problem regardless of the initial conditions: allocating more capacity to class 1 (given that conditions (ii) and (iii) are met) is optimal as long as no idleness is incurred. See Section EC.4.1 for details and an example.

We now consider the impact of condition (iii) over condition (ii). Assuming that condition (ii) holds with an inequality, condition (iii) directly holds if $h_1 \geq h_2$ or the shift-length τ is sufficiently large. To see this note that condition (iii) can be written as $(\tau/2)(h_1\mu_1 - h_2\mu_2) \geq h_2 - h_1$. While these are both relevant parameter regimes with respect to our motivating application, in general, conditions (ii) and (iii) could be misaligned and the statement of the theorem may not hold if only condition (ii) is satisfied. Condition (iii) is necessary (and sufficient) for the statement of the theorem to hold for the single-stage problem, and numerically we observe that emptying class 1 queue at the end of the stage can perform arbitrarily bad compared to the optimal allocation when (iii) is not satisfied. This is not the case for the multistage problem. In fact, we numerically observe that as long as there are enough stages in the horizon for the queues to empty, condition (ii) is sufficient for the statement of the theorem to hold and emptying the class 1 queue at the end of the stage is very close to optimal. (See Section EC.4.2 for more details.) As our interest in the transient control problem is primarily motivated by finding cost effective ways to recover back to 'normal' after 'shocks' (e.g. epidemics) to the system, we are mainly interested in the multi-stage problem with enough time in the horizon to drain the queues. Thus, condition (iii) is unlikely to be very restrictive in practice.

5. An Asymptotically Optimal Discrete-Review (DR) Policy

In this section, we present an asymptotically optimal discrete-review policy. The DR policy is based on a translation of the solution to Problem 2 for the original stochastic system with a finite number of servers and is derived from two main elements that we describe next.

First, the policy requires a solution of Problem 2, given initial condition, x, and the number of stages to go, $N-k$. Recall from Section 4.2.2 that the solution to Problem 2 may not be unique. For instance, if the initial condition satisfies $x < \bar{\rho}$ and $\bar{\rho}_i \cdot \mathrm{e} < 1$, using any allocation satisfying $u[k] \geq \bar{\rho}$ for all $k \in \mathcal{K}$ leads to an empty queue that remains empty during all stages and results in zero total cost. As in the statement of Theorem 3, in the description of the policy below we use $\hat{\phi}^k(x)$ to denote an arbitrary optimal allocation from the set of optimal allocations starting at x and with $N-k$ stages to go; that is, $\hat{\phi}^k(x) \in \phi^k(x)$ for all x and $k \in \mathcal{K}$.

Recall that our results hold in the limit as $n \to \infty$. As such, problems may occur when stochastic fluctuations cause system dynamics to deviate substantially from mean behavior. In particular, this can lead to suboptimal or excessive idleness in the higher-priority classes, which as seen in the discussion of Theorem 3 needs to be carefully accounted for. To address this issue, we introduce a *safety parameter* to protect lower-priority classes in "small" systems while still maintaining asymptotic optimality. The safety parameter is implemented by carefully modifying the observed state of high-priority classes to shift resources towards the low-priority ones in order to avoid excessive idleness. Specifically, we introduce a safety parameter to the formulation of the DR policy denoted by $\beta^n = [\beta_1^n, \beta_2^n, \ldots, \beta_I^n]$ satisfying $\beta^n = o(n)$ that is subtracted from the system state so that the server allocation decision is determined based on a modified state of the system observed at the beginning of each shift. The use of safety parameters (or safety stocks) for translation of fluid solutions has been previously proposed in the literature; see, e.g., Maglaras (2000) and Dai and Weiss (2002), although the implementation there is different. In our numerical experiments, we find that although using a safety parameter is necessary for obtaining good policies for small systems, the performance is quite robust to the choice of parameters (we simply use zero for the lowest priority class and the same safety for all other classes). We discuss this further in Section 7.

The DR policy. We are now ready to present the policy for the nth system. Under the DR policy the state of the system is reviewed at times $\{t_k; k \in \mathcal{K}\}$ with $t_0 = 0$ and $t_{k+1} - t_k = \tau$. Starting at $t_0 = 0$ with $\bar{X}^n_*(0) = \bar{X}^n(0)$, at each time instance t_k and given the observed state of the system $X^n_*(t_k)$, an allocation decision for that shift $U^n_*[k] \equiv U^n_*(t)$, $t \in [t_k, t_{k+1})$ is made according to the following plan:

1. First a fluid-scale solution $\tilde{U}^n_*[k]$ is constructed using the normalized state $\bar{X}^n_*(t_k) = X^n_*(t_k)/n$ and safety vector $\bar{\beta}^n = \beta^n/n$ using

$$\tilde{U}^n_*[k] \equiv \hat{\phi}^k\left((\bar{X}^n_*(t_k) - \bar{\beta}^n)^+\right). \tag{39}$$

2. The fluid-scale solution is then translated to a feasible assignment for the original n server system. This involves multiplying the assignment by n and then appropriately rounding the solution. For simplicity, we use

$$U^n_*[k] = \left\lfloor n\tilde{U}^n_*[k]\right\rfloor, \tag{40}$$

where $\lfloor \cdot \rfloor$ is the floor function applied element-wise.

We note that the rounding in (40) can be carried out more carefully, e.g., using the largest reminder method, to avoid wasting capacity (see Section 7). However, it is easy to see that since the rounding error remains bounded and hence negligible under fluid-scaling, the asymptotic optimality of the policy is unaffected. (See the proof of Theorem 4.)

Theorem 4. *If Assumptions 1 and 2 hold, then the sequence of control policies $\{U^n_*\}$ is asymptotically optimal, that is, it achieves the asymptotic lower-bound (27) in Theorem 1:*

$$\limsup_{n\to\infty} n^{-1}\mathbb{E}[C^n_T(U^n_*)] \leq \mathbb{E}[J_{T,\tau}(\bar{X}(0))]. \tag{41}$$

As our result holds in the limit as $n \to \infty$, we wish to examine the performance of the DR policy when n is finite. We consider an example corresponding to a two-class system. The service and arrival rates of the system are respectively $\mu = (0.5, 0.5)$ and $\lambda^n = n\bar{\lambda}$, with $\bar{\lambda} = (0.23, 0.20)$ and the shift-length is $\tau = 10$. The initial condition is set to $X^n_i(0) = nx^0_i$ for $i \in \{1, 2\}$, with $x^0 = (1.6, 0.9)$. The holding cost rate is assumed to be $h = (4, 2)$. Note that the scaled parameters (i.e., $\bar{\lambda}$ and

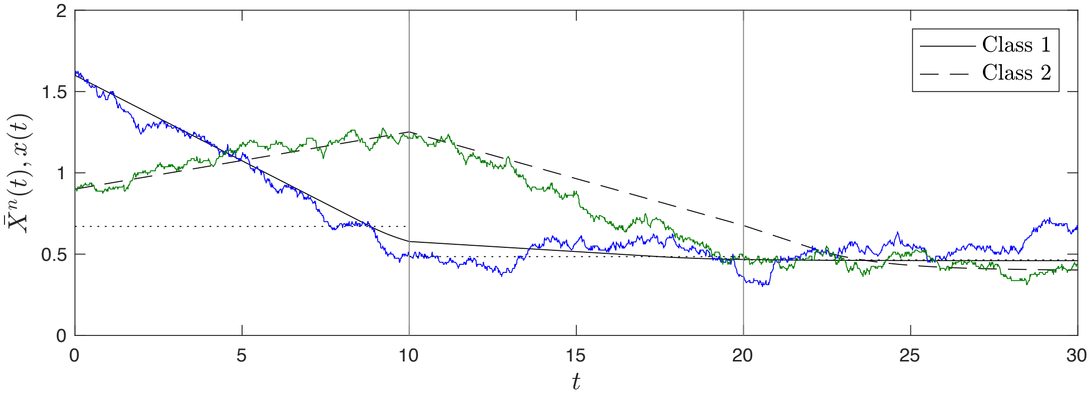

Figure 4 A sample path of $\bar{X}^n(t)$ under the DR policy overlaid on corresponding optimal trajectories for $I=2$ and system parameters $n=80$, $\mu=(0.5,0.5)$, $\lambda^n=(0.23n,0.20n)$, $X^n(0)=(1.6n,0.9n)$ and $\tau=10$.

Estimated cost $\mathbb{E}[n^{-1}C_T^n(U_*^n)]$			Lower bounds	
$n=20$	$n=80$	$n=300$	$J_{T,\tau}(x^0)$	$J_T(x^0)$
61.22 ± 3.5	52.20 ± 1.8	46.29 ± 1.0	42.02	33.48

Table 1 Performance of the DR policy U_*^n in terms of the expected (scaled) finite-horizon cost for different system sizes and in contrast to asymptotic lower-bounds of Theorem 1. The values after ± correspond to the %95 confidence intervals.

x^0) as well as the holding cost rate and shift length are the same as those in Figure 3. In this example, we set the safety parameter to $\beta^n=(0,0)$. Figure 4 plots the same (optimal) trajectories from Figure 3 with a (scaled) sample path of the stochastic $n=80$ server system overlaid on it. Starting from the initial condition, the DR policy solves for the optimal fluid allocations over a horizon of length $T=30$ ($N=3$). Since, due to stochastic fluctuations, the predicted state by the fluid model is different than the realized state of the stochastic system, the policy re-solves for the optimal allocations at the beginning of each stage. As formalized in Theorem 4, as n grows large, the fluid models become more accurate and the (scaled) sample paths and the resulting expected cost becomes closer to that under the optimal fluid trajectories and the lower-bound is (asymptotically) achieved.

We estimate the expected cost of the system under the DR policy by simulating 300 sample paths and for different system sizes. Table 1 presents the estimated costs as well as the asymptotic lower-bounds of Section 4. The first (tighter) lower-bound $J_{T,\tau}(x^0)$ corresponds to the optimal solution of the discrete-time fluid control problem (Problem 2) and the second lower-bound $J_T(x^0)$ corresponds to the optimal solution of the continuous-time control problem (Problem 1). The gap between $J_{T,\tau}(x^0)$ and $\mathbb{E}\left[n^{-1}C_T^n(U_*^n)\right]$ is due to stochastic fluctuations of $o(n)$, while the larger gap between $\mathbb{E}\left[n^{-1}C_T^n(U_*^n)\right]$ and $J_T(x^0)$ also includes the effect of restricting control to the beginning of shifts of length $\tau=10$.

6. Extensions and Heuristic Policies

To focus on the impact of the shift constraint, we began with a parsimonious model. That said, many real world problems exhibit more complex dynamics. In this section, we consider extensions of our model to include time-varying arrivals and customer abandonment and we leverage insights from the *base model* to develop heuristic policies for these more complex systems.

6.1. Time-Varying Arrivals

In many applications, including our motivating ED setting, the arrival process is non-stationary and exhibits significant temporal variation. Thus, we consider the case where the arrivals are driven by non-homogeneous Poisson processes with periodic time-varying rates (see, e.g., Kim and Whitt 2014 and Armony et al. 2015). We impose the following assumption on the arrival process of the nth system.

Assumption 3. *The instantaneous arrival rate for the nth system is given by a piecewise-continuous function* $\lambda^n(\cdot) \in \mathbb{R}_+^I$*. Further,* $\lambda^n(t)/n \to \bar{\lambda}(t)$ *u.o.c. as* $n \to \infty$*, where* $\bar{\lambda}(t)$ *is a bounded periodic piecewise-continuous function with period* p*, that is*

$$\bar{\lambda}(t+p) = \bar{\lambda}(t), \quad \textit{for all } t \geq 0,$$

and the period p *satisfies* $p = L\tau$ *for some integer* $L \geq 1$*.*

For example, in the ED setting, the period is a day ($p = 24$ hours) and the shift lengths are typically 8 or 12 hours ($L = 2$ or 3).

Under Assumption 3, the FSLLN for non-homogeneous Poisson processes implies that Theorem 1 directly extends to the case with time-varying arrivals on compact intervals. Further, solving Problem 2 with time-dependent dynamics, one can obtain the optimal fluid allocations and use it in the DR policy described in Section 5 to obtain an assignment for the original stochastic problem. (We discuss this approach further in Section 6.3.)

The challenge in the case of time-varying arrivals is in solving the fluid control problem. While this can be done numerically, characterizing the optimal policy in general is very hard. In fact, proving convexity of the value function itself is difficult. This is in part because closed-form expressions for the single-stage function cannot be obtained, even assuming a parametric family of arrival rate functions (e.g., sinusoidal or piecewise-constant). That said, numerical examples suggest the total cost is convex in the allocation as formally shown in the stationary case. For an intuitive argument, consider the cost-to-go function $\tilde{V}^k(x, u)$ defined in (38) with a fixed initial condition x and stage k. The definition of convexity requires that,

$$\tilde{V}^k(x, \delta u_1 + (1-\delta)u_2) \leq \tilde{V}^k(x, \delta u_1) + \tilde{V}^k(x, (1-\delta)u_2), \quad \delta \in [0,1]. \tag{42}$$

This implies that convexity in u is equivalent to having a lower optimal cost for a system with pooled capacity, compared to the sum of the optimal costs for two dedicated systems, under any allocation of capacity for the dedicated systems. Indeed, this should hold regardless of the arrival function.

Finally, we note that despite the more complicated dynamics introduced by time-varying arrivals, the underlying tradeoff between myopic cost reduction and idleness is the same as in the stationary case. The inclusion of time-varying arrivals can lead to non-monotone trajectories which change when idleness may occur during a stage – it no longer is always at the end or beginning[2] of the stage as in the case with stationary arrivals. Regardless of *when* the idleness is incurred, the optimal allocation involves balancing the amount of idleness incurred in higher-priority classes during the stage with the future cost of clearing the queue that builds up in low-priority classes. When excess capacity is scarce, avoiding "excessive" idleness becomes extremely important. We leverage this insight in design of the heuristic policies.

6.2. Customer Abandonment

Customer abandonment is also prevalent in many service systems and could significantly affect the system performance. With respect to our main motivating application, abandonment corresponds to patients Leaving the ED Without Being Seen (LWBS) (e.g. Batt and Terwiesch 2015). For other applications, e.g., the airport security application discussed in Martonosi (2011), customer abandonment is less relevant.

In the presence of abandonment, we consider a modified version of the cost function in (9),

$$\mathbb{E}\left[\int_0^T h\cdot Q(s)ds+\gamma\cdot M(s)\right], \tag{43}$$

where $\gamma=(\gamma_1,\ldots,\gamma_I)$ is the vector of abandonment cost with γ_i denoting the per customer abandonment cost for class i customers and $M(t)=(M_1(t),\ldots,M_I(t))$ is the counting process associated with abandoning customers with $M_i(t)$ denoting the number of class i customers who have abandoned by time t. The modified cost function measures the expected total holding and abandonment costs over a finite horizon of length T. We assume that customers wait an exponentially distributed amount of time in queue before abandoning the system, and denote by $\theta=(\theta_1,\ldots,\theta_I)$ the vector of abandonment rates with θ_i denoting the rate of abandonment for class i customers. In this case, (43) simplifies to

$$\mathbb{E}\left[\int_0^T \sum_{i=1}^I (h_i+\gamma_i\theta_i)Q_i(s)ds\right]=\mathbb{E}\left[\int_0^T \tilde{h}\cdot Q(s)ds\right], \tag{44}$$

where $\tilde{h}=(\tilde{h}_1,\ldots,\tilde{h}_I)$ with $\tilde{h}_i=h_i+\gamma_i\theta_i$ can be viewed as a modified holding cost rate vector that also captures the abandonment cost. Indeed, the headcount and queue processes are also affected by abandonment. We present the updated fluid dynamics in Section 6.3.

The scheduling problem with abandonment under continuous control has been the subject of recent research in the literature (see, e.g., Atar et al. 2010, Down et al. 2011, Larranaga et al. 2013). In general, the optimal policy may be state-dependent and as discussed in Down et al. (2011) involves managing a tradeoff between myopic cost reduction and avoiding future idleness due to abandonment. Characterizing the optimal policy becomes even more challenging under the shift constraint since, as highlighted by our base model, the inability to reallocate capacity continuously in time can lead to additional idleness. Given these challenges, we focus on understanding the impact of abandonment on the benefits of partial flexibility using heuristic policies.

6.3. Heuristic Policies

We now propose heuristic policies for the extended model involving both time-varying arrivals and customer abandonment. Our analysis of the base model highlighted the importance of balancing myopic cost reduction versus excessive idleness introduced by the shift constraint. Our heuristics will capture various levels of these competing factors and can be categorized as balancing both, focusing primarily on myopic cost reduction, or focusing on minimizing excessive idleness.

6.3.1. The Generalized DR (GDR) Policy The GDR policy heuristically extends the DR policy presented in Section 5 to the case with time-varying arrivals and abandonment. The policy is identical to that described for the base model except that the fluid allocation $\phi_*^k(\cdot)$ is obtained by solving a modified version of Problem 2 where the map $f^k(x[k],u[k],t)$, describing the dynamics during stage $(k+1)$, is the solution to the following IVP:

$$\begin{aligned} \dot{x}(t) &= \bar{\lambda}(t) - \hat{\mu}\cdot(x(t)\wedge u[k]) - \hat{\theta}\cdot(x(t)-u[k])^+, \quad t\in[\tau k, \tau(k+1)) \\ x(\tau k) &= x[k], \end{aligned} \tag{45}$$

where $\hat{\theta}\equiv \mathrm{diag}(\theta)$. Note that in addition to a time-varying input $\bar{\lambda}(t)$, the fluid abandons the queue at rate $\hat{\theta}\cdot(x(t)-u[k])^+$. Assuming stationary arrivals, the IVP can be solved analytically to find the transition and single-stage cost functions (see Section EC.2 in the E-Companion). In general, the IVP can be solved numerically. This carefully balances myopic cost reduction and idleness, but is computationally the most demanding.

6.3.2. Tracking Policies We now introduce policies that focus on cost reduction without explicitly accounting for the idleness introduced with the shift constraint. The main idea behind a tracking policy is to make allocation decisions in discrete-time with the goal of *tracking* trajectories known to be optimal for the continuous-time control problem (see, e.g., Maglaras 2000 and Bassamboo et al. 2006). More formally and within the context of our problem, one attempts to find a discrete-review policy for the nth system such that the fluid-scaled headcount process converges to the optimal continuous-time trajectory as $n \to \infty$, provided that the shift lengths are negligible in the limit. Therefore, intuitively one expects the policy to perform well if the system is large and the shift lengths are relatively small. For the base model, the above argument can be formalized by showing that the tracking policy described below achieves the asymptotic lower bound (26) in Theorem 1, assuming that in the sequence of systems described in Section 3 the shift length is also scaled and $\tau^n = o(1)$. We omit a formal proof here and instead use the idea to heuristically construct a policy for the extended model.

Consider an extended version of Problem 1 where the stage dynamics (12) are instead governed by the following ODE:

$$\dot{x}(t) = \bar{\lambda}(t) - \hat{\mu} \cdot (x(t) \wedge u(t)) - \hat{\theta} \cdot (x(t) - u(t))^+, \tag{46}$$

i.e., with a time-varying input and allowing for abandonment from the queues. Denote the optimal solution of the modified problem by $u^*(t)$. A natural way to translate this solution for the corresponding discrete-time control problem (Problem 2) is to use,

$$u_i[k] = \frac{1}{\tau} \int_{\tau k}^{\tau(k+1)} u_i^*(t)dt, \quad i \in \mathcal{I}, \tag{47}$$

in stage $(k+1)$, i.e., the average continuous allocation during that stage. In the following, we use this approach to construct a fluid-scale periodic allocation map, that takes the shift of the day and the initial state at the beginning of each shift, and returns an allocation for the shift. The allocation can then be translated to an integer assignment for the stochastic system using the same procedure described for the DR policy.

Without abandonment, i.e., with $\hat{\theta} = 0$, we use the $c\mu$ policy discussed in Section 4 as the continuous solution. Denote by $\{t_l; l \in \{0, \ldots, L-1\}\}$ the time-instances corresponding to the hour of the day when a shift starts. The fluid-scale allocation map, denoted by $\varphi(x,l)$, where $x \in \mathbb{R}_+^I$ is the state and $l+1 \in \{1, \ldots, L\}$ is the shift of the day, can be obtained using,

$$\varphi_i(x,l) = \frac{1}{\tau} \int_{t_l}^{t_{l+1}} \hat{u}_i(t)dt = \left(x_i + \int_{t_l}^{t_{l+1}} \bar{\lambda}_i(t)dt - x_i(t_{l+1}) \right) / (\tau \mu_i), \quad i \in \mathcal{I}, \tag{48}$$

where $\hat{u}(t)$ is the $c\mu$ allocation satisfying (29), and $x(t_{l+1}) \in \mathbb{R}_+^I$ is the state of the trajectory at the end of the shift under the $c\mu$ allocation, which can be obtained by numerically solving the ODE in (46) with allocation $\hat{u}(t)$ and starting from x.

With abandonment, the optimal continuous policy is unknown in general and may be state-dependent, although prioritizing according to the $c\mu$ and the $c\mu/\theta$ policy (prioritizing in the order of $\tilde{h}_i \mu_i / \theta_i$ indexes) is optimal in some problem instances (Atar et al. 2010, Larranaga et al. 2013). Given these complexities, we consider the static $c\mu$ and $c\mu/\theta$ policies as candidate continuous solutions. While the allocation does not admit an explicit form as in the case without abandonment, it can be obtained numerically using a similar approach as described above.

We refer to the policies that track the $c\mu$ trajectories (with or without abandonment) as $c\mu$-tracking (CMUT) policies, and the policy that tracks $c\mu/\theta$ trajectories as the $c\mu/\theta$-tracking (CMUθT) policy. We find that a naive implementation of these policies performs poorly, as they are obtained under the assumption of a negligible shift length. One issue is that the allocation does not necessarily add up to one (i.e., use all capacity) regardless of the initial condition. To deal with this issue, we equally divide the unused capacity among the classes. More importantly, since the allocations are computed assuming a negligible shift length, they ignore the potential idleness incurred during the shifts and tend to over-allocate capacity to higher-priority classes. To prevent excessive idleness during the shifts, we modify the normalized initial state using a safety parameter similar to that in the DR policy. As we demonstrate numerically in Section 7 the *refined* tracking policies, i.e., after the above modifications, perform well and comparable to the DR and GRD policies.

6.3.3. The Balancing Heuristic (BH) Policy In contrast to the tracking heuristics, which focus on cost reduction, the BH policy is designed to prevent excessive idleness. The BH is based on the insights regarding the importance of minimizing excessive idleness discussed in Section 4.2.3. Unlike the tracking policies, the BH ignores the cost parameters and instead aims to reallocate capacity to different classes in a way that none of them incur excessive idleness during the shift. As such, it could be far from optimal with respect to the cost function defined in (24). However, as we numerically illustrate in Section 7, by reallocating capacity when there are significant imbalances among different classes in terms of their congestion levels at the beginning of the shifts, it can achieve a significant reduction in the total expected queue length.

Similar to the tracking policies, the output of BH is a fluid-scale periodic allocation map, denoted by $\psi(x,l)$, which can then be translated back to the stochastic n server system using a similar procedure described for the DR policy. Recall that $\{t_l; l \in \{0,\dots,L-1\}\}$ denotes the time-instances corresponding to the hour of the day when a shift starts. Then, given a vector of initial conditions $x(t_l)$ and a piecewise-constant allocation vector $u(t) \geq 0$ satisfying $u(t)\cdot \mathrm{e} \leq 1$, the (fluid) dynamics of the system during shift $l+1$ are governed by the ODE in (46). In the following, we use $f^l(x,u,t), t \in [t_l, t_{l+1})$ to describe the dynamics of the system during shift $l+1$ given a fixed allocation $u(t) = u, t \in [t_l, t_{l+1})$ and starting from x at time t_l. Note that f^l can be computed using the IVP in (46).

The allocation map $\psi(x,l)$ is constructed in two steps. First, for each class an allocation vector $\tilde{\psi}(x,l)$ is computed using $f^l(x,u,t), t \in [t_l, t_{l+1})$, where $\tilde{\psi}_i(x,l)$ is the maximum allocation for class

i such that no idleness is incurred for that class during shift l. The final allocation $\psi(x,l)$ is then constructed according to the following procedure:

- If $\tilde{\psi}(x,l)\cdot \mathrm{e} < 1$, i.e., there is enough capacity such that at least one class must incur some idleness; then any remaining capacity is distributed equally among all I classes. This is achieved by setting

$$\psi(x,l) \leftarrow \tilde{\psi}(x,l) + (1-\tilde{\psi}(x,l)\cdot \mathrm{e})/I. \tag{49}$$

- If $\tilde{\psi}(x,l)\cdot \mathrm{e} \geq 1$, i.e., there is not enough capacity such that there exists an allocation where no class will have any idleness; then the allocation for each class is obtained by re-normalizing for each $i \in \mathcal{I}$ using

$$\psi_i(x,l) = \tilde{\psi}_i(x,l)/\tilde{\psi}(x,l)\cdot \mathrm{e}. \tag{50}$$

We conclude this section with two remarks. First, note that the BH does not require a safety parameter and as we show numerically in Section 7, is robust and performs well for a variety of system parameters. Second, we note that the only computational burden of the heuristic is computing the intermediary allocation vector $\tilde{\psi}(x,l)$ using $f^l(x,u,t), t\in[t_l,t_{l+1})$. This step can be carried out efficiently using a simple bi-section algorithm that searches for the target allocation by numerically computing the trajectory $f^l(x,u,t), t\in[t_l,t_{l+1})$ and checking whether idleness is incurred or not.

7. Numerical Study

In this section, we use numerical experiments to develop an understanding of the magnitude of the potential benefits gained by introducing partial flexibility as well as to examine the performance of our heuristics as we relax a number of assumptions of our base model. In doing so, we consider both transient and steady-state performance. The transient performance is useful when considering external shocks to the system, such as disease outbreaks or disasters, which can push the number of customers in system far from its steady-state distribution. The steady-state performance measures the benefits of partial flexibility in alleviating imbalances in congestion caused by temporal and stochastic fluctuations in arrival and service processes. Though our theoretical analysis focuses on transient control, as we discuss further below, it allows for obtaining stationary or periodic (in the case of time-varying arrivals) policies. Through simulation experiments we examine how the developed insights and heuristics extend to steady-state performance.

In Section 7.1, we start by considering our base model from Section 2. Next, we consider relaxations of modeling assumptions in Section 7.2. Finally, in Section 7.3 we report the results of a case study where we evaluate the performance of the proposed policies for a simulation model of the system in which we allow both non-stationary arrivals and abandonment, relax other assumptions of the queueing model, and calibrate the inputs using data from our partner hospital.

In all of our experiments, we implement a non-preemptive version of the policies so that the current service must be completed before reassigning a busy server to another pool. We note that the observations made in this section continue to hold under preemption. Further, although our theoretical results are obtained assuming preemption, one would expect the difference between non-preemption and preemption to be negligible in the limit under fluid-scaling.

Before turning to the examples, we comment on two numerical methods used in our examples. First, given a vector of (non-integer) assignment $U \geq 0$ such that $U \cdot \mathrm{e} = n$ we use the largest reminder method to obtain an integer assignment. That is, we first apply the floor function to the assignment and then sequentially assign the remaining servers in order of the largest difference between the floored assignment and the original one until an integer assignment is obtained. Second, stationary policies for the discrete-time fluid control problem are obtained by solving Problem 2 for a finite number of stages (to be specified later) using an implementation of the BFGS convex solver in Python.

7.1. Base Model

We consider a two-class system. Denote the offered load of class $i \in \mathcal{I} = \{1,2\}$ by $R_i = \lambda_i/\mu_i$. We fix the service rate vector μ and total utilization ρ of the system and change λ to obtain different load-ratios. That is, given a *load-ratio* r we have $r_i = R_i/\sum_{i\in\mathcal{I}} R_i$ with $\sum_{i\in\mathcal{I}} R_i = n\rho$. For both the DR and CMUT policies we set the safety parameter to $\beta_1^n = a\log(n)$ and $\beta_2^n = 0$ and vary the *safety factor* $a \in \{0,1,2,3,4,5\}$, unless otherwise stated. The reported cost reductions are the best amongst all safety factors considered.

Note that, in general, the solution of the DR policy is time-dependent. However, for a sufficiently large horizon T and a total cost that remains finite, we obtain a *stationary (time-independent) policy* (see, e.g., Maglaras 2000). In the case of base model, the existence of a stable dedicated allocation is sufficient for the optimal cost function to remain finite as $T \to \infty$. To obtain a stationary DR policy we solve Problem 2 with $N = 6$ stages at the beginning of each shift, which is typically enough for the solution to become stationary.

7.1.1. Value of partial flexibility.

We investigate the benefits of introducing partial flexibility in terms of the relative cost reduction compared to dedicated staffing. We focus on cases where a stable dedicated staffing exists and find the optimal dedicated staffing and the corresponding (exact) expected cost numerically using the Erlang-C formula.

Transient experiments. To examine the transient performance, we evaluate the expected cost of the DR and CMUT policies over a finite horizon, and for different system parameters and initial conditions. The initial headcount for each class is randomly selected from either a low-congestion

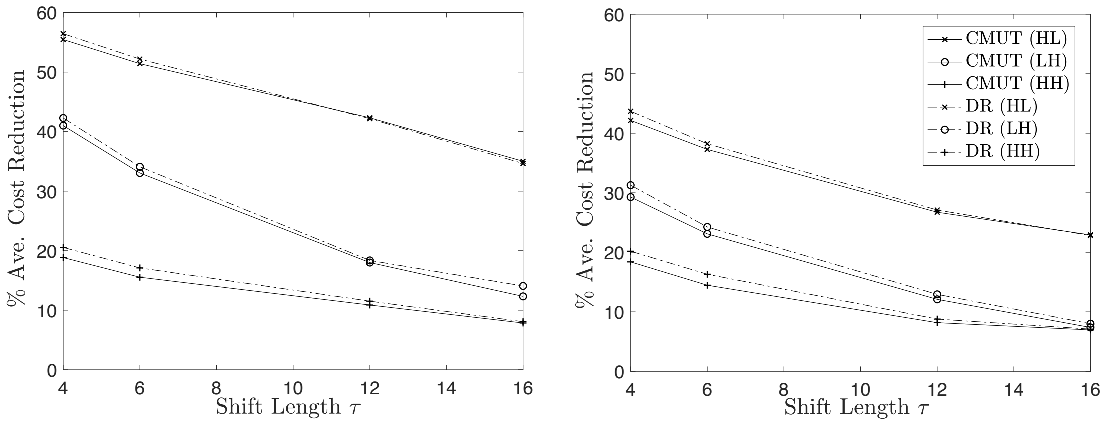

Figure 5 Performance of the policies in terms of estimated expected cost over a horizon of length $T = 48$ and for system parameters $\mu = (0.5, 0.5)$, $r = (1/2, 1/2)$, $\rho = 0.92$, $n = 48$ (left) and $n = 20$ (right).

set $L = \{0, \ldots, 0.5 \times n\}$ or a high-congestion set $H = \{n, \ldots, 1.5 \times n\}$. The results for two sets of experiments are summarized in Figure 5. The plots present the average (estimated) relative cost reduction with respect to dedicated staffing over 30 randomly generated initial conditions in 3 different regions, namely "HH", "HL", and "LH", and for different shift-lengths. For instance, region "HL" indicates that the initial headcount for class 1 and 2 belongs to set H and L, respectively. The system parameters are $\mu = (0.5, 0.5)$, $r = (1/2, 1/2)$, and $\rho = 0.92$. The holding cost rate is $h = (4, 2)$ and the horizon length is fixed at $T = 48$, which is typically long enough for the queues to empty by the end of the horizon. The left and right plots correspond to systems with $n = 20$ and $n = 40$ servers, respectively. In both examples, there are significant cost reductions (on average) over dedicated staffing (which splits the servers equally between the two classes) even when the shift-length is large. The benefits are higher when the system is initiated farther from the steady-state, and with a larger imbalance between the initial congestion of the two classes. The benefits are also higher when the system parameters are such that it takes longer for the system to reach its steady-state. The DR and CMUT policies achieve comparable performance. However, the performance of the CMUT policy is more sensitive to the choice of safety parameter. We further elaborate on the impact of system and safety parameters on the performance of the policies in the context of the steady-state experiments which we present next.

Steady-state experiments. The results for two sets of steady-state experiments are summarized in Figures 6 and 7. We plot the estimated percentage of cost reduction achieved over dedicated staffing for both policies and under different shift-lengths and system sizes. The service rate is $\mu = (0.5, 0.5)$ and the holding cost rate is $h = (4, 2)$. The left plot corresponds to utilization $\rho = 0.92$ and the right plot to utilization $\rho = 0.96$. The first example has equal load for each class, i.e., $r = (1/2, 1/2)$ and in the second example $r = (1/4, 3/4)$. In both examples, we observe significant

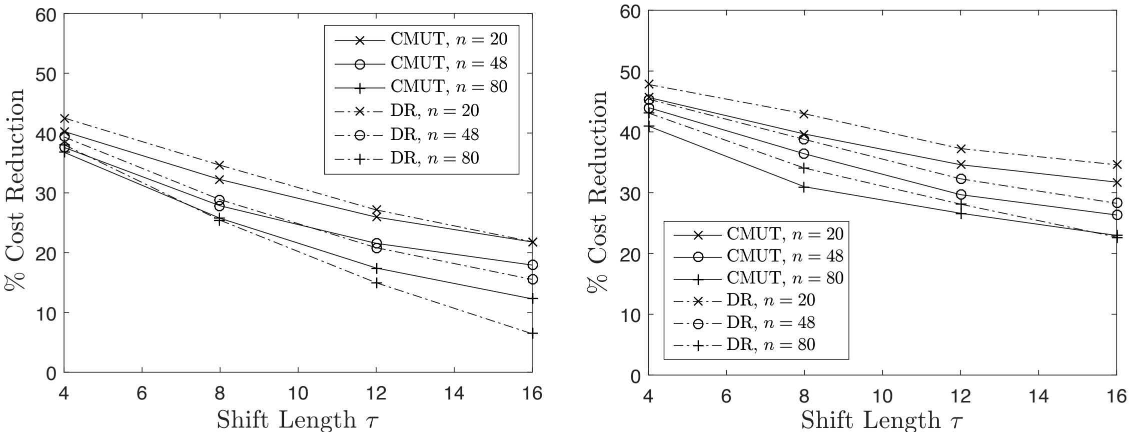

Figure 6 Performance of the policies in terms of estimated expected long-run average cost for system parameters $\mu = (0.5, 0.5)$, $r = (1/2, 1/2)$, $\rho = 0.92$ (left) and $\rho = 0.96$ (right).

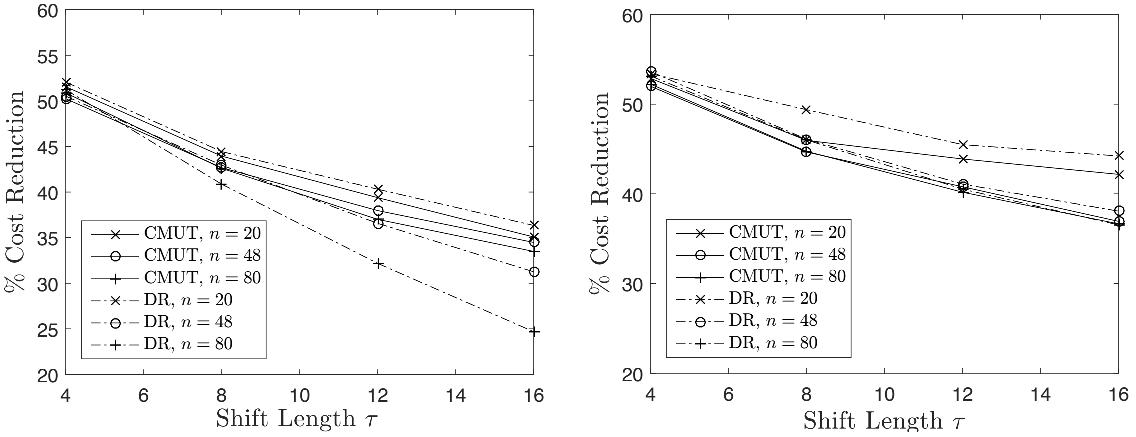

Figure 7 Performance of the policies in terms of estimated expected long-run average cost for system parameters $\mu = (0.5, 0.5)$, $r = (1/4, 3/4)$, $\rho = 0.92$ (left) and $\rho = 0.96$ (right).

benefits compared to dedicated staffing (ranging between 18% to 53% for $\tau = 12$). While the cost savings are not as large as the maximum reduction achieved under the continuous $c\mu$ policy (i.e., fully flexible design), which ranges between 68% to 78%, there are still substantial gains under the discrete-time control introduced by the shift constraint.

The results also illustrate the parameter regimes in which discrete-time control is more beneficial. Note that the percentage of the cost reduction for each policy decreases in the size of the system and shift length, and increases in utilization. This can be understood by noting that as the rate of service completion, $n\mu$, or shift-length, τ, increases, the queue-length processes reach steady-state faster during the shift. As a result, the average cost over shifts are close to that of the long-run average cost achieved under the staffing used in that shift, regardless of the initial condition. Therefore, the benefits become smaller in contrast to using the best dedicated staffing during all

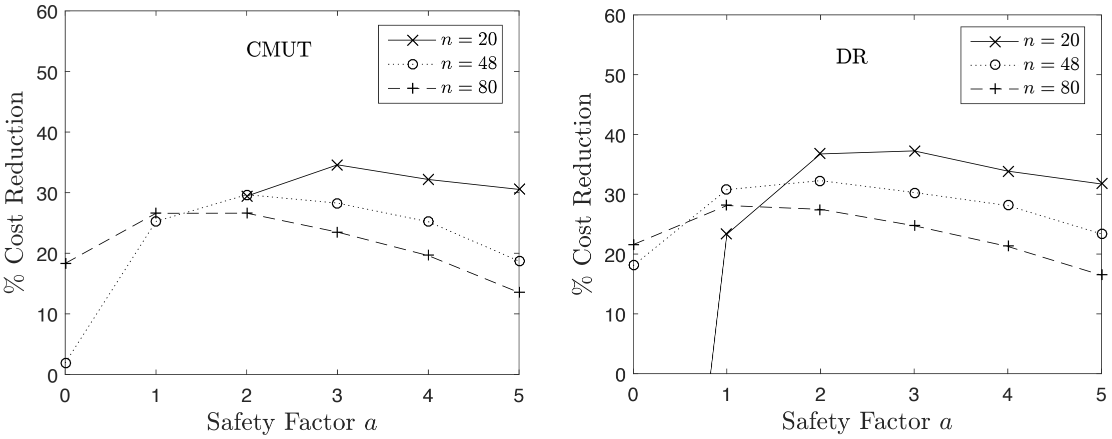

Figure 8 Performance of the policies in terms of estimated expected long-run average cost for different safety factors and parameters $\mu = (0.5, 0.5)$, $r = (2, 2)$, $\rho = 0.96$, and $\tau = 12$.

shifts. Similar to as in transient experiments, CMUT policy performs comparable to and in some cases even better than the DR policy. As we discuss next, however, the CMUT policy is more sensitive to the choice of safety parameter and can perform very poorly if it is not calibrated well.

7.1.2. Sensitivity to safety parameter. We find that the safety parameter is required for achieving good performance by both policies when the system size is "small" and the utilization is close to 1, especially with with respect to the steady-state performance. However, the performance of the DR policy is much more robust with respect to the choice of the safety parameter. We illustrate this by plotting the percentage of cost reduction for each policy under different safety factors in the steady-state example with $r = (1/2, 1/2)$ and $\rho = 0.96$ in Figure 8. Using a non-zero safety factor improves the performance of both policies but improvement is much larger under CMUT. Further, without use of the safety parameters the DR policy is stable in all examples. That said, when the system is small, i.e. $n = 20$, the DR policy performs 110.5% worse than dedicated. On the other hand, the CMUT becomes *unstable* without use of the safety parameters when $n = 20$. The poor performance for small systems can be attributed to two sources, namely the sub-optimality of the allocations at the fluid-scale[3] and the stochastic fluctuations of $o(n)$ not considered in the fluid control problems.

For larger systems with lower utilization, the role of the safety parameter becomes less important for both policies. Especially under the DR policy, the performance is fairly robust with respect to the safety parameter and typically using a safety factor equal to 1 for all classes except the lowest priority leads to significant improvements as demonstrated in Figure 8.

7.2. Model Extensions

We now consider the impact of the modeling extensions considered in Section 6, namely customer abandonment and time-varying arrival rates. We summarize our main observations here and refer the interested reader to the details of the numeric experiments in Sections EC.4.3 and EC.4.4 of the E-Companion, respectively. In order to isolate the impact of each of these extensions, we start by considering each of them individually. We then combine them, along with other relaxations, in our case study.

With a time-varying input and general stability due to abandonment, the cost of the system does not necessarily remain finite as $T \to \infty$. Therefore, for steady-state experiments, we solve the modified Problem 2 with a finite but large number of stages to obtain an (approximately) periodic GDR policy.

7.2.1. Customer abandonment. We start with the impact of abandonment. We observe that as long as the amount of abandonment is relatively small, allocating nurses according to a policy (i.e., CMUT, DR, or BH) that ignores abandonment can still lead to significant benefits in steady-state – with nearly 25% cost reductions compared to dedicated in some relevant parameter regimes. As the abandonment rate increases, the benefit over dedicated decreases substantially as the queue length processes reach steady-state very quickly. That said, the policies that explicitly take abandonment and the implied costs into account perform better and degrade more gradually. The LWBS rate in EDs is typically small (e.g., around 3% for our partner hospital), hence using a policy that ignores abandonment can still achieve significant gains over dedicated, while also being fairly easy to implement.

We observe that partial flexibility is most beneficial when the queue imbalances propagate past the shift boundaries and the system requires a number of shifts before reaching steady-state. This becomes more pronounced in the transient experiments as the system may not reach steady-state before the end of the horizon. Even when the steady-state benefits of partial flexibility are small, we observe that the proposed heuristics can significantly reduce the expected cost incurred over a finite horizon compared to dedicated staffing. This is most notable when starting from an initial condition that is “far” from the steady-state under the optimal dedicated staffing.

7.2.2. Time-Varying Arrivals We next turn our attention to systems with time-varying arrival rates. The observations on the benefits of partial flexibility are consistent with those for the stationary experiments. Further, we observe that since time-varying arrivals can move the system from underloaded to overloaded, the system can quickly move out of the well-behaved steady-state dynamics where partial flexibility has more significant benefits. This is even more pronounced when considering the transient performance.

We see that the GDR and CMUT heuristics perform comparably well. The BH achieves a lower cost reduction compared to the other heuristics since it ignores the cost parameters. Nevertheless, it still achieves a significant queue reduction by sharing capacity among different classes, without incurring excessive idleness.

7.3. Case Study

Finally, we evaluate the performance of the policies in Section 6.1 using a simulation model of the system calibrated using data from our partner ED. The purpose of this example is to illustrate the robustness of the proposed policies with respect to certain parametric and modeling assumptions made in our stylized model. More specifically, in the simulation model we relax the assumption of exponential service times and sinusoidal arrival rates, and incorporate key features of the ED including patients who LWBS, congestion dependent service times, and multi-tasking. We explain the details of the simulation model and its calibration next.

Simulation model. The simulation model for the case study has 4 customer classes that correspond to each area of the ED. Our partner ED has 4 areas and patients are primarily routed to specific areas based on their acuity level. More specifically, the majority of ESI level 1 and 2 patients (higher acuity) are treated in area 1; the majority of ESI level 4 and 5 patients (lower acuity) are treated in area 4; and the majority of ESI level 3 patients are treated in areas 2 and 3.

Customers of each class arrive to the system according to a non-homogeneous Poisson process with piecewise-constant (during each hour of the day) rates. Service times for each class are assumed to be log-normally distributed, with parameters that vary depending on the congestion in that class (e.g., as in Chan et al. 2014). More specifically, for each class, service times are generated from two different Log-Normal distributions, depending on whether the number of patients in that class is below or above a certain threshold, at the instances when service times are initiated. Finally, we assume that patients in each class abandon after waiting for an exponentially distributed time.

Calibration of the simulation input parameters. The hourly arrival rates for each class are estimated using the average number of arrivals for each hour of the day over the year in our dataset. The parameters of the service time distributions and the congestion threshold at which the parameters change for each class are estimated using Length of Stay (LOS) data (start of treatment to discharge) from each area. The threshold is set to the third quartile of the observed congestion in each area. The parameters of the log-normal distributions are then estimated using the maximum likelihood method and using LOS data filtered based on whether the congestion is above or below the estimated threshold, at time instances when treatments begin. The abandonment rates are set to $\theta = (1/120, 1/85, 1/85, 1/80)$ so that the percentage of abandonment under dedicated staffing is approximately the same as the LWBS percentage observed in the data (i.e., $\sim 3\%$). The shift-length is set to $\tau = 12$ hours with the first shift of the day starting at 7 am.

Since the total offered load of the system is equal to 42.71, we set the total capacity to $n = 48$. In the ED, nurses typically treat multiple patients at a time. Therefore, given a nurse-to-patient ratio $\varrho \in \mathbb{Z}_+$, we limit the assignments to groups of servers (of the given ratio) by rounding the

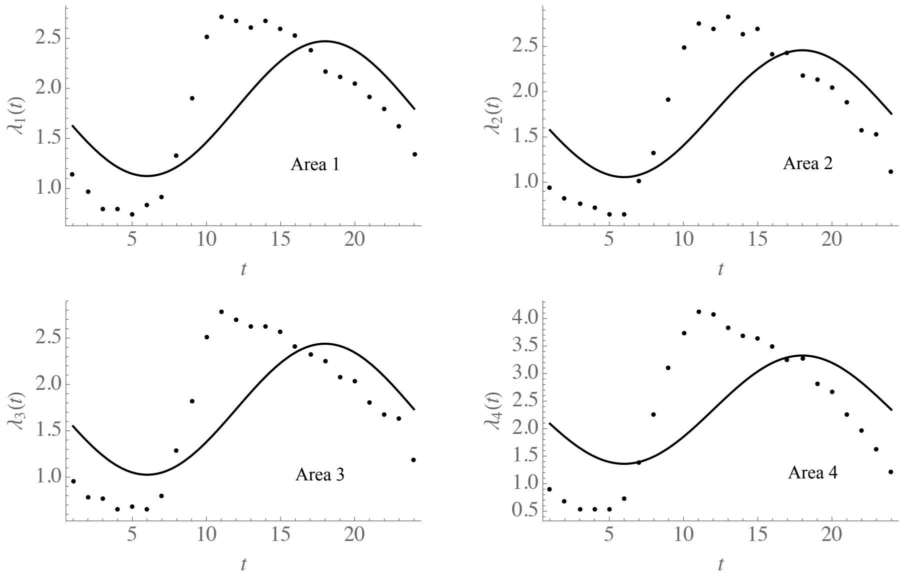

Figure 9 Empirical hourly arrival rates for each bay and the corresponding Sinusoidal fit.

allocations obtained from our heuristic policies to multiples of ϱ. We vary $\varrho \in \{1,2,4\}$ in our simulation experiments. Similarly, when determining the approximately optimal dedicated staffing using the Erlang-A formula, we find the best assignment of servers to different classes in multiples of ϱ.

Policy parameters. Note that the policies only require first-order statistics on (time-dependent) arrival rates and service requirements. For service times, the policies use the average LOS for each area. For arrival rates, we could use the empirical arrival rate function, which is the arrival rate function used to evaluate performance in the simulation; however, this makes computing the allocations for both heuristics more challenging. As such, the policies use a sinusoidal arrival rate $\lambda_i(t) = \alpha_{i,1} + \alpha_{i,2}\sin(\pi t/12)$ with parameters $\alpha_{i,1}, \alpha_{i,2}$ for $i \in \{1,2,3,4\}$ estimated by minimizing the squared error with respect to the empirical rates from the data. Figure 9 illustrates the arrival rates and the fitted Sinusoidal for each class. Note that curves fit the data coarsely. We will see that despite this poor fit, the heuristics still achieve reasonable gains.

A summary of the simulation model and policy parameters is given in Table 2. Note that the system parameters correspond to a regime with a "small" number of service completions during each shift. This, together with the low abandonment rate and high utilization, implies, using our previous insights and observations, that there are significant gains in using a partially flexible design.

Results and discussion. The results of the case study are presented in Table 3. Both policies achieve a significant reduction in terms of long-run average total cost over dedicated staffing assuming holding cost $h = (4,3,2,1)$ and abandonment cost $\gamma = (24,18,12,6)$. Specifically, we observe

			Area 1	Area 2	Area 3	Area 4
Service time	Congestion Thresh.		18	22	21	7
	Below Thresh.	(LogMean, LogSd)	$(1.77, 0.55)$	$(1.70, 1.89)$	$(1.75, 0.55)$	$(0.71, 0.81)$
	Above Thresh.	(LogMean, LogSd)	$(1.92, 0.50)$	$(1.89, 0.52)$	$(1.92, 0.50)$	$(0.85, 0.72)$
	Mean (hrs)		7.00	6.67	6.89	2.76
Arrival rate		$(\alpha_{i,1}, \alpha_{i,2})$	$(1.79, -0.67)$	$(1.75, -0.70)$	$(1.73, -0.70)$	$(2.34, -0.98)$
		Mean (/hr)	1.80	1.76	1.73	2.34
Offered load			12.58	11.72	11.94	6.47
Dedicated staffing ($\varrho = 1$)			13	12	12	7
Dedicated staffing ($\varrho = 4$)			12	12	12	8

Table 2 Summary of the inputs for the case study.

	1 to 1			1 to 4		
Perf. meas.	Ded.	BH	CMUT(1)	Ded.	BH	CMUT(1)
$\mathbb{E}[Q_1]$	3.53 ± 0.05	3.41 ± 0.02	3.16 ± 0.02	7.04 ± 0.10	4.04 ± 0.03	3.67 ± 0.02
$\mathbb{E}[Q_2]$	3.62 ± 0.05	3.31 ± 0.03	3.11 ± 0.02	3.62 ± 0.05	3.91 ± 0.03	3.66 ± 0.02
$\mathbb{E}[Q_3]$	4.07 ± 0.06	3.31 ± 0.02	3.47 ± 0.05	4.07 ± 0.06	3.92 ± 0.03	4.71 ± 0.07
$\mathbb{E}[Q_4]$	7.30 ± 0.10	4.33 ± 0.03	3.96 ± 0.02	2.92 ± 0.03	4.86 ± 0.04	4.55 ± 0.02
$\sum_i \mathbb{E}[Q_i]$	18.52 ± 0.15	14.36 ± 0.09	13.70 ± 0.10	17.65 ± 0.14	16.72 ± 0.10	16.59 ± 0.12
$\sum_i \tilde{h}_i \mathbb{E}[Q_i]$	75.59 ± 0.74	60.78 ± 0.38	58.07 ± 0.43	80.31 ± 0.68	71.06 ± 0.44	70.39 ± 0.53
% Aband.	0.0312 ± 0.0002	0.0242 ± 0.0001	0.023 ± 0.0001	0.0297 ± 0.0001	0.0282 ± 0.0001	0.0279 ± 0.0001
Ave. dev. from ded.	-	5.02 ± 0.02	4.22 ± 0.01	-	5.14 ± 0.02	4.64 ± 0.02
% Queue red.	-	22.45%	26.02%	-	5.21%	6.00%
% Cost red.	-	19.59%	23.18%	-	11.52%	12.35%

Table 3 Detailed output of the case study simulation experiments.

reductions of around 20% when the nurse-to-patient ratio is $\varrho = 1$. The benefits are smaller when the nurse-to-patient ratio is greater than 1, as the policy is further limited with respect to reassignments. However, we still observe reductions of over 10% in the total cost when the nurse-to-patient ratio equals 4. In addition, the abandonment probabilities are reduced under all policies. We also report the average number of servers that are reassigned to another area (compared to dedicated staffing) at each shift. More specifically, the values are the long-run average (absolute) deviation from dedicated staffing for all classes combined, then divided by 2 (because moving 1 server will result in deviations of 1 in 2 areas). We observe that the benefits are achieved with a modest number of reassignments under both policies; about 5 servers (out of 48) when the nurse-to-patient ratio is 1, and about 4 servers when the nurse-to-patient ratio is equal to 4, which is approximately equivalent to 1 nurse. Thus, it may be possible to achieve these gains without requiring too many reassignments of nurses.

EDs are incredibly complex, so while our case study does not capture every nuanced featured in such settings, it does relax a number of the assumptions needed for our theoretical results. In doing so, we find that the qualitative benefits of our approach are quite robust to specific modeling assumptions. Moreover, we estimate that the potential savings in practice are likely to be quite substantial.

8. Discussion and Future Research

In this paper, we study a partially flexible design for service systems that have multiple customer classes and staff who work in shifts. The partially flexible design allows for dynamic assignment of servers to different areas (or classes) of the service system at the beginning of discrete shifts. Such a design may be valuable in hospital EDs and knowledge-intensive services where switching from one customer class to another is costly and may come with loss of productivity for the servers or increase service times. We consider a queueing model with multiple customer classes and server pools and exploit a fluid asymptotic regime that allows us to obtain effective server assignment policies and obtain insights on the impact of the shift constraint on the structure of control policies. Our analysis highlights a key insight in discrete-time assignment of servers to different customer classes: when assigning more servers (compared to dedicated staffing) to high-priority classes, one must be cautious of excess idleness mid-shift. The incurred idleness comes at the cost of building queues in other areas that have to be cleared in future shifts. Therefore, myopic assignments based on the observed congestion or priority level of the areas that do not take the incurred idleness into account, can result in worse performance compared to dedicated staffing, or even lead to instability. When this trade-off is properly taken into account and the parameters of the system are such that the system does not reach steady-state quickly during each shift, the partial flexibility could significantly reduce the expected waiting cost of the system compared to dedicated staffing.

In order to study the assignment of servers at discrete shift intervals, we consider a queueing model that stylizes the operations of complex service systems, such as the ED. Hence, like any other model, it has limitations. In particular, our model assumes a single type of resource that can process work at a specified rate at any of the areas. This may not hold in a multi-resource setting like the ED where the bottleneck resource may change depending on the availability of different resources (e.g., doctors, beds, labs). The First-Come-First-Served discipline and identically distributed service requirements within each class may also be violated in some service systems. Additionally, as discussed in Section 7, ED nurses handle multiple patients simultaneously, which may impact the processing rate depending on how many patients are being treated. Despite these limitations, our numerical results and the case study in Section 7 suggests that the performance of the proposed policies are fairly robust with respect to different system configurations. Since our proposed policies only depend on first-order statistics one can easily estimate the required input parameters, which makes them desirable from a practical standpoint. Finally, our model assumes that all staff are available for reassignment at the beginning of all shifts. Some service systems use staggered or overlapping shifts, that is, different servers may start their shifts at different times. While we do not directly consider staggered shifts in our model, the impact can be estimated using the results of this paper. Introducing staggered shifts allows for more frequent reassignments, but only for the subset of servers starting at each shift. As intuition suggests and we observe numerically in Section 6, more frequent reassignments increase the benefits of the partially flexible design. Further, since the benefits of the partially flexible design are achieved by reassigning a small number of staff at each shift, the limitation on the number of available nurses for reassignment is not expected to have a significant impact on the performance of the policies. We also note that the extension of the proposed policies to consider different number of servers at each shifts is straightforward. A thorough investigation of the impact of staggered shifts and coordination of reassignments with different shift-scheduling approaches is left for future research.

The introduction of the partially flexible design leads to new research directions that could be pursued in future. In the following, we discuss a few of these directions.

Stability region under the partially flexible design. The benefits of partial flexibility can also be quantified with respect to the stability region of the system, i.e., the set of parameters for which the system is stabilizable under the partially flexible design (in the absence of abandonment). Although intuitively adding partial flexibility is expected to increases the size of the stability region compared to dedicated staffing (with the fully flexible design achieving the largest stability region), future work can focus on full characterization of the stability region for a given shift-length and system size.

Choosing safety parameters for small systems. The discrete-review policy proposed in this work uses a safety parameter to achieve robust performance for all system sizes. We calibrate the safety parameters heuristically and rely on insights and numerical experiments for finding the right parameters. Although the experiments suggest that the performance of the policies are fairly robust with respect to the choice of safety parameters, a more fundamental understanding of the effect of safety parameters is lacking and requires analyzing associated Brownian control problems or characterizing large deviations from the mean path predicated by the fluid models.

Impact of customer abandonment. Future work should investigate the impact of abandonment on the structure of dynamic assignment policies under the shift constraint, and identify the parameter regime where despite significant abandonment partial flexibility can still be beneficial. Due to the complexity of the transient dynamics in the presence of abandonment, this is however likely to be a very challenging task. An alternative would be to only consider the cost of abandonment (and not the holding cost; see Puha and Ward 2019) for which the optimal policy has a simpler structure under full flexibility.

A general parallel server system. The proposed partially flexible design is suitable for service systems offering specialized service to different classes of customers. For such systems, serving customers in non-specialized server pools can compromise the quality of service, or increase service times as empirically demonstrated in the case of inpatient wards of a hospital in Song et al. (2015). If customers of certain classes can be served in multiple server pools without degradation of service, server reassignment can be considered jointly with dynamic routing. An extension of the model could be to consider a general parallel server system under the shift constraint and investigate the benefits of joint dynamic routing and server assignment policies. In the general setting, arrivals can be dynamically routed to a subset of customer classes. In addition, each server is capable of handling a subset of customer classes and the assignment can be adjusted to any of the eligible classes at the beginning of discrete shifts.

Balancing the cost of reassignment with gained benefits. In our model, we restrict control to the beginning of shifts. This corresponds to the case where mid-shift reassignments are infeasible or have a prohibitively high cost. An alternative model would be to associate a cost to reassignments and investigate policies that balance the cost of reassignment with the reduction of future waiting cost. This cost formulation may be more natural in other settings, such as a department store or airport security, where there may be more flexibility to reassign mid-shift. In the event that mid-shift reassignments can only happen at periodic intervals (e.g. at the half-way point), our formulation can account for this by appropriately defining the shift interval (e.g. half the shift length).

Our work here provides an important first step in understanding how the staffing constraint, which only allows staff to be reallocated at the beginning of discrete intervals, introduces interdependencies across shifts and how one must account for these dynamics when making operational decisions. As shift staffing is common in many service systems, much work remains to optimize tactical decisions including server allocations, customer routing, timing of shifts, etc.

Endnotes

1. Note that because of this structure, where areas are typically differentiated by the severity of the patients they treat, rerouting patients to different areas is rare.

2. Note that the idleness can also occur at the beginning of the shift in the case where the initial condition of class 1 is below the equilibrium. See also the numerical example in EC.4.1.

3. In the case of the DR policy, this error can technically be eliminated by solving Problem 2 with sufficiently large number of stages. In our experiments, using $N = 6$ is typically sufficient to obtain an stationary policy.

References

Anderson R, Gamarnik D (2015) Scheduling interns in hospitals: Queueing models and fluid approximations. Technical report, Working Paper, Massachusetts Institute of Technology, Cambridge, Massachusetts.

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

Atar R, Giat C, Shimkin N (2010) The $c\mu/\theta$ rule for many-server queues with abandonment. *Operations Research* 58(5):1427–1439.

Atlason J, Epelman MA, Henderson SG (2008) Optimizing call center staffing using simulation and analytic center cutting-plane methods. *Management Science* 54(2):295–309.

Avram F, Bertsimas D, Ricard M (1995) Fluid models of sequencing problems in open queueing networks; an optimal control approach. *Institute for Mathematics and its Applications* 71:199.

Bassamboo A, Harrison JM, Zeevi A (2005) Dynamic routing and admission control in high-volume service systems: Asymptotic analysis via multi-scale fluid limits. *Queueing Systems* 51(3):249–285.

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.

Batt RJ, Kc DS, Staats BR, Patterson BW (2019) The effects of discrete work shifts on a nonterminating service system. *Production and Operations Management* 28(6):1528–1544.

Batt RJ, Terwiesch C (2015) Waiting patiently: An empirical study of queue abandonment in an emergency department. *Management Science* 61(1):39–59.

Bäuerle N (2000) Asymptotic optimality of tracking policies in stochastic networks. *Annals of Applied Probability* 1065–1083.

Brusco MJ, Jacobs LW, Bongiorno RJ, Lyons DV, Tang B (1995) Improving personnel scheduling at airline stations. *Operations Research* 43(5):741–751.

Burke EK, De Causmaecker P, Berghe GV, Van Landeghem H (2004) The state of the art of nurse rostering. *Journal of scheduling* 7(6):441–499.

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.

Chen RR, Meyn S (1999) Value iteration and optimization of multiclass queueing networks. *Queueing Systems* 32(1-3):65–97.

Dai J, Shi P (2019) Inpatient overflow: An approximate dynamic programming approach. *Manufacturing & Service Operations Management* .

Dai J, Tezcan T (2011) State space collapse in many-server diffusion limits of parallel server systems. *Mathematics of Operations Research* 36(2):271–320.

Dai JG (1995) On positive harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *The Annals of Applied Probability* 5(1):49–77.

Dai JG, Weiss G (2002) A fluid heuristic for minimizing makespan in job shops. *Operations Research* 50(4):692–707.

Down DG, Koole G, Lewis ME (2011) Dynamic control of a single-server system with abandonments. *Queueing Systems* 67(1):63–90.

Green L, Kolesar P (1989) Testing the validity of a queueing model of police patrol. *Management Science* 35(2):127–148.

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

Green LV (2010) *Using Queueing Theory to Alleviate Emergency Department Overcrowding* (John Wiley & Sons, Inc.).

Green LV, Soares J, Giglio JF, Green RA (2006) Using queueing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine* 13(1):61–68.

Gurvich I, Whitt W (2010) Service-level differentiation in many-server service systems via queue-ratio routing. *Operations Research* 58(2):316–328.

Harrison JM (1996) The bigstep approach to flow management in stochastic processing networks. *Stochastic Networks: Theory and Applications* 4:147–186.

Harrison JM (1998) Heavy traffic analysis of a system with parallel servers: asymptotic optimality of discrete-review policies. *Annals of applied probability* 822–848.

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.

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.

Ingolfsson A, Haque MA, Umnikov A (2002) Accounting for time-varying queueing effects in workforce scheduling. *European Journal of Operational Research* 139(3):585–597.

Jouini O, Dallery Y, Nait-Abdallah R (2008) Analysis of the impact of team-based organizations in call center management. *Management Science* 54(2):400–414.

Kim SH, Whitt W (2014) Are call center and hospital arrivals well modeled by nonhomogeneous poisson processes? *Manufacturing & Service Operations Management* 16(3):464–480.

Kolesar PJ, Rider KL, Crabill TB, Walker WE (1975) A queuing-linear programming approach to scheduling police patrol cars. *Operations Research* 23(6):1045–1062.

Larranaga M, Ayesta U, Verloop IM (2013) Dynamic fluid-based scheduling in a multi-class abandonment queue. *Performance Evaluation* 70(10):841–858.

Maglaras C (2000) Discrete-review policies for scheduling stochastic networks: Trajectory tracking and fluid-scale asymptotic optimality. *The Annals of Applied Probability* 10(3):897–929.

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

Mandelbaum A, Stolyar AL (2004) Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized cμ-rule. *Operations Research* 52(6):836–855.

Martonosi SE (2011) Dynamic server allocation at parallel queues. *IIE Transactions* 43(12):863–877.

Meyn S (1997) Stability and optimization of queueing networks and their fluid models. *Lectures in applied mathematics-American Mathematical Society* 33:175–200.

Puha AL, Ward AR (2019) Scheduling an overloaded multiclass many-server queue with impatient customers. *Tutorials in Operations Research* .

Saghafian S, Hopp WJ, Van Oyen MP, Desmond JS, Kronick SL (2012) Patient streaming as a mechanism for improving responsiveness in emergency departments. *Operations Research* 60(5):1080–1097.

Song H, Tucker AL, Murrell KL (2015) The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Science* 61(12):3032–3053.

Staats BR, Gino F (2012) Specialization and variety in repetitive tasks: Evidence from a japanese bank. *Management science* 58(6):1141–1159.

Tirdad A, Grassmann WK, Tavakoli J (2016) Optimal policies of $M(t)/M/c/c$ queues with two different levels of servers. *European Journal of Operational Research* 249(3):1124–1130.

Van den Bergh J, Beliën J, De Bruecker P, Demeulemeester E, De Boeck L (2013) Personnel scheduling: A literature review. *European Journal of Operational Research* 226(3):367–385.

Van Mieghem JA (1995) Dynamic scheduling with convex delay costs: The generalized $c\mu$ rule. *The Annals of Applied Probability* 809–833.

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

Whitt W (1991) The pointwise stationary approximation for $M(t)/M(t)/s$ queues is asymptotically correct as the rates increase. *Management Science* 37(3):307–314.

Yankovic N, Green LV (2011) Identifying good nursing levels: A queuing approach. *Operations research* 59(4):942–955.

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

Supplementary Results, Proofs, and Additional Numerical Experiments

EC.1. Details of the Asymptotic Framework and Convergence to Fluid Limits

Here we provide details of the asymptotic framework introduced in Section 3 and formally establish the convergence of the sequence of fluid-scaled processes. We begin by formally defining a sequence of admissible control policies. Under the fully flexible design, a sequence of dynamic control policies $\{U^n\}$ is said to be admissible if for each n there exists unique processes X^n, Q^n, and Z^n satisfying the following relations:

$$X_i^n(t) = X_i^n(0) + A_i(\lambda_i^n t) - S_i\left(\mu_i T_i^n(t)\right) \geq 0, \quad \text{for all } i \in \mathcal{I}, \tag{EC.1}$$

$$T_i^n(t) = \int_0^t Z_i^n(s)ds, \quad \text{for all } i \in \mathcal{I}, \tag{EC.2}$$

$$Q^n(t) = X^n(t) - Z^n(t) \geq 0, \tag{EC.3}$$

$$U^n(t) - Z^n(t) \geq 0, \tag{EC.4}$$

$$Q_i^n(t)(U_i^n(t) - Z_i^n(t)) = 0, \quad \text{for all } i \in \mathcal{I}, \tag{EC.5}$$

$$U^n(t) \cdot \mathrm{e} \leq n, \tag{EC.6}$$

$$U^n(t) \geq 0. \tag{EC.7}$$

Note that the relations are the same as those in (1)–(6) except that $T_i^n(t)$ is used to express the cumulative amount of time that all n servers have spent on class i up to time t.

A sequence of dynamic control policies $\{U^n\}$ is said to be admissible under the partially flexible design, or discrete-review, if in addition to (EC.1)–(EC.7) and for almost all sample paths,

$$U^n(t) = \sum_{k=0}^{\bar{k}(t)-1} U^n[k] \mathbf{1}_{[t_k, t_{k+1})}(t), \tag{EC.8}$$

for some sequence $\{U^n[k]; 0 \leq k \leq \bar{k}(t) - 1\}$ where $\mathbf{1}_B$ is the indicator function of set B, t_k denotes the starting time of shift $k+1$ with $t_{k+1} - t_k = \tau > 0$, $k \in \mathbb{N}$ and $t_0 = 0$, and finally $\bar{k}(t) = \min\{k; t_k \geq t\}$. Note that (EC.8) limits the set of admissible policies under the partially flexible design to those with piecewise-constant sample paths.

Let $Y^n(t)$ and $\Upsilon^n(t)$ denote the cumulative assignment and queue length processes, that is

$$Y^n(t) = \int_0^t U^n(s)ds, \quad \text{and} \quad \Upsilon^n(t) = \int_0^t Q^n(s)ds, \tag{EC.9}$$

and define the fluid-scaled processes,

$$\bar{X}^n = n^{-1} X^n, \quad \bar{Q}^n = n^{-1} Q^n, \quad \bar{Z}^n = n^{-1} Z^n, \quad \bar{U}^n = n^{-1} U^n,$$
$$\bar{T}^n = n^{-1} T^n, \quad \bar{Y}^n = n^{-1} Y^n, \quad \bar{\Upsilon}^n = n^{-1} \Upsilon^n.$$

The following proposition establishes the existence of a converging subsequence for the sequence of the scaled processes $\{(\bar{X}^n, \bar{T}^n, \bar{Y}^n, \bar{\Upsilon}^n)\}$ associated with any sequence of control policies as $n \to \infty$. The limiting dynamics of the sample paths are characterized by equations (EC.11)–(EC.17). Equation (EC.11) is similar to the evolution of fluid limits under general scheduling policies expressed in terms of cumulative time allocation to each class (see e.g., Dai and Tezcan 2011, Appendix B) and holds for any sequence of admissible controls subject to the limiting relations (EC.12)–(EC.15). Equation (EC.17) is specific to discrete-review policies and establishes the preservation of the discrete-review structure in the limit. Note that although discrete-review controls do not have continuous sample paths in the limit, since time remains unscaled, the convergence in (EC.16) is uniform. We use this proposition later in developing asymptotic lower bounds and establishing the asymptotic optimality of our proposed policy.

Proposition EC.1. *Consider any sequence of admissible controls $\{U^n\}$ and the associated processes satisfying (EC.1)–(EC.7). Assume that Assumptions 1 and 2 hold. Then, for almost all sample paths, the sequence $\{(\bar{X}^n, \bar{T}^n, \bar{Y}^n, \bar{\Upsilon}^n)\}$ has a convergent subsequence $\{(\bar{X}^{n_j}, \bar{T}^{n_j}, \bar{Y}^{n_j}, \bar{\Upsilon}^{n_j})\}$. That is,*

$$\left(\bar{X}^{n_j}(t), \bar{T}^{n_j}(t), \bar{Y}^{n_j}(t), \bar{\Upsilon}^{n_j}(t)\right) \to \left(\bar{X}(t), \bar{T}(t), \bar{Y}(t), \bar{\Upsilon}(t)\right), \quad \textit{u.o.c.}, \tag{EC.10}$$

where $\bar{X}(t)$ and $\bar{T}(t)$ satisfy

$$\bar{X}_i(t) = \bar{X}_i(0) + \bar{\lambda}_i t - \mu_i \bar{T}_i(t) \geq 0, \quad \textit{for all } i \in \mathcal{I}. \tag{EC.11}$$

Furthermore, the functions $\bar{X}(t), \bar{T}(t), \bar{Y}(t), \bar{\Upsilon}(t)$ are absolutely continuous and time derivatives $\bar{Z}(t) \equiv \dot{\bar{T}}(t)$, $\bar{U}(t) \equiv \dot{\bar{Y}}(t)$, and $\bar{Q}(t) \equiv \dot{\bar{\Upsilon}}(t)$ exist a.e. and satisfy the following:

$$\bar{Q}(t) = \bar{X}(t) - \bar{Z}(t) \geq 0, \tag{EC.12}$$

$$\bar{Z}(t) = \bar{U}(t) \wedge \bar{X}(t) \tag{EC.13}$$

$$\bar{U}(t) \cdot e \leq 1, \tag{EC.14}$$

$$\bar{U}(t) \geq 0. \tag{EC.15}$$

Finally, if $\{U^n\}$ is a sequence of discrete-review policies, i.e., it satisfies (EC.8), then for almost all sample paths,

$$\bar{U}^{n_j}(t) \to \bar{U}(t), \quad \textit{u.o.c.}, \tag{EC.16}$$

where

$$\bar{U}(t) = \sum_{k=0}^{\bar{k}(t)-1} \bar{U}[k] \mathbf{1}_{[t_k, t_{k+1})}(t), \tag{EC.17}$$

for some $\{\bar{U}[k]; 0 \leq k \leq \bar{k}(t) - 1\}$.

Proof of Proposition EC.1. To prove the first part we need to show that, for almost all sample paths ω, the sequence $\{(\bar{X}^n(\cdot,\omega), \bar{T}^n(\cdot,\omega), \bar{Y}^n(\cdot,\omega), \bar{\Upsilon}^n(\cdot,\omega))\}$ is pre-compact as $n \to \infty$ in the Skorohod path space endowed with the *u.o.c.* topology. In the following, we fix a sample path and omit ω from the notation for brevity; all statements hold a.s.

First, the pre-compactness of $\{(\bar{X}^n(\cdot), \bar{T}^n(\cdot)\}$ and the relation (EC.11) can be shown similarly as in Theorem B.1. of Dai and Tezcan (2011) using the fact that for all n and $i \in \mathcal{I}$,

$$\bar{T}_i^n(t_2) - \bar{T}_i^n(t_1) \leq t_2 - t_1, \quad 0 \leq t_1 \leq t_2. \tag{EC.18}$$

It follows that there exists a further subsequence, say $\{n_j; j \in \mathbb{N}\}$, such that

$$\left(\bar{X}^{n_j}(\cdot), \bar{T}^{n_j}(\cdot)\right) \to \left(\bar{X}(\cdot), \bar{T}(\cdot)\right), \quad u.o.c.,$$

as $j \to \infty$ and $\bar{X}(t)$ and $\bar{T}(t)$ satisfy (EC.11). Next, consider the sequence of scaled cumulative assignment processes $\{\bar{Y}^{n_j}(\cdot)\}$. Then, for all n and $i \in \mathcal{I}$, and noting that $\bar{U}_i^n(t) \leq 1$, we have

$$\bar{Y}_i^n(t_2) - \bar{Y}_i^n(t_1) \leq t_2 - t_1, \quad 0 \leq t_1 \leq t_2. \tag{EC.19}$$

Therefore, $\{\bar{Y}^{n_j}(\cdot)\}$ is also pre-compact and there exists a further subsequence that converges u.o.c. to some function $\bar{Y}(t)$. To simplify notation, we assume that the convergence occurs on the same subsequence, that is

$$\bar{Y}^{n_j}(t) \to \bar{Y}(t), \quad u.o.c. \tag{EC.20}$$

Finally, consider the scaled sequence of cumulative queue length processes $\{\bar{\Upsilon}^{n_j}(\cdot)\}$. We proceed using a similar argument to that in the proof of Theorem B.1. in Dai and Tezcan 2011. Fix $T > 0$ and pick j_0 large enough so that for $j > j_0$,

$$\max_{i \in \mathcal{I}} \sup_{t \in [0,T]} \left| n_j^{-1} A_i(\lambda_i^{n_j} t) - \bar{\lambda}_i t \right| < \epsilon,$$

for $\epsilon < 1$ (where j_0 exists by FSLLN for A_i). It follows that for $i \in \mathcal{I}$ and $j > j_0$, $n_j^{-1} A_i(\lambda_i^{n_j} t) \leq M' \equiv 1 + T\bar{\lambda}_i$. Therefore, using (EC.1) and (EC.3),

$$\begin{aligned} \bar{\Upsilon}_i^{n_j}(t_2) - \bar{\Upsilon}_i^{n_j}(t_1) &= \int_0^{t_2} \bar{Q}_i^{n_j}(s) ds - \int_0^{t_1} \bar{Q}_i^{n_j}(s) ds \\ &\leq \int_{t_1}^{t_2} n_j^{-1} A_i(\lambda_i^{n_j} s) ds \\ &\leq (t_2 - t_1) M', \quad 0 \leq t_1 \leq t_2 \leq T, \end{aligned} \tag{EC.21}$$

and hence (again simplifying notation by assuming that the convergence occurs on n_j) we have $\bar{\Upsilon}^{n_j}(\cdot) \to \bar{\Upsilon}(\cdot)$ u.o.c., which concludes the first part of the proof.

Next, it follows respectively from (EC.18), (EC.19) and (EC.21) that the functions $\bar{T}(\cdot)$, $\bar{Y}(\cdot)$ and $\bar{\Upsilon}(\cdot)$ are Lipschitz and therefore they are absolutely continuous and their time-derivatives denoted by $\bar{Z}(\cdot)$, $\bar{U}(\cdot)$, and $\bar{Q}(\cdot)$ exists a.e. as claimed. The absolute continuity of $\bar{X}(t)$ then follows from the relation between $\bar{X}(\cdot)$ and $\bar{T}(\cdot)$ in (EC.11) which was established earlier.

We next turn to establishing (EC.12)–(EC.15). Define the *cumulative idle time process* for all servers assigned to pool $i \in \mathcal{I}$ up to time t as

$$I_i^n(t) \equiv \int_0^t (U_i^n(s) - Z_i^n(s))ds = Y_i^n(t) - T_i^n(t),$$

with $\bar{I}^n(t) = I^n(t)/n$ and note that, $\bar{I}^{n_j}(t) \to \bar{I}(t) \equiv \bar{Y}(t) - \bar{T}(t)$ u.o.c. where $\bar{I}(t)$ is Lipschitz. Now, observe that by (EC.3)–(EC.4),

$$\begin{aligned}\bar{\Upsilon}_i^{n_j}(t) &= \int_0^t (\bar{X}_i^{n_j}(s) - \bar{Z}_i^{n_j}(s))ds,\\ \bar{I}_i^{n_j}(t) &= \int_0^t (\bar{U}_i^{n_j}(s) - \bar{Z}_i^{n_j}(s))ds,\end{aligned}$$

are non-decreasing in t for each n_j. It follows that the limits $\bar{\Upsilon}_i(\cdot)$ and $\bar{I}_i(\cdot)$ are also non-decreasing for $i \in \mathcal{I}$, implying that

$$\bar{Q}(t) = \bar{X}(t) - \bar{Z}(t) \geq 0, \tag{EC.22}$$

$$\bar{U}(t) - \bar{Z}(t) \geq 0, \tag{EC.23}$$

a.e.. Next, (EC.5) implies that for $i \in \mathcal{I}$,

$$\int_0^t \bar{Q}_i^{n_j}(s) d\bar{I}_i^{n_j}(s) = 0,$$

and hence using Lemma 4.4. in Dai (1995), we have

$$\int_0^t \bar{Q}_i(s) d\bar{I}_i(s) = 0.$$

It follows that for $i \in \mathcal{I}$,

$$\bar{Q}_i(t)(\bar{U}_i(t) - \bar{Z}_i(t)) = 0, \tag{EC.24}$$

holds a.e., which together with (EC.22)–(EC.23) establishes (EC.13). Furthermore, observing that

$$\int_0^t (1 - \bar{U}^{n_j}(s) \cdot \mathrm{e})ds = t - \sum_{i \in \mathcal{I}} \bar{Y}_i^{n_j}(t),$$

is non-decreasing by (EC.6) we can conclude that (EC.14) holds a.e.. A similar argument establishes (EC.15) a.e..

Finally, assume that (EC.8) is satisfied. Then for a fixed $T > 0$ we have

$$\bar{U}^{n_j}(t) = \sum_{k=0}^{\bar{k}(T)-1} \bar{U}^{n_j}[k]\mathbf{I}_{[t_k,t_{k+1})}(t), \quad t \in [0,T],$$

for some $\{(\bar{U}^{n_j}[0], \dots, \bar{U}^{n_j}[\bar{k}(T)-1])\}$. Since for each k $\bar{U}^{n_j}[k]$ is uniformly bounded, there exists a further subsequence, say again n_j, such that as $j \to \infty$

$$(\bar{U}^{n_j}[0], \dots, \bar{U}^{n_j}[\bar{k}(T)-1]) \to (\bar{U}[0], \dots, \bar{U}[\bar{k}(T)-1]), \tag{EC.25}$$

for some $(\bar{U}[0], \dots, \bar{U}[\bar{k}(T)-1])$. Now, use this vector to define

$$\bar{U}(t) = \sum_{k=0}^{\bar{k}(T)-1} \bar{U}[k]\mathbf{1}_{[t_k,t_{k+1})}(t), \quad t \in [0,T].$$

We need to show that $\left|\left|\bar{U}^{n_{j'}}(t) - \bar{U}(t)\right|\right|_T \to 0$ as $j \to \infty$. To this end, observe that for each $i \in \mathcal{I}$,

$$\begin{aligned}
\left|\bar{U}_i^{n_{j'}}(t) - \bar{U}_i(t)\right|_T &= \sup_{t\in[0,T]} \left| \sum_{k=0}^{\bar{k}(T)-1} \left(\bar{U}_i^{n_j}[k] - \bar{U}_i[k]\right) \mathbf{1}_{[t_k,t_{k+1})}(t) \right| \\
&\leq \sup_{t\in[0,T]} \sum_{k=0}^{\bar{k}(T)-1} \left| \left(\bar{U}_i^{n_j}[k] - \bar{U}_i[k]\right)\right) \mathbf{1}_{[t_k,t_{k+1})}(t) \right| \\
&\leq (\bar{k}(T)-1) \max_{0\leq k\leq(\bar{k}(T)-1)} \left| \left(\bar{U}_i^{n_j}[k] - \bar{U}_i[k]\right) \right|,
\end{aligned}$$

which vanishes as $j \to \infty$ by (EC.25) and noting that $\bar{k}(T) < \infty$. It follows that

$$\left|\left|\bar{U}^{n_j}(t) - \bar{U}(t)\right|\right|_T = \sup_{t\in[0,T]} \sum_{i=0}^{I} \left|\bar{U}_i^{n_j}(t) - \bar{U}_i(t)\right| \leq \sum_{i=0}^{I} \left|\bar{U}_i^{n_j}(t) - \bar{U}_i(t)\right|_T,$$

also goes to zero as $j' \to \infty$, which proves the claim. □

EC.2. Single-stage Cost and Transition Functions in the Presence of Abandonment and under Stationary Arrivals

Starting with initial condition x at the beginning of a stage and under allocation vector u, the dynamics of the system over a shift are described using the following IVP:

$$\begin{aligned}
\dot{x}(t) &= \bar{\lambda} - \hat{\mu} \cdot (x(t) \wedge u[k]) - \hat{\theta} \cdot (x(t) - u[k])^+, \\
x(0) &= x.
\end{aligned} \tag{EC.26}$$

Similar to the case without abandonment, considering different cases depending on whether the queue will empty during the stage or not, one can solve the IVP. The solution then allows us to find the state of the fluid at the beginning of the next stage, as well as the incurred cost during the stage. With a slight abuse of notation we use the same notation as in the case without abandonment to express the dynamics over a stage and the single-stage cost function. The state of class i at the beginning of the next stage is given by,

$$f_i(x,u,\tau)=\begin{cases}\frac{1}{\theta_i}\left[e^{-\theta_i\tau}\left(x_i\theta_i+(e^{\theta_i\tau}-1)\left(\bar{\lambda}_i+u_i(\theta_i-\mu_i)\right)\right)\right], & ,u_i<x_i \text{ and } u_i<\hat{u}_i,\\ \bar{\rho}_i+e^{-\mu_i(\tau-\sigma_i)}(u_i-\bar{\rho}_i), & ,u_i<x_i \text{ and } u_i\geq\hat{u}_i,\\ \frac{1}{\theta_i}\left[\bar{\lambda}_i(1-e^{-(\tau-\nu_i)\theta_i})+u_i(\theta_i-\mu_i(1-e^{-(\tau-\nu_i)\theta_i})\right], & ,u_i\geq x_i \text{ and } u_i<\check{u}_i,\\ \bar{\rho}_i+e^{-\mu_i\tau_i}(x_i-\bar{\rho}_i), & ,u_i\geq x_i \text{ and } u_i\geq\check{u}_i,\end{cases}$$

where,

$$\hat{u}_i=\frac{x_i+(\bar{\lambda}_i/\theta_i)(e^{\tau\theta_i}-1)}{1+(\mu_i/\theta_i)(e^{\tau\theta_i}-1)},\quad \check{u}_i=\bar{\rho}_i+e^{-\mu_i\tau}(x-\bar{\rho}_i),$$

and

$$\sigma_i=\frac{1}{\theta_i}\log\left[\frac{(u_i-x_i)\theta_i+\bar{\lambda}_i-u_i\mu_i}{\bar{\lambda}_i-u_i\mu_i}\right],\quad \nu_i=\frac{1}{\mu_i}\log\left(\frac{\bar{\rho}_i-x_i}{\bar{\rho}_i-u_i}\right).$$

The single-stage cost function is $g(x,u)=\tilde{h}\cdot c(x,u)$, where $c(x,u)$ is given by,

$$c_i(x,u)=\begin{cases}\frac{e^{-\theta_i\tau}\left(\bar{\lambda}_i+\theta_iu_i-\mu_iu_i+e^{\theta_i\tau}(\bar{\lambda}_i(\theta_i\tau-1)-u_i(\theta_i\mu_i\tau+\theta_i-\mu_i)+\theta_ix_i)-\theta_ix_i\right)}{\theta_i^2}, & u_i<x_i,u_i<\hat{u}_i,\\ \frac{(\bar{\lambda}_i-\mu_iu_i)\log\left(\frac{\theta_i(x_i-u_i)}{\mu_iu_i-\bar{\lambda}_i}+1\right)+\theta_i(x_i-u_i)}{\theta_i^2}, & u_i<x_i,u_i\geq\hat{u}_i,\\ \frac{(\bar{\lambda}_i-\mu_iu_i)\left(\theta_i\mu_i\tau-\mu_i+\mu_ie^{-\theta_i\tau}\left(\frac{\bar{\lambda}_i-\mu_ix_i}{\bar{\lambda}_i-\mu_iu_i}\right)^{\theta_i/\mu_i}-\theta_i\log\left(\frac{\bar{\lambda}_i-\mu_ix_i}{\bar{\lambda}_i-\mu_iu_i}\right)\right)}{\theta_i^2\mu_i}, & u_i\geq x_i,u_i<\check{u}_i,\\ 0, & u_i\geq x_i,u_i\geq\check{u}_i.\end{cases}$$

EC.3. Proofs

EC.3.1. Proof of Theorem 1: Asymptotic Lower-Bounds

The proof of Theorem 1 extends the argument in Maglaras (2000) to a multiserver setting under our modified specification of control policies.

Proof of Theorem 1. Consider any sequence of admissible controls $\{U^n\}$. Then for almost all sample paths ω, since $\{n^{-1}C_T^n(U^n)\}$ is a sequence in $\mathbb{R}_+$ it has a subsequence, say $\{n_j; j\in\mathbb{N}\}$, that converges to its lim inf, i.e.,

$$\liminf_{n\to\infty} n^{-1}C_T^n(U^n)=\lim_{j\to\infty} n_j^{-1}C_T^{n_j}(U^{n_j}). \tag{EC.27}$$

Applying Proposition EC.1 to this subsequence we have that there exists a further subsequence say $\{n_{j'}; j'\in\mathbb{N}\}$ such that the scaled sequence of corresponding processes $\{\left(\bar{X}^{n_{j'}}(\cdot),\bar{T}^{n_{j'}}(\cdot),\bar{Y}^{n_{j'}}(\cdot),\bar{\Upsilon}^{n_{j'}}(\cdot)\right)\}$ converges to $\left(\bar{X}(\cdot),\bar{T}(\cdot),\bar{Y}(\cdot),\bar{\Upsilon}(\cdot)\right)$ u.o.c. with the limiting dynamics satisfying (EC.11)–(EC.15) a.e. as well as (EC.16) and (EC.17) if $U^n\in\Pi_\tau^n$ for all n. This together with absolute continuity of $\bar{X}$ implies that $(\bar{X}(t),\bar{U}(t))$ is an admissible pair for Problem 1

(10–15) starting from $x^0 = \bar{X}(0,\omega)$. Further, if $U^n \in \Pi^n_\tau$ then $\bar{U}(t)$ is piecewise-constant and hence $(\bar{X}(t), \bar{U}(t))$ is also admissible for Problem 2 (17–21). Therefore, we have a.s.,

$$h \cdot \bar{\Upsilon}(T) = \int_0^T h \cdot \bar{Q}(s)ds \geq J_T(\bar{X}(0)), \tag{EC.28}$$

$$h \cdot \bar{\Upsilon}(T) = \int_0^T h \cdot \bar{Q}(s)ds \geq J_{T,\tau}(\bar{X}(0)), \text{ if } U^n \in \Pi^n_\tau. \tag{EC.29}$$

Going back to (EC.27) and noting that $n_{j'}$ is a convergent subsequence we have

$$\lim_{j\to\infty} n_j^{-1} C_T^{n_j}(U^{n_j}) = \lim_{j'\to\infty} n_{j'}^{-1} C_T^{n_{j'}}(U^{n_{j'}}) = \lim_{j'\to\infty} h \cdot \bar{\Upsilon}^{n_{j'}}(T) = h \cdot \bar{\Upsilon}(T). \tag{EC.30}$$

It follows from (EC.27) and (EC.28)–(EC.29) that a.s.,

$$\liminf_{n\to\infty} n^{-1} C_T^n(U^n) \geq J_T(\bar{X}(0)), \tag{EC.31}$$

$$\liminf_{n\to\infty} n^{-1} C_T^n(U^n) \geq J_{T,\tau}(\bar{X}(0)), \text{ if } U^n \in \Pi^n_\tau. \tag{EC.32}$$

Further, since any solution to Problem 2 is feasible for Problem 1, we have a.s.,

$$J_{T,\tau}(\bar{X}(0)) \geq J_T(\bar{X}(0)). \tag{EC.33}$$

Taking expectation in (EC.31) and (EC.32) and applying Fatou's Lemma we get that

$$\liminf_{n\to\infty} n^{-1}\mathbb{E}[C_T^n(U^n)] \geq \mathbb{E}[\liminf_{n\to\infty} n^{-1} C_T^n(U^n)] \geq \mathbb{E}[J_T(\bar{X}(0))],$$

$$\liminf_{n\to\infty} n^{-1}\mathbb{E}[C_T^n(U^n)] \geq \mathbb{E}[\liminf_{n\to\infty} n^{-1} C_T^n(U^n)] \geq \mathbb{E}[J_{T,\tau}(\bar{X}(0))], \text{ if } U^n \in \Pi^n_\tau,$$

establishing (26) and (27). Taking expectation in (EC.33) we have $\mathbb{E}[J_{T,\tau}(\bar{X}(0))] \geq \mathbb{E}[J_T(\bar{X}(0))]$ as claimed in (28). □

EC.3.2. Proof of Proposition 1 and Theorem 2: Properties of the Value Function

Proof of Proposition 1. To prove the claims we apply the "Maximum Theorem" (see Ok 2007, Page 306) to the minimization problem in (36), i.e.,

$$V^k(x) = \min_{u\in\mathcal{U}} \tilde{V}^k(x,u) \equiv g(x,u) + V^{k+1}(f(x,u,\tau)), \tag{EC.34}$$

for $k \in \mathcal{K}$. For our problem, the Maximum Theorem states that if $\tilde{V}^k(x,u)$ is continuous and $\mathcal{U}$ is compact, then $V^k(x)$ is also continuous and the minimizer, $\phi^k(x)$, is a non-empty, compact-valued, and upper hemicontinuous correspondence as claimed. Note that $\mathcal{U}$ is clearly compact and recall from Section 4.2.1 that $g(x,u)$ and $f(x,u,\tau)$ are both continuous in (x,u). For $k = N-1$, we have $\tilde{V}^{N-1}(x,u) \equiv g(x,u)$ which is continuous and, therefore, the claims directly follow. Now assume that $V^{k+1}(\cdot)$ is continuous for some $k \leq N-2$. It follows from (EC.34) and using the continuity of $f(x,u,\tau)$ and $g(x,u)$ that $\tilde{V}^k(x,u)$ is also continuous. Therefore, the claims follow by induction for all $k \in \mathcal{K}$. The proof is complete. □

We need the following lemma before presenting the proof of Theorem 2.

Lemma EC.1. *The function $f_i(x,u,t)$ is,*

(i) convex in (x,u) for all $i\in\mathcal{I}$, $t\geq 0$,

(ii) nondecreasing in x for all $i\in\mathcal{I}$, $t\geq 0$, $u\in\mathcal{U}$,

(iii) nonincreasing in u for all $i\in\mathcal{I}$, $t\geq 0$, $x\in\mathbb{R}_+^I$.

Proof. Since for fixed t the function f_i only depends on $(x_i,u_i)\in\mathbb{R}_+^2$, it suffices to prove the claims for the piecewise-defined $\mathbb{R}_+\times[0,1]\to\mathbb{R}_+$ map,

$$\tilde{f}(x_i,u_i)=\begin{cases} x_i+(\bar{\lambda}_i-\mu_i u_i)t, & (x_i,u_i)\in\mathcal{A}_1,\\ \bar{\rho}_i+e^{-\mu_i(t-\sigma_i)}(u_i-\bar{\rho}_i), & (x_i,u_i)\in\mathcal{A}_2,\\ u_i+(\bar{\lambda}_i-u_i\mu_i)(t-\nu_i), & (x_i,u_i)\in\mathcal{A}_3,\\ \bar{\rho}_i+e^{-\mu_i t}(x_i-\bar{\rho}_i), & (x_i,u_i)\in\mathcal{A}_4,\end{cases}\tag{EC.35}$$

with

$$\begin{aligned}\mathcal{A}_1&\equiv\{(x_i,u_i);0\leq u_i\leq x_i \text{ and } u_i\leq(x_i+t\bar{\lambda}_i)/(1+t\mu_i)\},\\ \mathcal{A}_2&\equiv\{(x_i,u_i);0\leq u_i\leq x_i \text{ and } u_i\geq(x_i+t\bar{\lambda}_i)/(1+t\mu_i)\},\\ \mathcal{A}_3&\equiv\{(x_i,u_i);0\leq x_i\leq u_i\leq 1 \text{ and } u_i\leq\bar{\rho}_i+e^{-\mu_i t}(x_i-\bar{\rho}_i)\},\\ \mathcal{A}_4&\equiv\{(x_i,u_i);0\leq x_i\leq u_i\leq 1 \text{ and } u_i\geq\bar{\rho}_i+e^{-\mu_i t}(x_i-\bar{\rho}_i)\},\end{aligned}\tag{EC.36}$$

obtained by fixing $t>0$ in (31) and substituting $\hat{u}$ and $\check{u}$ from (32) to define the four regions $\mathcal{A}_j, j=1,2,3,4$. The lines specified in the four regions intersect at $(\bar{\rho}_i,\bar{\rho}_i)$ creating a maximum of four convex subsets of $\mathbb{R}^2$. Note that in defining the regions we have used the continuity of $\tilde{f}$ (recall that f_i is continuous in (x,u)) and included the boundaries of each region (compared to the 4 cases in (31)) such that the resulting subsets are also closed. (This will be useful later in the proof.) To make the functions $\tilde{f}^j, j=2,3$, well-defined at the boundaries, we set $\nu_i(\bar{\rho}_i,\bar{\rho}_i)=0$ and $\sigma_i(\bar{\rho}_i,\bar{\rho}_i)=0$.

To show convexity, we use a result from Bauschke et al. (2016) which provides sufficient conditions for convexity of piecewise-defined functions, provided that each component is convex in the interior of the corresponding region. To this end, we first compute the gradient of $\tilde{f}$ in the interior (int) of each region. Denote the function specified in (EC.35) for region $\mathcal{A}_j$ by $\tilde{f}^j$ and the gradient of $\tilde{f}$ in the interior of region A_j by $\nabla\tilde{f}^j$. We have,

$$\nabla\tilde{f}^j(x_i,u_i)=\begin{cases}(1,-\mu_i t), & (x_i,u_i)\in\text{int }\mathcal{A}_1,\\ (e^{-\mu_i(t-\sigma_i)},-\mu_i\sigma_i e^{-\mu_i(t-\sigma_i)}), & (x_i,u_i)\in\text{int }\mathcal{A}_2,\\ (e^{-\mu_i\nu_i},-\mu_i(t-\nu_i)) & (x_i,u_i)\in\text{int }\mathcal{A}_3,\\ (e^{-\mu_i t},0) & (x_i,u_i)\in\text{int }\mathcal{A}_4.\end{cases}$$

The convexity of $\tilde{f}^j$ in the interior of each region can be easily verified by checking that for each j the Hessian of $\tilde{f}^j$ is positive semi-definite. To establish the convexity of the function over the whole domain we verify the conditions specified in Theorem 5.4 of Bauschke et al. (2016). First, the functions $\{\tilde{f}^j\}$ in (EC.35) form a system of *compatible functions* (Definition 4.1 in Bauschke et al. 2016) since $\tilde{f}$ is continuous and $\tilde{f}^j(x_i,u_i) = \tilde{f}^{j'}(x_i,u_i)$ for $(x_i,u_i) \in \mathcal{A}_j \cap \mathcal{A}_{j'}$ with $j \neq j'$. Second, $\tilde{f}$ is differentiable in the interior of each region. Third, the domain of $\tilde{f}$, i.e., $\mathbb{R}_+ \times [0,1] = \cup_j \mathcal{A}_j$ is convex and two dimensional. Fourth, $\{\mathcal{A}_j\}, j = 1,2,3,4$ is a *compatible system of sets* (Definition 3.1 in Bauschke et al. 2016) since it is comprised of a finite number of closed convex subsets of $\mathbb{R}^2$. Finally, we verify that except at a finite subset of the interior points, i.e., the intersection of all regions $\{(\bar{\rho}_i,\bar{\rho}_i)\}$, the following limits of the gradients exist along the boundaries of the regions:

$$\lim_{\substack{(x_i,u_i)\to(x_i',y_i')\in(A_2\cap A_4)\setminus(\bar{\rho}_i,\bar{\rho}_i),\\ (x_i,u_i)\in\mathrm{Int}\mathcal{A}_2}} \nabla\tilde{f}^2(x_i,u_i) = \lim_{\substack{(x_i,u_i)\to(x_i',y_i')\in(A_2\cap A_4)\setminus(\bar{\rho}_i,\bar{\rho}_i),\\ (x_i,u_i)\in\mathrm{Int}\mathcal{A}_4}} \nabla\tilde{f}^4(x_i,u_i) = (e^{-\mu_i t},0),$$

$$\lim_{\substack{(x_i,u_i)\to(x_i',y_i')\in(A_1\cap A_3)\setminus(\bar{\rho}_i,\bar{\rho}_i),\\ (x_i,u_i)\in\mathrm{Int}\mathcal{A}_1}} \nabla\tilde{f}^1(x_i,u_i) = \lim_{\substack{(x_i,u_i)\to(x_i',y_i')\in(A_1\cap A_3)\setminus(\bar{\rho}_i,\bar{\rho}_i),\\ (x_i,u_i)\in\mathrm{Int}\mathcal{A}_3}} \nabla\tilde{f}^3(x_i,u_i) = (1,-\mu_i t),$$

$$\lim_{\substack{(x_i,u_i)\to(x_i',y_i')\in(A_1\cap A_2)\setminus(\bar{\rho}_i,\bar{\rho}_i),\\ (x_i,u_i)\in\mathrm{Int}\mathcal{A}_1}} \nabla\tilde{f}^1(x_i,u_i) = \lim_{\substack{(x_i,u_i)\to(x_i',y_i')\in(A_1\cap A_2)\setminus(\bar{\rho}_i,\bar{\rho}_i),\\ (x_i,u_i)\in\mathrm{Int}\mathcal{A}_2}} \nabla\tilde{f}^2(x_i,u_i) = (1,-\mu_i t),$$

$$\lim_{\substack{(x_i,u_i)\to(x_i',y_i')\in(A_4\cap A_3)\setminus(\bar{\rho}_i,\bar{\rho}_i),\\ (x_i,u_i)\in\mathrm{Int}\mathcal{A}_4}} \nabla\tilde{f}^4(x_i,u_i) = \lim_{\substack{(x_i,u_i)\to(x_i',y_i')\in(A_4\cap A_3)\setminus(\bar{\rho}_i,\bar{\rho}_i)\\ (x_i,u_i)\in\mathrm{Int}\mathcal{A}_3}} \nabla\tilde{f}^3(x_i,u_i) = (e^{-\mu_i t},0).$$

It follows that $\tilde{f}(x_i,u_i)$ is convex in (x_i,u_i) and continuously differentiable away from $(\bar{\rho}_i,\bar{\rho}_i)$.

Next, noting that $\sigma_i < t$ and $\nu_i < t$, it is easy to see that for each j, $\nabla\tilde{f}^j_1 \geq 0$ and $\nabla\tilde{f}^j_2 \leq 0$. This together with continuity and a.e. differentiability of $\tilde{f}$ implies its monotonicity in x_i and u_i as claimed. □

Proof of Theorem 2. We first show monotonicity using induction on k. For $k = N-1$, $V^{N-1}(x) = \min_{u\in\mathcal{U}} g(x,u)$. From Lemma EC.1 ($i$) we have that $f_i(\cdot,u,t)$ is nondecreasing. It follows directly that $f(\cdot,u,t)$ and using (16) that $g(\cdot,u)$ is also nondecreasing. Now consider $x,x' \in \mathbb{R}^I_+$ such that $x \geq x'$ and let $\hat{\phi}^k(x)$ denote an arbitrary member of the correspondence $\phi^k(x)$, i.e., $\hat{\phi}^k(x) \in \phi^k(x)$. Observe that

$$\begin{aligned} V^{N-1}(x) &= g(x,\hat{\phi}^{N-1}(x)) \\ &\geq g(x',\hat{\phi}^{N-1}(x)) \\ &\geq \min_{u\in\mathcal{U}} g(x',u) = V^{N-1}(x'), \end{aligned}$$

and hence $V^{N-1}(\cdot)$ is nondecreasing. Next assuming that $V^{k+1}(\cdot)$ is nondecreasing we show that $V^k(\cdot)$ is also nondecreasing. To this end again consider $x \geq x'$ and observe that, using monotonicity of $V^{k+1}(\cdot)$ and $f(\cdot,u,t)$, we have

$$V^k(x) = \min_{u\in\mathcal{U}} \left[g(x,u) + V^{k+1}(f(x,u,\tau))\right]$$

$$
\begin{aligned}
&= g(x,\hat{\phi}^k(x)) + V^{k+1}(f(x,\hat{\phi}^k(x),\tau)) \\
&\geq g(x',\hat{\phi}^k(x)) + V^{k+1}(f(x',\hat{\phi}^k(x),\tau)) \\
&\geq \min_{u\in\mathcal{U}} \left[g(x',u) + V^{k+1}(f(x',u,\tau))\right] \\
&= V^k(x'),
\end{aligned}
$$

which proves the claim.

Next, we show convexity again using induction. Consider $V^{N-1}(x)$. Since by Lemma EC.1 $f_i(\cdot,\cdot,\tau)$ is convex for all $i\in\mathcal{I}$ we have that $f(\cdot,\cdot,\tau)$ is convex. Further, since convexity is preserved under summation and pointwise maximization it follows from (16) that $g(\cdot,\cdot)$ is also convex. Finally, convexity of $V^{N-1}(x)$ follows since convexity is preserved under partial minimization. Next, assume that $V^{k+1}(\cdot)$ is convex. Then since $f(\cdot,\cdot,\tau)$ is convex and $V^{k+1}(\cdot)$ is convex nondecreasing, the composition $V^{k+1}(f(\cdot,\cdot,\tau))$ is also convex (see, e.g., Dattorro 2005), which together with convexity of $g(\cdot,\cdot)$ implies that

$$\tilde{V}^k(x,u) \equiv g(x,u) + V^{k+1}(f(x,u,\tau)), \tag{EC.37}$$

is convex. Therefore, we can conclude from (36) and preservation of convexity under partial minimization that $V^k(\cdot)$ is convex. The claim follows by induction. □

EC.3.3. Proof of Proposition 2 and Theorem 3: Properties of an optimal solution to Problem 2.

Proof of Proposition 2. Recall that an optimal policy (possibly non-unique) always exists. Now assume there exists an optimal policy, say $\pi'\equiv\{u'[k];k\in\mathcal{K}\}$, such that $u'[j]\cdot\mathrm{e}<1$ for some $j\in\mathcal{J}$ where $\mathcal{J}\subseteq\mathcal{K}$ is nonempty. Denote the vector of fluid headcount under policy π' by $\{x'[k];k\in\mathcal{K}\}$ and note that $x[0]=x'[0]=x^0$. Pick $j\in\mathcal{J}$ and construct the policy $\pi\equiv\{u[k];k\in\mathcal{K}\}$ as follows: $u[k]=u'[k]$ for all $k\in\mathcal{K}\setminus\{j\}$, $u_i[j]=u'_i[j]$ for all $i\in\mathcal{I}\setminus\{l\}$ and $u_l[j]=u'_l[j]+(1-u'[j]\cdot\mathrm{e})$ where $l\in\mathrm{argmin}_{i\in\mathcal{I}}u'[j]$. That is the policy π uses the same allocation as π' except at $k=j$ where it allocates the unused capacity $1-u'[j]\cdot\mathrm{e}$ to the class with minimum allocation (ties can be broken arbitrarily). Note that $u[j]\cdot\mathrm{e}=1$. We next compare the total cost of the two policies. To this end, first note that $x[k]=x'[k]$ and $u[k]=u'[k]$ for $k\in\{0,\dots,j-1\}$ and at $k=j$ we have $u[j]>u'[j]$ and $x[j]=x'[j]$. This implies using Lemma EC.1 part (*iii*) that $x[j+1]=f(x[j],u[j],\tau)\leq f(x'[j],u'[j],\tau)=x'[j+1]$. It follows using Lemma EC.1 (*ii*) and noting that by construction $u[k]=u'[k]$ for $k\in\{j+1,\dots,N-1\}$ that $x[k]\leq x'[k]$ and $u[k]=u'[k]$ for $k\in\{j+1,\dots,N-1\}$. Therefore, since by Lemma EC.1 and (16) the stage cost $g(x,u)$ is nonincreasing in x and nondecreasing in u, policy π achieves the same or a lower cost as π in all stages. If j is the only stage where π' does not use all capacity or the cost of π' is strictly greater than that of π then we are done. Otherwise, set $\pi'\leftarrow\pi$ and apply the same procedure to the next member of $\mathcal{J}$ until either a contradiction is reached or another policy with the same cost as π' is constructed. The proof is complete. □

The following lemma will be useful in the proof of Theorem 3 and states that starting above the stationary point $\bar{\rho}_i$, the class i trajectory $x_i(t)$ will remain above $\bar{\rho}_i$ under any admissible policy. We note that the trajectory could get arbitrarily close to the stationary point, but can never cross below it.

Lemma EC.2. *Assume that the initial condition x^0 in Problem 2 satisfies $x_i^0 > \bar{\rho}_i$ for $i \in \mathcal{I}' \subseteq \mathcal{I}$. Then under any admissible policy the corresponding trajectory satisfies $x_i[k] > \bar{\rho}_i$ for all $i \in \mathcal{I}'$ and $k \in \mathcal{K}$.*

Proof of Lemma EC.2. For each $i \in \mathcal{I}'$ satisfying $x_i^0 > \bar{\rho}_i$ we show using the transition function f_i in (31) that regardless of the allocation $x_i[k] > \bar{\rho}_i$ implies $x_i[k+1] > \bar{\rho}_i$. The claim then follows by induction. We do this by checking that in each case of (31), $x_i > \bar{\rho}_i$ implies $f_i(x,u,\tau) > \bar{\rho}_i$.

Case 1. Note that $u_i \le \hat{u}_i(x,\tau) \equiv (x_i + \tau\bar{\lambda}_i)/(1+\tau\mu_i)$ together with $x_i > \bar{\rho}_i$ implies that $x_i > \hat{u}_i(x,\tau) \ge u_i$. Hence, we have

$$f_i(x,u,\tau) = x_i + (\bar{\lambda}_i - \mu_i u_i)\tau \ge x_i(1-\mu_i\tau) + \bar{\lambda}_i\tau > \bar{\rho}_i.$$

Case 2. Clearly, it suffices to show $u_i - \bar{\rho}_i > 0$. This however follows noting that $u_i > \hat{u}_i(x,\tau) \equiv (x_i + \tau\bar{\lambda}_i)/(1+\tau\mu_i)$ and $u_i < x_i$ imply $u_i > (u_i + \tau\bar{\lambda}_i)/(1+\tau\mu_i)$, which yields $u_i > \bar{\rho}_i$.

Case 3. The conditions of this case, i.e., $x_i \le u_i < \check{u}_i(x,\tau) \equiv \bar{\rho}_i + e^{-\mu_i\tau}(x_i - \bar{\rho}_i)$ cannot be satisfied together with $x_i > \bar{\rho}_i$. To see this note that $x_i \le u_i < \check{u}_i(x,\tau)$ implies $x_i - \bar{\rho}_i \le u_i - \bar{\rho}_i < e^{-\mu_i\tau}(x_i - \bar{\rho}_i)$, which clearly cannot hold with $x_i - \bar{\rho}_i > 0$. Hence, starting from $x_i > \bar{\rho}_i$ we cannot be in Case 3 under any allocation.

Case 4. The claim follows directly using the assumption $x_i > \bar{\rho}_i$. This concludes the proof. □

Proof of Theorem 3. We show that the property holds for the first stage (at $k=0$) for any number of stages $N \in \mathbb{N}$ and starting from $x^0 = x > \bar{\rho} \in \mathbb{R}_+^2$. Then since the dynamics and stage cost are independent of k, the problem at stage $k+1$ is equivalent to that in the first stage for a $N-k$ stage problem and hence the claim follows for all k. Note that by Lemma EC.2 $x > \bar{\rho}$ implies, regardless of the allocations, that $x[k] > \bar{\rho}$ for all $k \in \mathcal{K}$.

For simplicity and with a slight abuse of notation, in the rest of the proof we write $f(x,u)$ for $f(x,u,\tau)$, $\hat{u}_i(x)$ for $\hat{u}_i(x,\tau)$, and $\check{u}_i(x)$ for $\check{u}_i(x,\tau)$.

Consider the value function at $k=0$ for the N stage problem denoted by $V_N^0(\cdot)$. From Proposition 2 we know that there exists an optimal policy where $u \cdot \mathrm{e} \le 1$ is satisfied with equality in every stage. Adjoining the constraint $u \cdot \mathrm{e} = 1$ in (EC.37) we have

$$V_N^0(x) = \min_{u \ge 0} \left[g(x,u) + V_N^1(f(x,u)) + \beta(u \cdot \mathrm{e} - 1) \right], \tag{EC.38}$$

where $\beta \geq 0$ is the Lagrangian multiplier. The necessary and sufficient optimality conditions then imply that a minimizer of (EC.38) $\hat{\phi}^0(x)$ satisfies the following for $i = 1, 2$,

$$\frac{\partial g(x, \hat{\phi}^0(x))}{\partial u_i} + \frac{\partial V_N^1(f(x, \hat{\phi}^0(x)))}{\partial x_i} \cdot \frac{\partial f_i(x, \hat{\phi}^0(x))}{\partial u_i} + \beta \geq 0, \tag{EC.39}$$

$$\left(\frac{\partial g(x, \hat{\phi}^0(x))}{\partial u_i} + \frac{\partial V_N^1(f(x, \hat{\phi}^0(x)))}{\partial x_i} \cdot \frac{\partial f_i(x, \hat{\phi}^0(x))}{\partial u_i} + \beta \right) \hat{\phi}_i^0(x) = 0, \tag{EC.40}$$

$$\hat{\phi}^0(x) \cdot \mathrm{e} - 1 = 0, \tag{EC.41}$$

$$\hat{\phi}_i^0(x) \geq 0, \tag{EC.42}$$

assuming that the derivatives in (EC.39) and (EC.40) exist on $\{(x, u); x > \bar{\rho}\}$. (Note that the complementary condition in (EC.40) is introduced to handle the nonnegativity constraint $u \geq 0$.) The partials $\partial f_i / \partial u_i$ and $\partial f_i / \partial x_i$ are given by

$$\left(\frac{\partial}{\partial x_i}, \frac{\partial}{\partial u_i}\right) f_i = \begin{cases} (1, -\mu_i t), & u_i < x_i \text{ and } u_i \leq \hat{u}_i(x), \\ (e^{-\mu_i(t-\sigma_i)}, -\mu_i \sigma_i e^{-\mu_i(t-\sigma_i)}), & u_i < x_i \text{ and } u_i > \hat{u}_i(x), \\ (e^{-\mu_i \nu_i}, -\mu_i(t - \nu_i)), & u_i \geq x_i \text{ and } u_i < \check{u}_i(x), \\ (e^{-\mu_i t}, 0), & u_i \geq x_i \text{ and } u_i \geq \check{u}_i(x). \end{cases} \tag{EC.43}$$

Further, using (34) and differentiating (35) we have

$$\frac{\partial g(x, u)}{\partial u_i} = \begin{cases} -h_i(\tau + \mu_i \tau^2/2), & u_i < x_i \text{ and } u_i \leq \hat{u}_i(x), \\ -h_i(\sigma_i + \mu_i \sigma_i^2/2), & u_i < x_i \text{ and } u_i > \hat{u}_i(x), \\ -h_i\big((\tau - \nu_i) + \mu_i(\tau - \nu_i)^2/2\big), & u_i \geq x_i \text{ and } u_i < \check{u}_i(x), \\ 0, & u_i \geq x_i \text{ and } u_i \geq \check{u}_i(x), \end{cases} \tag{EC.44}$$

and

$$\frac{\partial g(x, u)}{\partial x_i} = \begin{cases} h_i \tau, & u_i < x_i \text{ and } u_i \leq \hat{u}_i(x), \\ h_i \sigma_i, & u_i < x_i \text{ and } u_i > \hat{u}_i(x), \\ h_i(\tau - \nu_i) e^{-\nu_i \mu_i}, & u_i \geq x_i \text{ and } u_i < \check{u}_i(x), \\ 0, & u_i \geq x_i \text{ and } u_i \geq \check{u}_i(x). \end{cases} \tag{EC.45}$$

It is straightforward to verify using the above that for all points (x, u) with $x > \bar{\rho}$ the partial derivatives exist and are continuous. Hence, f and g are both differentiable at any point (x, u) with $x > \bar{\rho}$. The following result establishes the required differentiability of the value function.

Proposition EC.2. *The value function $V_N^k(x)$ is continuously differentiable in the interior of the set $\{x \in \mathbb{R}_+^I; x > \bar{\rho}\}$ for all $k \in \mathcal{K}$.*

To prove the proposition we use the following result from Stokey and Lucas (1989) originally proved in Benveniste and Scheinkman (1979).

Lemma EC.3. *(Benveniste and Scheinkman 1979) Let $Y \subseteq \mathbb{R}^I$ be a convex set, let $V : Y \to \mathbb{R}$ be concave, let $y \in int(Y)$, and let $\mathcal{D}$ be a neighborhood of y. If there is a concave, differentiable function $\omega : \mathcal{D} \to \mathbb{R}$, with $\omega(y) = V(y)$ and with $\omega(y) \leq V(y)$ for all $y \in \mathcal{D}$, then V is differentiable at y.*

Proof of Proposition EC.2. The proof is by induction on k. For $k = N-1$ we have from (36) that $V_N^{N-1}(x) = g(x, \hat{\phi}^0(x))$, where $\hat{\phi}^0(x) \in \phi^0(x)$ is an optimal allocation at stage $k = 0$. Let $x_0 > \bar{\rho}$ and denote by $\mathcal{D}$ a neighborhood of x_0. Define the function ω^{N-1} on $\mathcal{D}$ as $\omega^{N-1}(x) = g(x, \hat{\phi}^0(x_0))$ and note that ω is differentiable and convex, with $\omega(x_0) = V_N^{N-1}(x_0)$ and $\omega(x) \geq V_N^{N-1}(x)$. It follows using Lemma EC.3 (applied to the concave function $-V_N^{N-1}$) that V_N^{N-1} is differentiable in x_0. Next, assume that $V_N^{k+1}(x)$ is differentiable in all points with $x > \bar{\rho}$ for some $k \leq N-2$. Using (36) we have $V_N^k(x) = g(x, \hat{\phi}^k(x)) + V_N^{k+1}(f(x, \hat{\phi}^k(x))$. Similar to the case with $k = N-1$, let $\mathcal{D}$ denote a neighborhood of $x_0 > \bar{\rho}$, and define $\omega^k(x) = g(x, \hat{\phi}^k(x_0)) + V_N^{k+1}(f(x, \hat{\phi}^k(x_0))$ on $\mathcal{D}$. Note that g and f are both differentiable on $\mathcal{D}$ and by Lemma EC.2 $f(x, \hat{\phi}^k(x_0)) > \bar{\rho}$. It follows by the induction assumption that $\omega^k(x)$ is differentiable. Further, since g and V_N^{k+1} are convex, ω^k is also convex. Noting that $\omega^k(x_0) = V_N^k(x_0)$ and $\omega^k(x) \geq V_N^k(x)$ and again applying Lemma EC.3 we have that $V_N^k(x)$ is differentiable at x_0. The proof is complete. □

Define for x with $x > \bar{\rho}$, the set of allocations

$$\widehat{\mathcal{U}}(x) = \{(u_1, u_2); u_1, u_2 \geq 0 \text{ and } u_1 + u_2 = 1 \text{ and } u_1 \geq \hat{u}_1(x) \wedge 1\}.$$

We prove the existence of an optimal solution $\phi^0(x) \in \widehat{\mathcal{U}}(x)$ by showing that any non-wasteful allocation $u \notin \widehat{\mathcal{U}}(x)$ (i.e., with $u_1 + u_2 = 1$) either does not satisfy the optimality equations (EC.39)–(EC.42), or if it does, then there exists another optimal allocation that belongs to $\widehat{\mathcal{U}}(x)$. The proof is by induction on the number of stages N.

Base case. First, consider the *single-stage problem* ($N = 1$). In this case (EC.39) and (EC.40) are simply

$$\frac{\partial g(x, \hat{\phi}^0(x))}{\partial u_i} + \beta \geq 0, \quad \text{and} \quad \left(\frac{\partial g(x, \hat{\phi}^0(x))}{\partial u_i} + \beta\right) \hat{\phi}_i^0(x) = 0, \quad i = 1, 2.$$

Therefore, depending on whether the solution is in the interior or at a corner point it satisfies one of the following:

$$-\frac{\partial g(x, \hat{\phi}^0(x))}{\partial u_1} = -\frac{\partial g(x, \hat{\phi}^0(x))}{\partial u_2} = \beta, \quad \hat{\phi}_1^0(x) > 0, \hat{\phi}_2^0(x) > 0, \tag{EC.46}$$

$$-\frac{\partial g(x, \hat{\phi}^0(x))}{\partial u_1} = \beta, -\frac{\partial g(x, \hat{\phi}^0(x))}{\partial u_2} < \beta, \quad \hat{\phi}_1^0(x) = 1, \hat{\phi}_2^0(x) = 0, \tag{EC.47}$$

$$-\frac{\partial g(x, \hat{\phi}^0(x))}{\partial u_1} < \beta, -\frac{\partial g(x, \hat{\phi}^0(x))}{\partial u_2} = \beta, \quad \hat{\phi}_1^0(x) = 0, \hat{\phi}_2^0(x) = 1, \tag{EC.48}$$

together with $\hat{\phi}_1^0(x) + \hat{\phi}_2^0(x) = 1$. Now consider a non-wasteful allocation $u \notin \widehat{\mathcal{U}}(x)$ (i.e., with $u_1 < \hat{u}_1(x) \wedge 1$) and observe that starting from an initial condition $x > \bar{\rho}$ we must have $u_1 < x_1$. To see this, consider two cases. If $u_1 \leq \bar{\rho}_1$ then trivially $x_1 > \bar{\rho}_1 \geq u_1$, and if $u_1 > \rho_1$ then $u_1 < \hat{u}_1(x) \wedge 1 \leq \hat{u}_1(x) = (x_1 + \tau\bar{\lambda}_1)/(1 + \tau\mu_1)$ implies $x_1 > u_1 + \tau\mu_1(u_1 - \bar{\rho}_1) > u_1$. Therefore, for an allocation

$u \notin \widehat{\mathcal{U}}(x)$ we have $u_1 < x_1$ and $u_1 \leq \hat{u}_1(x)$ and hence we are in Case 1 of (EC.44). Considering the four possible cases for class 2 we observe that

$$-\frac{\partial g(x,u)}{\partial u_1} = h_1(\tau + \mu_1\tau^2/2) \geq -\frac{\partial g(x,u)}{\partial u_2} = \begin{cases} h_2(\tau + \mu_2\tau^2/2), & u_2 < x_2 \text{ and } u_2 \leq \hat{u}_2(x), \\ h_2(\sigma_2 + \mu_2\sigma_2^2/2), & u_2 < x_2 \text{ and } u_2 \leq \hat{u}_2(x), \\ h_2\big((\tau-\nu_2) + \mu_2(\tau-\nu_2)^2/2\big), & u_2 \geq x_2 \text{ and } u_2 < \check{u}_2(x), \\ 0, & u_2 \geq x_2 \text{ and } u_2 \geq \check{u}_2(x), \end{cases}$$

where the inequality follows from assumption (*iii*) in the first case, noting that $\sigma_2 < \tau$ in the second case, and since $0 < \nu_2 < \tau$ in the third case. In words, the inequality implies that the marginal value of allocating capacity to class 1 (in terms of cost *reduction*) is greater than or equal to that of class 2 for any allocation $u \notin \widehat{\mathcal{U}}(x)$. Considering the possible cases in (EC.46)–(EC.48) it follows that the allocation u can only be optimal if it satisfies (EC.46). But since $\sigma_2 < \tau$ and $\nu_2 > 0$, u can only be an optimal solution if

$$-\frac{\partial g(x,u)}{\partial u_1} = h_1(\tau + \mu_1\tau^2/2) = -\frac{\partial g(x,u)}{\partial u_2} = h_2(\tau + \mu_2\tau^2/2),$$

i.e., for class 2 we are also in Case 1 and assumption (*iii*) is satisfied with an equality. In this case however the optimality equation is easily seen to also hold at $(\hat{u}_1(x), 1 - \hat{u}_1(x)) \in \widehat{\mathcal{U}}(x)$ which proves the claim for $N = 1$.

Remark EC.1. As mentioned in the body of the paper, for the single-stage problem, we can show regardless of the initial condition that the marginal value of allocating capacity to class 1 is equal or greater than that of allocating to class 2 as long as no idleness is incurred in class 1. This is easily seen noting that as long as no idleness is incurred, we are in Case 1 of (EC.44) for class 1, which as shown above is an upper bound for the marginal value of adding capacity to class 2 in all possible four cases.

Induction hypothesis. Now assume that the claim holds for the $N-1$ period problem with $N \geq 2$, i.e., there exists an allocation $\hat{\phi}^0(x) \in \widehat{\mathcal{U}}(x)$ that satisfies the optimality equations (EC.39)–(EC.41) with N replaced by $N-1$. Further, differentiating the value function with respect to the initial condition x_i at the optimal allocation $\hat{\phi}^0(x)$ we obtain the following *Envelop Condition* (EC):

$$\frac{\partial V_{N-1}^0(x)}{\partial x_i} = \frac{\partial g(x, \hat{\phi}^0(x))}{\partial x_i} + \frac{\partial f_i(x, \hat{\phi}^0(x))}{\partial x_i} \cdot \frac{\partial V_{N-1}^1(f(x, \hat{\phi}^0(x))}{\partial x_i}, \quad i = 1, 2, \tag{EC.49}$$

which will be instrumental in the next step of the proof.

Induction step. We now turn to the N period problem. We want to show that there exists an optimal allocation $\hat{\phi}^0(x) \in \widehat{\mathcal{U}}(x)$ satisfying (EC.39)–(EC.42). Consider a non-idling allocation $u \notin \widehat{\mathcal{U}}(x)$, i.e., an allocation that does not empty the queue for class 1 by the end of the first stage.

Similar to the single-stage problem we show that for any such allocation the marginal value of allocating capacity to class 1 is no less than that of allocating to class 2; the difference here being that one needs to also consider the effect on future stages. For the N period problem this translates to

$$-\frac{\partial g(x,u)}{\partial u_1} - \frac{\partial V_N^1(f(x,u))}{\partial x_1} \cdot \frac{\partial f_1(x,u)}{\partial u_1} \geq -\frac{\partial g(x,u)}{\partial u_2} - \frac{\partial V_N^1(f(x,u))}{\partial x_2} \cdot \frac{\partial f_2(x,u)}{\partial u_2}. \tag{EC.50}$$

If the above inequality is strict then it follows directly that u does not satisfy the optimality conditions (EC.39)–(EC.41). In the case where (EC.50) is satisfied with equality, we show that there exists another optimal solution which belongs to $\widehat{\mathcal{U}}(x)$.

We proceed by obtaining the terms in (EC.50). First, similar to the single-stage problem, $u \notin \widehat{\mathcal{U}}(x)$ implies that in the first stage we have

$$-\frac{\partial g(x,u)}{\partial u_1} = h_1(\tau + \mu_1\tau^2/2), \quad -\frac{\partial g(x,u)}{\partial u_2} \leq h_2(\tau + \mu_2\tau^2/2). \tag{EC.51}$$

Further, from (EC.43) at $t = \tau$ and with $u \notin \widehat{\mathcal{U}}(x)$ we have

$$-\frac{\partial f_1(x,u)}{\partial u_1} = \mu_1\tau, \quad -\frac{\partial f_2(x,u)}{\partial u_2} \leq \mu_2\tau, \tag{EC.52}$$

where the inequality follows easily recalling that $\sigma_i < \tau$ and $\nu_i > 0$. We next obtain explicit expressions for $\partial V_N^1(f(x,u))/\partial x_i, i = 1,2$ required for verifying (EC.50). Denote by $\tilde{x} \equiv f(x,u)$ the initial position at the beginning of the second stage (at $k = 1$) and recall from Lemma EC.2 that the assumption $x > \bar{\rho}$ implies, regardless of the allocation u, that $\tilde{x} > \bar{\rho}$. It follows by the induction assumption, and noting $V_N^1(\tilde{x}) = V_{N-1}^0(\tilde{x})$, that at $k = 1$ there exists an optimal allocation $\phi^1(\tilde{x}) \in \widehat{\mathcal{U}}(\tilde{x})$ satisfying the optimality equations

$$-\frac{\partial g(\tilde{x},\hat{\phi}^1(\tilde{x}))}{\partial u_1} - \frac{\partial V_{N-1}^0(f(\tilde{x},\hat{\phi}^1(\tilde{x})))}{\partial x_1} \cdot \frac{\partial f_1(\tilde{x},\hat{\phi}^1(\tilde{x}))}{\partial u_1} = \beta, \tag{EC.53}$$

$$-\frac{\partial g(\tilde{x},\hat{\phi}^1(\tilde{x}))}{\partial u_2} - \frac{\partial V_{N-1}^0(f(\tilde{x},\hat{\phi}^1(\tilde{x})))}{\partial x_2} \cdot \frac{\partial f_2(\tilde{x},\hat{\phi}^1(\tilde{x}))}{\partial u_2} \leq \beta, \tag{EC.54}$$

where (EC.54) is satisfied with equality if $\hat{\phi}^1(\tilde{x})$ is an interior solution and is strict if it is the corner solution $\hat{\phi}_1^1(\tilde{x}) = 1, \hat{\phi}_2^1(\tilde{x}) = 0$. Noting that $\partial f_2(\tilde{x},\hat{\phi}^1(\tilde{x}))/\partial u_2 \leq 0$ (see (EC.43)) this implies that when $\partial f_2(\tilde{x},\hat{\phi}^1(\tilde{x}))/\partial u_2 \neq 0$,

$$\frac{\partial V_{N-1}^0(f(\tilde{x},\hat{\phi}^1(\tilde{x})))}{\partial x_1} = \left(-\beta - \frac{\partial g(\tilde{x},\hat{\phi}^1(\tilde{x}))}{\partial u_1}\right) \Bigg/ \frac{\partial f_1(\tilde{x},\hat{\phi}^1(\tilde{x}))}{\partial u_1},$$

$$\frac{\partial V_{N-1}^0(f(\tilde{x},\hat{\phi}^1(\tilde{x})))}{\partial x_2} \leq \left(-\beta - \frac{\partial g(\tilde{x},\hat{\phi}^1(\tilde{x}))}{\partial u_2}\right) \Bigg/ \frac{\partial f_2(\tilde{x},\hat{\phi}^1(\tilde{x}))}{\partial u_2}.$$

Inserting the above in the EC (EC.49) starting from $\tilde{x}$ and with optimal allocation $\hat{\phi}^1(\tilde{x})$, and using $V_N^1(\tilde{x}) = V_{N-1}^0(\tilde{x})$ we get that when $\partial f_2(\tilde{x}, \hat{\phi}^1(\tilde{x}))/\partial u_2 \neq 0$,

$$\frac{\partial V_N^1(\tilde{x})}{\partial x_1} = \frac{\partial g(\tilde{x}, \hat{\phi}^1(\tilde{x}))}{\partial x_1} + \left(-\beta - \frac{\partial g(\tilde{x}, \hat{\phi}^1(\tilde{x}))}{\partial u_1}\right)\left(\frac{\partial f_1(\tilde{x}, \hat{\phi}^1(\tilde{x}))}{\partial x_1} \Bigg/ \frac{\partial f_1(\tilde{x}, \hat{\phi}^1(\tilde{x}))}{\partial u_1}\right), \tag{EC.55}$$

$$\frac{\partial V_N^1(\tilde{x})}{\partial x_2} \leq \frac{\partial g(\tilde{x}, \hat{\phi}^1(\tilde{x}))}{\partial x_2} + \left(-\beta - \frac{\partial g(\tilde{x}, \hat{\phi}^1(\tilde{x}))}{\partial u_2}\right)\left(\frac{\partial f_2(\tilde{x}, \hat{\phi}^1(\tilde{x}))}{\partial x_2} \Bigg/ \frac{\partial f_2(\tilde{x}, \hat{\phi}^1(\tilde{x}))}{\partial u_2}\right), \tag{EC.56}$$

where we note again that the inequality in (EC.56) is strict in case of a corner solution ($\hat{\phi}_1^1(\tilde{x}) = 1, \hat{\phi}_2^1(\tilde{x}) = 0$) and is an equality in case of an interior solution ($\hat{\phi}_1^1(\tilde{x}) > 0, \hat{\phi}_2^1(\tilde{x}) > 0$ and $\hat{\phi}_1^1(\tilde{x}) + \hat{\phi}_2^1(\tilde{x}) = 1$). Explicit expressions for the terms in (EC.55) and (EC.56) are readily available depending on which case the solution $\hat{\phi}^1(\tilde{x})$ belongs to. For class 1, since $\tilde{x}_1 > \bar{\rho}_1$ and $\hat{\phi}^1(\tilde{x}) \in \widehat{\mathcal{U}}(\tilde{x})$ we are either in Case 2 where $\hat{\phi}_1^1(\tilde{x}) < \tilde{x}_1$ and $\hat{\phi}_1^1(\tilde{x}) > \hat{u}_1(\tilde{x})$ and the queue empties at time $0 < \sigma_1 < \tau$ during the stage, or we are in Case 1 with $\hat{\phi}_1^1(\tilde{x}) = \hat{u}_1(\tilde{x}) \wedge 1$ where the optimal policy is emptying the queue τ time units into the stage, or allocating all capacity to it in the case of a corner solution. Therefore, we have from (EC.44) and (EC.45) that

$$\frac{\partial g(\tilde{x}, \hat{\phi}^1(\tilde{x}))}{\partial x_1} = h_1\theta, \quad -\frac{\partial g(\tilde{x}, \hat{\phi}^1(\tilde{x}))}{\partial u_1} = h_1(\theta + \mu_1\theta^2/2), \tag{EC.57}$$

and from (EC.43) at $t = \tau$ that

$$\frac{\partial f_1(\tilde{x}, \hat{\phi}^1(\tilde{x}))}{\partial x_1} = e^{-\mu_1(\tau-\theta)}, \quad -\frac{\partial f_1(\tilde{x}, \hat{\phi}^1(\tilde{x}))}{\partial u_1} = \mu_1\theta e^{-\mu_1(\tau-\theta)}, \tag{EC.58}$$

where we have used parameter θ to unify the two cases: $\theta = \sigma_1 < \tau$ if $\hat{\phi}_1^1(\tilde{x}) > \hat{u}_1(\tilde{x})$, and $\theta = \tau$ if $\hat{\phi}_1^1(\tilde{x}) = \hat{u}_1(\tilde{x}) \wedge 1$. We can therefore write (EC.55) as

$$\frac{\partial V_N^1(\tilde{x})}{\partial x_1} = h_1\theta + \big(\beta - h_1(\theta + \mu_1\theta^2/2)\big)\big(1/(\mu_1\theta)\big). \tag{EC.59}$$

For class 2, we could be in Case 1, 2 or 4 but not 3 since $\tilde{x}_2 > \bar{\rho}_2$ (see the proof of Lemma EC.2). This in addition to the case where $\hat{\phi}^1(\tilde{x})$ is a corner solution leads to four cases that we consider separately below.

1. *$\hat{\phi}^1(\tilde{x})$ is a corner solution with $\hat{\phi}_1^1(\tilde{x}) = 1$ and $\hat{\phi}_2^1(\tilde{x}) = 0$:* Since $\hat{\phi}_2^1(\tilde{x}) = 0$ we are in Case 1 for class 2 and $\theta = \tau$ for class 1. Substituting the corresponding terms from (EC.44), (EC.45), and (EC.43) (at $t = \tau$) in (EC.56) and noting that $\hat{\phi}^1(\tilde{x})$ is a corner solution we have

$$\frac{\partial V_N^1(\tilde{x})}{\partial x_2} < h_2\tau + \big(\beta - h_2(\tau + \mu_2\tau^2/2)\big)\left(1/(\mu_2\tau)\right).$$

This together with (EC.59) at $\theta = \tau$ yields

$$-\frac{\partial g(x,u)}{\partial u_1} - \frac{\partial V_N^1(f(x,u))}{\partial x_1} \cdot \frac{\partial f_1(x,u)}{\partial u_1} = h_1(\tau + \mu_1\tau^2/2) + h_1\mu_1\tau^2 + \beta - h_1(\tau + \mu_1\tau^2/2) = h_1\mu_1\tau^2 + \beta,$$

$$-\frac{\partial g(x,u)}{\partial u_2} - \frac{\partial V_N^1(f(x,u))}{\partial x_2} \cdot \frac{\partial f_2(x,u)}{\partial u_2} < h_2(\tau + \mu_2\tau^2/2) + h_2\mu_2\tau^2 + \beta - h_2(\tau + \mu_1\tau^2/2) = h_2\mu_2\tau^2 + \beta,$$

which by assumption (ii) implies that (EC.50) is satisfied with a strict inequality.

2. *$\hat{\phi}^1(\tilde{x})$ is an interior solution and $\hat{\phi}^1_2(\tilde{x}) < \tilde{x}_2$ and $\hat{\phi}^1_2(\tilde{x}) \le \hat{u}_2$*: We are in Case 1 where class 2 starts with a positive queue at stage 2 that remains positive during the stage. Substituting the corresponding terms from (EC.44), (EC.45), and (EC.43) (at $t=\tau$) in (EC.56) we have

$$\frac{\partial V_N^1(\tilde{x})}{\partial x_2} = h_2\tau + \big(\beta - h_2(\tau+\mu_2\tau^2/2)\big)\,(1/(\mu_2\tau))\,.$$

This together with (EC.59) yields

$$-\frac{\partial g(x,u)}{\partial u_1} - \frac{\partial V_N^1(f(x,u))}{\partial x_1}\cdot\frac{\partial f_1(x,u)}{\partial u_1} = h_1(\tau+\mu_1\tau^2/2) + h_1\mu_1\tau\theta + (\tau/\theta)\big(\beta - h_1(\theta+\mu_1\theta^2/2)\big),$$
$$-\frac{\partial g(x,u)}{\partial u_2} - \frac{\partial V_N^1(f(x,u))}{\partial x_2}\cdot\frac{\partial f_2(x,u)}{\partial u_2} \le h_2(\tau+\mu_2\tau^2/2) + h_2\mu_2\tau^2 + \beta - h_2(\tau+\mu_2\tau^2/2).$$

We need to consider the two cases for θ separately.

(a) We first show that when $\theta=\sigma_1<\tau$ (EC.50) holds with a strict inequality. After simplifying, this is equivalent to showing

$$\beta\big((\tau/\sigma_1)-1\big) + h_1\mu_1\tau(\tau+\sigma_1)/2 - h_2\mu_2\tau^2 > 0. \tag{EC.60}$$

From the optimality equation (EC.54) we know that $\beta \ge -\partial g(\tilde{x},\hat{\phi}^1(\tilde{x}))/\partial u_2 = h_2(\tau+\mu_2\tau^2/2)$ and hence the LHS of (EC.60) is no less than

$$\xi(\sigma_1) \equiv h_2(\tau+\mu_2\tau^2/2)\big((\tau/\sigma_1)-1\big) + h_1\mu_1\tau(\tau+\sigma_1)/2 - h_2\mu_2\tau^2. \tag{EC.61}$$

We show that $\xi(\sigma_1)>0$ for any $\sigma_1\in(0,\tau)$ and hence the claim follows. To this end we need to consider two cases for σ_1. First, assume that $\sigma_1 \ge \tau(h_2\mu_2)/(h_1\mu_1)$. Then it is easy to see that $\xi(\sigma_1) > h_1\mu_1\tau(\tau+\sigma_1)/2 - h_2\mu_2\tau^2 \ge 0$. Next assume that $\sigma_1 < \tau(h_2\mu_2)/(h_1\mu_1)$. Combined with assumption (ii), we have $(\tau/\sigma_1)^2 > (h_1\mu_1)/(h_2\mu_2)$. This implies that $\xi(\sigma_1)$ is strictly decreasing on $\sigma_1 \in \big(0,\tau(h_2\mu_2)/(h_1\mu_1)\big)$ since $dh(\sigma_1)/d\sigma_1 = h_1\mu_1\tau/2 - h_2(\tau/\sigma_1)^2 - h_2\mu_2(\tau/\sigma_1)^2(\tau/2) < 0$. The claim then follows noting that $\xi(\sigma_1)$ is continuous and as we saw in the first case positive at $\sigma_1 = \tau(h_2\mu_2)/(h_1\mu_1)$.

(b) Next consider the case with $\theta=\tau$. We have

$$-\frac{\partial g(x,u)}{\partial u_1} - \frac{\partial V_N^1(f(x,u))}{\partial x_1}\cdot\frac{\partial f_1(x,u)}{\partial u_1} = h_1(\tau+\mu_1\tau^2/2) + h_1\mu_1\tau^2 + \beta - h_1(\tau+\mu_1\tau^2/2) = h_1\mu_1\tau^2+\beta,$$
$$-\frac{\partial g(x,u)}{\partial u_2} - \frac{\partial V_N^1(f(x,u))}{\partial x_2}\cdot\frac{\partial f_2(x,u)}{\partial u_2} \le h_2(\tau+\mu_2\tau^2/2) + h_2\mu_2\tau^2 + \beta - h_2(\tau+\mu_2\tau^2/2) = h_2\mu_2\tau^2+\beta.$$

Therefore, if assumption (ii) holds with an inequality, i.e., $h_1\mu_1 > h_2\mu_2$, (EC.50) also holds with an inequality and we are done. However, if $h_1\mu_1 = h_2\mu_2$ and also (EC.51) and (EC.52) are both satisfied with equality (i.e., we are in Case 1 for class 2 in the first stage) then (EC.50) holds with an equality and hence the allocation u belongs to some optimal policy, say π. In this case, however, we can construct another optimal policy that uses an allocation from $\widehat{\mathcal{U}}(x)$ in the first stage as follows: in stage 1, instead of u we use $(\hat{u}_1(x), 1-\hat{u}_1(x)) \in \widehat{\mathcal{U}}(x)$. In stage 2 we return the extra capacity allocated to class 1, i.e., $\hat{u}_1(x) - u_1$, to class 2 and use

$$\big(\hat{u}_1(\tilde{x}) - (\hat{u}_1(x) - u_1), 1 - \hat{u}_1(\tilde{x}) + (\hat{u}_1(x) - u_1)\big).$$

Finally, we use the same allocations in all future periods as the policy π. It is then easy to verify that we incur the same cost in the first two stages under the two policies and the trajectories *couple* at the end of the second stage. Therefore, both policies incur the same total minimal cost.

3. $\hat{\phi}^1(\tilde{x})$ *is an interior solution and* $\hat{u}_2 < \hat{\phi}^1_2(\tilde{x}) < \tilde{x}_2$: We are in Case 2 where class 2 starts with a positive queue at stage 2 that empties at some time $\sigma_2 < \tau$ into the stage. Since by the induction hypothesis queue 1 also empties by the end of the stage, using the same allocation $\hat{\phi}^1(\tilde{x})$ in future periods will keep both queues empty and hence future cost at zero. Therefore, the value function at the beginning of the second stage is simply,

$$V^1_N(\tilde{x}) = g(\tilde{x}, \hat{\phi}^1(\tilde{x})),$$

yielding

$$\frac{\partial V^1_N(\tilde{x})}{\partial x_1} = h_1\theta, \qquad \frac{\partial V^1_N(\tilde{x})}{\partial x_2} = h_2\sigma_2,$$

and thus

$$-\frac{\partial g(x,u)}{\partial u_1} - \frac{\partial V^1_N(f(x,u))}{\partial x_1} \cdot \frac{\partial f_1(x,u)}{\partial u_1} = h_1(\tau + \mu_1\tau^2/2) + h_1\mu_1\tau\theta, \tag{EC.62}$$

$$-\frac{\partial g(x,u)}{\partial u_2} - \frac{\partial V^1_N(f(x,u))}{\partial x_2} \cdot \frac{\partial f_2(x,u)}{\partial u_2} \leq h_2(\tau + \mu_2\tau^2/2) + h_2\mu_2\tau\sigma_2. \tag{EC.63}$$

It follows by assumptions (*ii*) and (*iii*) that if $\theta = \tau$ (EC.50) holds with a strict inequality. In the following we show that when $\theta = \sigma_1 < \tau$ either the inequality is again strict, or there exists another optimal policy that uses an allocation in the first stage that belongs to $\widehat{\mathcal{U}}(x)$. Assume that $\theta = \sigma_1 < \tau$ and note that the optimality equations (EC.53)–(EC.54) for $\hat{\phi}^1(\tilde{x})$ reduce to

$$-h_1(\sigma_1 + \mu_1\sigma_1^2/2) = -h_2(\sigma_2 + \mu_2\sigma_2^2/2) = \beta, \tag{EC.64}$$

where as before we have suppressed the dependence of σ_i on $\tilde{x}$ and $\hat{\phi}^1(\tilde{x})$. We want to show that (EC.50) holds when $\theta = \sigma_1$, i.e., using (EC.62) and (EC.63),

$$h_1(\tau + \mu_1\tau^2/2) + h_1\mu_1\tau\sigma_1 \geq h_2(\tau + \mu_2\tau^2/2) + h_2\mu_2\tau\sigma_2. \tag{EC.65}$$

To simplify notation, we define $z \equiv h_2/h_1$. Note that assumptions (*ii*) and (*iii*) can be written as $\mu_1 \geq z\mu_2$ and $(\tau + \mu_1\tau^2/2) \geq z(\tau + \mu_2\tau^2/2)$, respectively, and the optimality equation (EC.64) is $\sigma_1 + \mu_1\sigma_1^2/2 = z(\sigma_2 + \mu_2\sigma_2^2/2)$ which can be written as

$$\mu_1\sigma_1 = -1 + \sqrt{1 + 2\mu_1 z\sigma_2 + \mu_1\mu_2 z\sigma_2^2}. \tag{EC.66}$$

Finally, re-writing (EC.65), we want to show

$$\tau+\mu_1\tau^2/2+\mu_1\tau\sigma_1 \geq z(\tau+\mu_2\tau^2/2+\mu_2\tau\sigma_2). \tag{EC.67}$$

First assume that $z<1$. Then we can use (EC.66) to write

$$\begin{aligned}\tau+\mu_1\tau^2/2+\mu_1\tau\sigma_1 &= \mu_1\tau^2/2+\tau\sqrt{1+2\mu_1 z\sigma_2+\mu_1\mu_2 z\sigma_2^2}\\ &\geq z\mu_2\tau^2/2+\tau\sqrt{1+2z^2\mu_2\sigma_2+z^2\mu_2^2\sigma_2^2}\\ &> z\mu_2\tau^2/2+\tau\sqrt{z^2(1+2\mu_2\sigma_2+\mu_2^2\sigma_2^2)}\\ &= z\mu_2\tau^2/2+\tau z(1+\mu_2\sigma_2)\\ &= z(\tau+\mu_2\tau^2/2+\tau\sigma_2\mu_2),\end{aligned}$$

where the first inequality follows from the assumption $\mu_1\geq z\mu_2$ and the second from $z<1$. Next assume that $z\geq 1$ and note that this together with assumption (ii) $\mu_1\geq z\mu_2$ implies that $\mu_1\geq\mu_2$. Again using (EC.66) we have

$$\begin{aligned}\mu_1\sigma_1 &= -1+\sqrt{1+2\mu_1 z\sigma_2+\mu_1\mu_2 z\sigma_2^2}\\ &\geq -1+\sqrt{1+2\mu_2 z\sigma_2+\mu_2^2 z^2\sigma_2^2}\\ &= \mu_2 z\sigma_2,\end{aligned}$$

where the inequality follows using $\mu_1\geq\mu_2$ in the first term in the squared root and $\mu_1\geq z\mu_2$ in the second term. This, together with assumption (iii) $(\tau+\mu_1\tau^2/2)\geq z(\tau+\mu_2\tau^2/2)$ proves the claim. Now note that in the first case, i.e., $z<1$ the inequality is strict. The same holds in the second case as long as $z>1$ or $\mu_1>\mu_2$. Therefore, (EC.50) can only hold with equality if $h_1=h_2$ and $\mu_1=\mu_2$ and also both (EC.52) and (EC.51) hold with equality (i.e., we are in Case 1 for class 2 in the first stage and (EC.63) holds with equality). This further implies using (EC.66) that $\sigma_1=\sigma_2$. That is, under an optimal policy we start with positive queues for both classes that remain positive in the first stage and empty at the same time during the second stage. It is however easy to see that in this case by using $(\hat{u}_1(x),1-\hat{u}_1(x))\in\widehat{\mathcal{U}}(x)$ in stage 1 and emptying the queues at the same time in the second stage, the optimality equations are still satisfied and hence we are done.

4. *$\hat{\phi}^1(\tilde{x})$ is an interior solution and $\hat{\phi}^1_2(\tilde{x})\geq\tilde{x}_2$ and $\hat{\phi}^1_2(\tilde{x})\geq\check{u}_2$:* We are in Case 4 where class 2 starts with an empty queue in stage 2 that remains empty during the stage. In this case again since both queues empty at the second stage we can keep the cost zero in future periods and hence,

$$V_N^1(\tilde{x})=g(\tilde{x},\hat{\phi}^1(\tilde{x})).$$

It is then easy to see that $\hat{\phi}^1(\tilde{x})$ cannot be optimal since the optimality equation $-h_1(\theta+\mu_1\theta^2/2)=0$ cannot have a solution for $\theta>0$. The proof is complete. □

EC.3.4. Proof of Theorem 4: Asymptotic Optimality of the DR Policy

We need the following basic lemma before presenting the proof.

LEMMA EC.4. *For any sequence of admissible control policies $\{U^n\}$ the corresponding sequence of scaled total costs $\{n^{-1}C_T^n(U^n)\}$ is uniformly integrable and*

$$\limsup_{n\to\infty} \mathbb{E}\left[n^{-1}C_T^n(U^n)\right] \le \mathbb{E}\left[\limsup_{n\to\infty} n^{-1}C_T^n(U^n)\right]. \tag{EC.68}$$

Proof. First, a simple bound for the sequence of scaled total costs can be obtained by bounding the queue length by the total arrivals similar to that in Lemma 4.5. of Dai (1995). We have,

$$\begin{aligned} n^{-1}C_T^n(U^n) = \int_0^T h\cdot \bar{Q}^n(s)ds &\le \int_0^T \sum_{i\in\mathcal{I}} h_i(\bar{X}_i^n(0) + n^{-1}A_i(\lambda_i^n s))ds \\ &\le \int_0^T \sum_{i\in\mathcal{I}} h_i(M + n^{-1}A_i(\lambda_i^n T))ds \\ &\le T\bar{h}(M + \sum_{i\in\mathcal{I}} n^{-1}A_i(\lambda_i^n T)), \end{aligned}$$

where $\bar{h} = \max_{i\in\mathcal{I}} h_i$. It is easy to verify that under Assumption 2 the sequence of scaled Poisson random variables $\{n^{-1}A_i(\lambda_i^n T)\}$ is uniformly integrable for each $i \in \mathcal{I}$. It follows that $\{n^{-1}C_T^n(U^n)\}$ is also uniformly integrable. Next, we show (EC.68) by proving that it holds for any uniformly integrable sequence $\{B^n\}$. To this end observe that,

$$\begin{aligned} \limsup_{n\to\infty} \mathbb{E}[B^n] &\le \limsup_{n\to\infty} \mathbb{E}[B^n \wedge b] + \sup_n \mathbb{E}[B^n \mathbf{1}\{B^n \ge b\}] \\ \text{(Fatou's Lemma)} &\le \mathbb{E}\left[\limsup_{n\to\infty} B^n \wedge b\right] + \sup_n \mathbb{E}[B^n \mathbf{1}\{B^n \ge b\}] \\ &\le \mathbb{E}\left[\limsup_{n\to\infty} B^n\right] + \sup_n \mathbb{E}[B^n \mathbf{1}\{B^n \ge b\}]. \end{aligned}$$

Taking $b \to \infty$ the second term on the RHS vanishes by uniform integrability of $\{B^n\}$ proving the claim. This completes the proof. □

Proof of Theorem 4. Consider a sample path ω and denote by $\{n_j; j \in \mathbb{N}\}$ the subsequence under which the lim sup is achieved for $n^{-1}C_T^n(U_*^n)$ (i.e., under the DR policy). Note that since by Lemma EC.4 $\{n^{-1}C_T^n(U_*^n)\}$ is uniformly integrable the lim sup is finite. By Proposition EC.1 for almost all sample paths there exists a further subsequence, say $\{n_{j'}; j' \in \mathbb{N}\}$, such that the sequence of corresponding scaled processes $\{(\bar{X}_*^{n_{j'}}, \bar{T}_*^{n_{j'}}, \bar{Y}_*^{n_{j'}}, \bar{\Upsilon}_*^{n_{j'}})\}$ converges to some $(\bar{X}_*, \bar{T}_*, \bar{Y}_*, \bar{\Upsilon}_*)$ satisfying (EC.11)–(EC.15) a.e. It follows that a.s.,

$$\begin{aligned} \limsup_{n\to\infty} n^{-1}C_T^n(U_*^n) &= \lim_{j'\to\infty} n_{j'}^{-1}C_T^{n_{j'}}(U_*^{n_{j'}}) \\ &= h\cdot\bar{\Upsilon}_*(T) = \int_0^T h\cdot\bar{Q}_*(s)ds = \int_0^T h\cdot\left(\bar{X}_*(s) - \bar{U}_*(s)\right)^+ ds. \end{aligned} \tag{EC.69}$$

Next, note using (EC.11) and (EC.13) that $\bar{X}_*$ and $\bar{U}_*$ satisfy

$$\dot{\bar{X}}_*(t) = \bar{\lambda} - \hat{\mu} \cdot (\bar{X}_*(t) \wedge \bar{U}_*(t)). \tag{EC.70}$$

Further, since $U_*^n \in \Pi_\tau^n$ for all n, (EC.16) and (EC.17) hold, that is $\bar{U}_*$ is piecewise-constant and

$$\bar{U}_*(t) = \sum_{k=0}^{\bar{k}(T)-1} \bar{U}_*[k] \mathbf{1}_{[t_k, t_{k+1})}(t), \tag{EC.71}$$

for some $\{\bar{U}_*[k]; 0 \le k \le \bar{k}(T)-1\}$ with $t_{k+1} - t_k = \tau$. This, together with (EC.14) and (EC.15), implies that $(\bar{X}_*, \bar{U}_*)$ is an admissible pair for Problem 2 starting from $x^0 = \bar{X}(0,\omega)$. In the following, we first show that $\{\bar{U}_*[k]; 0 \le k \le N-1\}$ is an optimal policy for Problem 2 given the starting point $X(0,\omega)$ and hence the total cost in (EC.69) is the optimal cost starting from $X(0,\omega)$, i.e., $J_{T,\tau}(X(0,\omega))$, for almost all ω. We then take expectation and apply Lemma EC.4 to get the final result in (41).

The first part of the proof is by (forward) induction on k. First, consider $k = 0$ and note that by Assumption 1 the initial condition converges to the random variable $\bar{X}(0)$ a.s., so we have $\bar{X}_*^n(0,\omega) \to \bar{X}(0,\omega)$ for almost all ω. Now substituting from (39) in (40) we can write the scaled allocation at time $k = 0$ as

$$\begin{aligned} \bar{U}_*^{n_{j'}}[0] \equiv n_{j'}^{-1} U_*^{n_{j'}}[0] &= n_{j'}^{-1} \left\lfloor n_{j'} \tilde{U}_*^{n_{j'}}[0] \right\rfloor \\ &= n_{j'}^{-1} \left\lfloor n_{j'} \hat{\phi}^0 \left((\bar{X}_*^{n_{j'}}(0) - \bar{\beta}^{n_{j'}})^+ \right) \right\rfloor \\ &= \hat{\phi}^0 \left((\bar{X}_*^{n_{j'}}(0) - \bar{\beta}^{n_{j'}})^+ \right) + n_{j'}^{-1} \varepsilon, \end{aligned} \tag{EC.72}$$

where the rounding error $\varepsilon \in \mathbb{R}_+^I$ satisfies $|\varepsilon_i| \le 1$ for all $i \in \mathcal{I}$. Next, taking the limit $j' \to \infty$ in (EC.72) we claim that

$$\bar{U}_*[0] = \lim_{j' \to \infty} \bar{U}_*^{n_{j'}}[0] = \hat{\hat{\phi}}^0\big(\bar{X}(0,\omega)\big) \in \phi^0(\bar{X}(0,\omega)), \tag{EC.73}$$

where $\hat{\hat{\phi}}^k(x)$ denotes an arbitrary member of the correspondence $\phi^k(x)$ which may or may not be the same as $\hat{\phi}^k(x)$. To establish the claim first consider (EC.72) and note that since the safety vector $\beta^n = o(n)$ is negligible under fluid scaling and the $(\cdot)^+$ function is continuous, we have $(\bar{X}_*^{n_{j'}}(0) - \bar{\beta}^{n_{j'}})^+ \to \bar{X}(0,\omega)$ as $j' \to \infty$. Indeed, if $\hat{\phi}^0(\cdot)$ were a continuous function then the claim would directly follow. However, the existence of a continuous "selection" from the optimal correspondence is not generally guaranteed for convex optimization problems (unlike, e.g., linear programs as shown and used in Bassamboo et al. 2006). Therefore, here we use a different argument based on the properties of the correspondence established in Proposition 1. Since $\phi^k(\cdot)$ is upper hemicontinuous and closed-valued, it also satisfies the closed-graph property (Ok 2007, Proposition 3). The closed-graph property of the correspondence $\phi^k(\cdot)$ implies that for any convergent sequences $\{x_n\}$ and $\{y_n\}$ with $x_n \to x$ and $y_n \to y$, we have $y \in \phi^k(x)$ whenever $y_n \in \phi^k(x_n)$ for all n (Ok 2007, Page 294). Since by definition, $\hat{\phi}^0\big((\bar{X}_*^{n_{j'}}(0) - \bar{\beta}^{n_{j'}})^+\big) \in \phi^0\big((\bar{X}_*^{n_{j'}}(0) - \bar{\beta}^{n_{j'}})^+\big)$ for all j', and $\{n_{j'}; j' \in \mathbb{N}\}$ is a convergent subsequence, the closed-graph property of $\phi^k(\cdot)$ establishes the claim in (EC.73).

Next, assume that the sequence $\{\bar{U}_*[\kappa]; 0 \le \kappa \le k\}$ is optimal up to time k. A similar argument to that for $k = 0$ yields

$$\bar{U}_*^{n_{j'}}[k+1] \equiv n_{j'}^{-1} U_*^{n_{j'}}[k+1] = n_{j'}^{-1}\left(n_{j'}\hat{\phi}^{k+1}\left((\bar{X}_*^{n_{j'}}(t_{k+1}) - \bar{\beta}^{n_{j'}})^+\right) + \varepsilon\right),$$

which implies, using the closed-graph property of $\phi^{k+1}(\cdot)$ that,

$$\bar{U}_*[k+1] = \lim_{j' \to \infty} \bar{U}_*^{n_{j'}}[k+1] = \hat{\hat{\phi}}^{k+1}\big(\bar{X}_*(t_{k+1})\big) \in \phi^{k+1}\big(\bar{X}_*(t_{k+1})\big).$$

For $\bar{U}_*[k+1]$ to be the optimal allocation at time $k+1$, we need $\bar{X}_*(t_{k+1})$ to be on the optimal trajectory. This however follows noting that by the induction assumption $\bar{X}_*(t)$ is indeed the optimal trajectory for $t \in [0, t_{k+1})$ and hence by the continuity of the trajectories (see Proposition EC.1) $\bar{X}_*(t_{k+1}) = \bar{X}_*(t_{k+1}-)$ is also on the optimal trajectory.

We have shown that for almost all sample paths ω, $\limsup_{n\to\infty} n^{-1} C_T^n(U_*^n) = J_{T,\tau}(X(0,\omega))$. Therefore taking expectation and applying Lemma EC.4 we have

$$\limsup_{n\to\infty} \mathbb{E}\left[n^{-1} C_T^n(U_*^n)\right] \le \mathbb{E}\left[\limsup_{n\to\infty} n^{-1} C_T^n(U_*^n)\right] = \mathbb{E}\left[J_{T,\tau}(X(0))\right].$$

The proof is complete. □

EC.4. Additional Numerical Experiments

EC.4.1. On the conjecture formed in Section 4.2.3

Here we elaborate more on the conjecture formed in Section 4.2.3 and present a numerical example. Our conjecture implies that regardless of the initial condition x^0, assuming that conditions (ii) and (iii) in Theorem 3 are met, there exists an optimal policy with class 1 allocations that satisfy the following in each stage k,

$$\hat{\phi}_1^k(x[k]) \ge \hat{u}_1(x[k], \tau) \wedge 1, \quad x_1[k] > \bar{\rho}_1, \tag{EC.74}$$

$$\hat{\phi}_1^k(x[k]) \ge x_1[k] \wedge 1, \quad x_1[k] \le \bar{\rho}_1. \tag{EC.75}$$

Note that in each case, the lower-bound is the maximum allocation such that no idleness is incurred in class 1. The lower-bound in (EC.74) is the same as in Theorem 3 stated under the assumption $x[k] > \bar{\rho}$. (Recall that by Lemma EC.2, $x^0 > \bar{\rho}$ ensures $x[k] > \bar{\rho}$ for all $k \in \mathcal{K}$.) To understand the

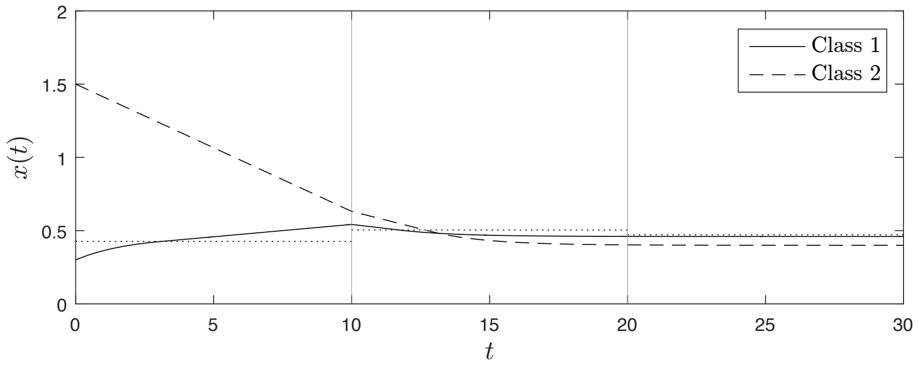

Figure EC.1 Optimal class 1 (solid) and class 2 (dashed) trajectories for Problem 2 with $\bar{\lambda} = (0.23, 0.20)$, $\mu = (0.5, 0.5)$, $x^0 = (0.3, 1.5)$, $h = (4, 2)$, and $\tau = 10$. The dotted horizontal lines correspond to optimal class 1 allocations for each stage.

lower-bound in (EC.75) note that when $x_1[k] \le \bar{\rho}$, any allocation smaller than $x_1[k]$ results in a positive queue at the beginning of the stage that builds up linearly during the stage and hence incurs no idleness. We demonstrate the above using an additional example. In Figure EC.1, we plot the optimal trajectories for the same parameters as in Figure 3, but with initial condition $x^0 = (0.3, 1.5)$ which is below the equilibrium point, i.e., $x_1 < \bar{\rho}_1$. We observe that the optimal allocation in stage 1 satisfies $x_1 \le u_1 < \bar{\rho}_1$. That is, although less capacity is allocated to class 1 in order to drain the large class 2 queue, the allocation is such that some idleness is incurred in class 1 at the beginning of the first stage. Note that in the second stage, where the conditions of the theorem are met, the allocation to class 1 is such that the queue is emptied during the stage.

EC.4.2. On the relationship between conditions of Theorem 3

Here we numerically demonstrate that condition (ii) $(h_1\mu_1 \ge h_2\mu_2)$ is sufficient for the statement of Theorem 3 to hold, as long the horizon is large enough for the queues to drain.

We consider a system with parameters $(h_1, h_2) = (2, 6)$, $(\mu_1, \mu_2) = (2, 0.5)$ and with $\tau = 4$. Note that $h_1\mu_1 - h_2\mu_2 = 1$ and $h_2 - h_1 = 4$, therefore condition (ii) is satisfied but condition (iii) does not hold since $(\tau/2)(h_1\mu_1 - h_2\mu_2) < h_2 - h_1$. Solving the problem with initial condition $x^0 = (1.6, 0.9)$ and with one stage $(N = 1)$, the optimal allocation for class 1 is $\hat{\phi}_1^0(x^0) = 0.43 < \hat{u}_1 = 0.587$ (minimum allocation required to empty the queue) which does not empty the queue during the stage. However, when considering 2 (or more) stages the optimal allocation empties the class 1 queue in stage 1. In Table EC.1 we report the optimal allocation in stage 1 when considering $N = 1, 2, 3,$ stages in the horizon. We also compare the total optimal cost with that of the policy which empties queue 1 exactly at the end of each stage. Observe that the optimality gap quickly goes down as the number of stages increases. Numerical experiments suggest that the observations hold regardless of the parameters and initial conditions.

Number of stages (N)	1	2	3
Optimal allocation in stage 1	0.419	0.589	0.589
Optimal total cost	14.133	20.922	21.492
Total cost when emptying at the end of each stage	15.413	21.179	21.528
Optimality gap	9.06%	1.23%	0.17%

Table EC.1 Comparing optimal allocations and total cost with those of the policy that empties the queue at the end of each stage for an example with $\bar{\lambda} = (0.92,\ 0.20)$, $\mu = (2, 0.5)$, $x^0 = (1.6, 0.9)$, $h = (2, 6)$, and $\tau = 4$.

EC.4.3. Impact of Customer Abandonment

Here we study the impact of customer abandonment on the value of partial flexibility and demonstrate the performance of the heuristics proposed in Section 6.3. To isolate the impact of abandonment, we assume a stationary arrival rate in the experiments. We evaluate the performance of the heuristics for a two-class system with respect to both expected transient and steady-state cost, and for different system parameters. In the presence of abandonment, we find that considering an extended range of safety parameters can improve the performance of the policies. Hence, in our experiments we vary the safety factor for class 1 in $a \in \{-2, -1, 0, 1, 2\}$ and set the safety factor for class 2 to 0 as before. To find the optimal dedicated staffing, we use the Erlang-A formula to exactly find the fixed assignment of servers that minimizes the total expected cost of the system. We use system parameters $\mu = (0.5, 0.5)$, $r = (1/2, 1/2)$, $\rho = 0.96$, and vary the shift-length and system size. We assume the same holding cost $h = (4, 2)$ vector and set the abandonment cost to be 6 times the holding cost, i.e., $\gamma = (24, 12)$ in the experiments, unless otherwise stated.

Steady-state experiments. We illustrate the observations summarized in the paper using two sets of representative examples. In Figure EC.2 we plot the percentage of cost reduction compared to dedicated staffing for increasing abandonment rates (starting with $\theta_1 = \theta_2 = 1/256$ and increasing the rates by a factor of two each time) and two system sizes $n \in \{20, 48\}$ with shift length $\tau = 4$. The cost reductions are plotted as a function of the abandonment probability under the optimal dedicated staffing. In addition to the heuristics which explicitly take abandonment into account, we also plot the performance of the DR policy (which ignores abandonment and the associated cost parameters). As we expect, with higher abandonment rates, the benefits of partial flexibility decrease since the system reaches steady-state early into the shifts. However, with small abandonment there could still be significant benefits in steady-state. The CMUT policy achieves a comparable performance to that of GDR while the cost reduction under the BH heuristic is smaller as it ignores the cost parameters. We also observe that when the abandonment rate is relatively small, the performance of the DR policy is very similar to the GDR. As the abandonment rate increases, GDR outperforms the DR policy, but this typically happens in the region where the benefits are relatively small. The same observation holds when comparing the heuristics with their counterparts which ignore abandonment.

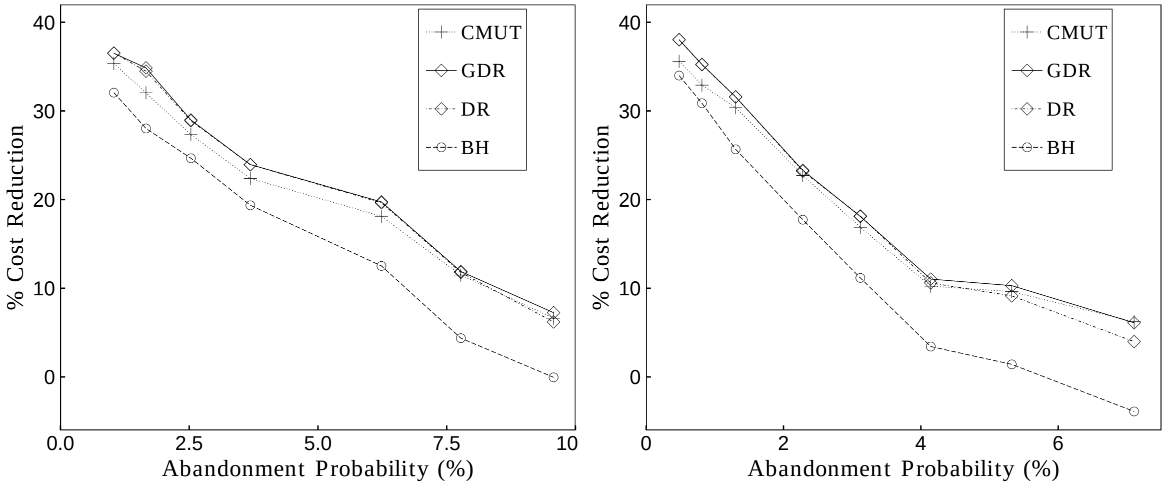

Figure EC.2 Performance of the heuristics in terms of estimated expected long-run average cost for increasing abandonment rates and with cost parameters $h = (4,2)$ and $\gamma = (24,12)$, and system parameters $\mu = (0.5, 0.5)$, $r = (1/2, 1/2)$, $\rho = 0.96$, $\tau = 4$; $n = 20$ (left) and $n = 48$ (right).

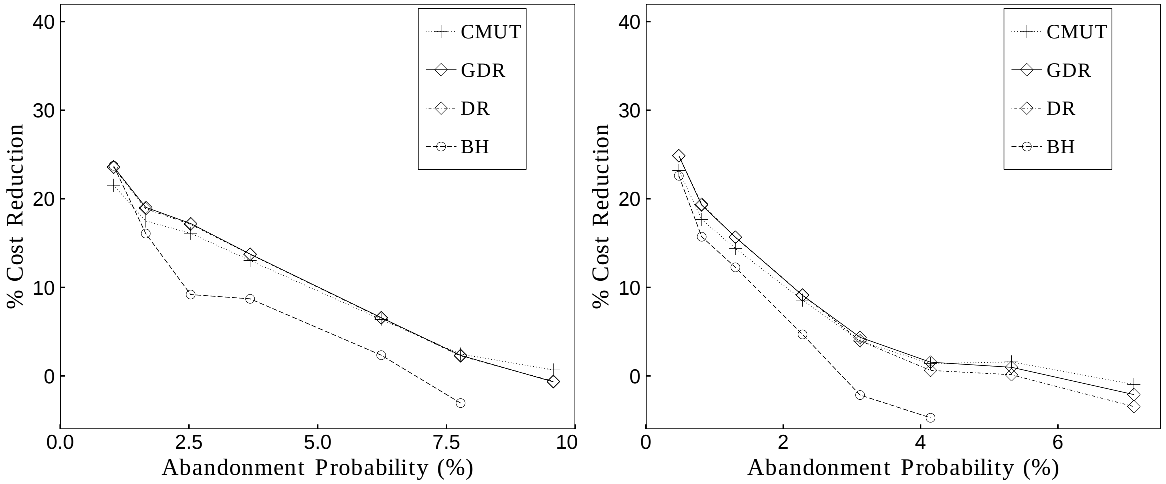

Figure EC.3 Performance of the heuristics in terms of estimated expected long-run average cost for increasing abandonment rates and with cost parameters $h = (4,2)$ and $\gamma = (24,12)$; system parameters $\mu = (0.5, 0.5)$, $r = (1/2, 1/2)$, $\rho = 0.96$, $\tau = 12$; $n = 20$ (left) and $n = 48$ (right).

The effect of other system parameters on the benefits of partial flexibility are consistent with those in the absence of abandonment. In Figure EC.3 we plot the results of the same experiments for shift length $\tau = 12$. The cost reductions are decreasing in the shift length, but larger for the smaller system similar to Figure EC.2.

While the cost parameters directly affect the GDR assignments, they only determine the priority ranking among different classes for the CMUT and CMUθT policies. In the above experiments, the $c\mu$ and $c\mu/\theta$ indexes are aligned and lead to the same ranking, and hence the two heuristics are identical. When the indexes are not aligned, the performance of the heuristics could significantly

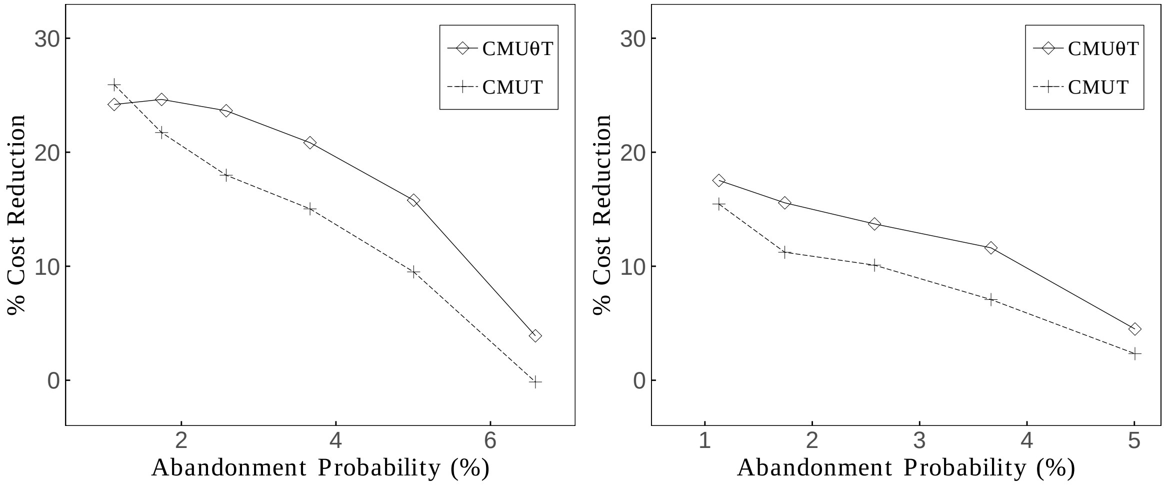

Figure EC.4 Performance of the CMUT and CMUθT policies in terms of estimated expected long-run average cost for increasing abandonment rates and with cost parameters $h=(4,2)$ and $\gamma=(24,12)$; system parameters $\mu=(0.5,0.5)$, $r=(1/2,1/2)$, $\rho=0.96$, $n=20$; $\tau=4$ (left) and $\tau=12$ (right).

differ. In Figure EC.4, we plot the performance of the two heuristics for a system with $n=20$ and the same cost parameters, but abandonment rates that lead to the opposite priority ranking. More specifically, we set $\theta_1 = 10 \times \theta_2$ and increase θ_2 by a factor of two starting from $\theta_2 = 1/960$. In the left plot (with $\tau=4$) when the abandonment probability is small, the CMUT policy performs better, whereas for high abandonment the CMUθT achieves a higher cost reduction. In the right plot (with $\tau=12$) the CMUθT policy achieves a better performance for all abandonment rates. We also find that while using the "right" priority ranking could have a significant effect, ignoring the abandonment in the trajectories is less important. This can be explained noting that with relatively small abandonment (where there are still benefits) the abandonment rate does not significantly affect the dynamics of the system during a shift and hence its impact on the server assignment (after rounding) is typically zero.

Transient experiments. We next demonstrate the performance of the heuristics with respect to their expected cost over a finite horizon. We consider a two-class system with $n=48$ servers and even load $r_1=r_2=0.5$. The cost parameters are $h=(4,2)$ and $\gamma=(24,12)$ (unless otherwise stated) and the shift-length is $\tau=12$. The initial conditions are randomly generated in 4 different regions with "high" (H) and "low" (L) initial congestion in each class. Similar to the experiments for the base model, "low" congestion refers to starting with a headcount in $\{0,\ldots,0.5\times n\}$ and a "high" congestion to a headcount in $\{n,\ldots,1.5\times n\}$. We test the performance of the policies for varying length of the horizon $T\in\{\tau,2\tau,3\tau\}$ and abandonment rates $\theta_i\in\{0.033,0.067,1\}$, for $i=1,2$, and in comparison with the optimal dedicated staffing (with respect to the steady-state cost).

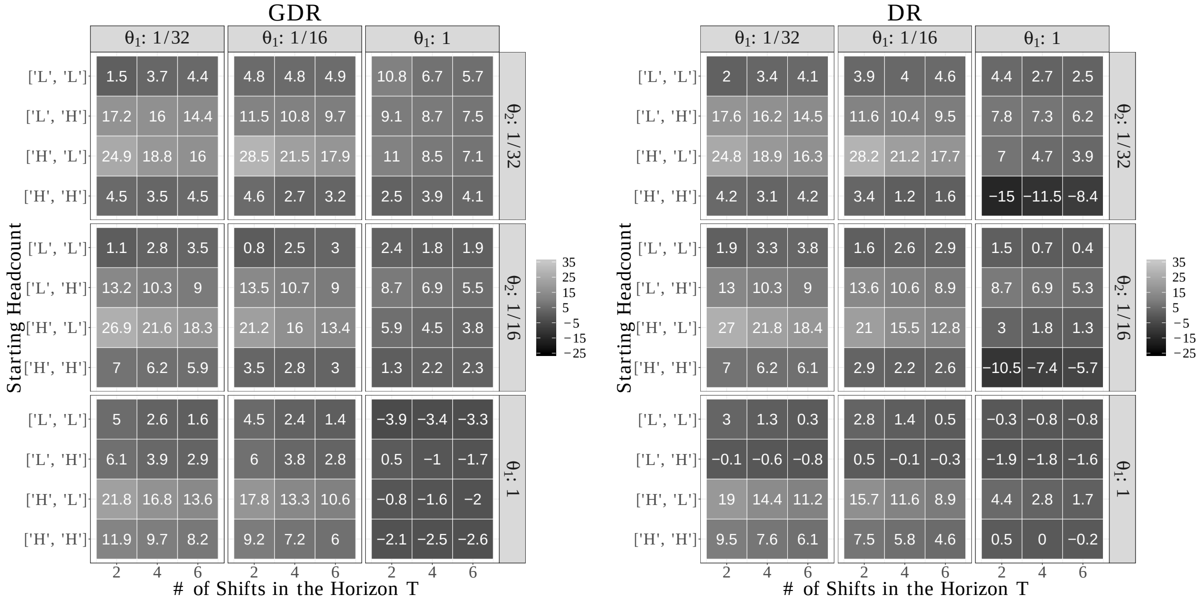

Figure EC.5 Average transient cost reduction over different initial conditions compared to dedicated staffing under the GDR heuristic (left) and the DR policy (right) for varying horizon lengths and abandonment rates. Cost parameters are $h = (4,2)$ and $\gamma = (24,12)$ and system parameters are $\mu = (0.5,0.5)$, $r = (1/2,1/2)$, $\rho = 0.96$, and $\tau = 12$.

Figure EC.5 presents the performance of the GDR heuristic and the DR policy. The reported values are average estimated cost reduction percentages across the 30 initial conditions considered for each region. We observe that the GDR heuristic can achieve a significant cost reduction over dedicated staffing, especially when the initial conditions are in different regions. The benefits are considerable even for abandonment rates for which steady-state benefits are close to zero. In addition, the GDR policy achieves a higher cost reduction compared to the DR policy (which ignores abandonment) especially when the abandonment rates are high and far from one another.

Figure EC.6 presents the same plots for CMUT (which in the examples coincides with CMUθT) and BH heuristics. The CMUT heuristic achieves comparable cost reductions to GDR. The BH heuristic does not perform well except when the abandonment rate is not too high and the initial conditions are significantly imbalanced. This can be explained noting that BH does not take cost parameters into account and only tries to balance the queue lengths, which does not necessarily lead to good performance with respect to the expected incurred cost. In Figure EC.7 we present the performance of the BH with respect to both cost and queue length reduction for the same examples, but with identical holding $h = (1,1)$ and abandonment cost $\gamma = (12,12)$ rates. We observe that the BH achieves significant cost reduction except when the abandonment rates are high. In all cases a positive queue length reduction is observed. We conclude that the BH should be used when the customer classes are sufficiently close with respect to their holding and abandonment costs,

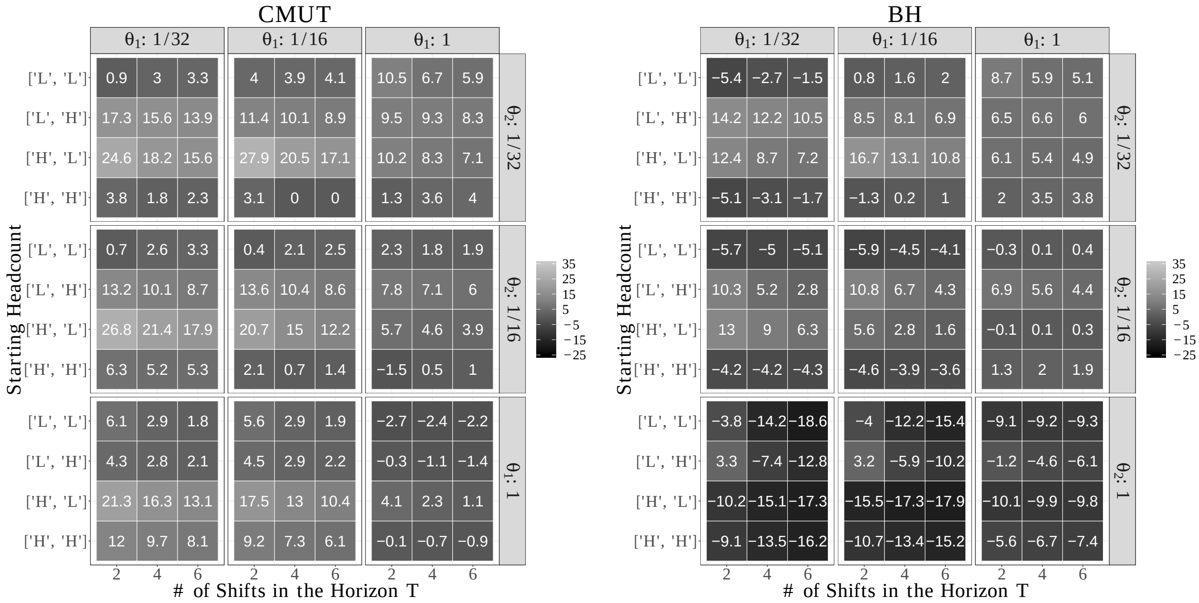

Figure EC.6 Average transient cost reduction over different initial conditions compared to dedicated staffing under the CMUT heuristic (left) and the BH heuristic (right) for varying horizon lengths and abandonment rates. Cost parameters are $h = (4, 2)$ and $\gamma = (24, 12)$ and system parameters are $\mu = (0.5, 0.5)$, $r = (1/2, 1/2)$, $\rho = 0.96$, and $\tau = 12$.

and achieving a balanced queue length is a desirable outcome. Otherwise, CMUT (or CMUθT) heuristics should be used.

EC.4.4. Experiments with Time-Varying Arrivals

Focusing on the long-run average cost and queue length, here we evaluate the performance of the heuristic policies proposed in Section 6.1 using simulation experiments and under a variety of system parameters. Similar to the stationary case we compare the performance to that under dedicated staffing and the fully flexible $c\mu$ policy. In all experiments we consider the sinusoidal arrival rate function $\lambda_i(t) = \lambda_i - (\lambda_i/2)\sin(\pi t/12)$ for $i \in \mathcal{I}$ and vary the average rates λ_i. The arrival rate function has period 24. In all experiments, we assume that the first shift starts at 7 a.m. We note that the sinusoidal arrival rates only allow for closed-form computation of the trajectories depending on whether the queue length is positive or at zero, and hence require numerical computation of the time-instances during the shift when the queue empties or starts building up.

Figure EC.8 illustrates the result for a 4-class system with service rate $\mu_i = 0.5, i \in \mathcal{I}$ and equal offered load for all classes, i.e., $r_i = 1/4, i \in \mathcal{I}$. The left plot corresponds to total utilization $\rho = 0.92$ and the right plot to $\rho = 0.96$ (computed using the average arrival rates during a 24-hour period). Each plot illustrates the percentage of cost reduction with respect to the cost under optimal dedicated staffing for different system sizes $n \in \{32, 48, 80\}$. In addition to the CMUT and BH

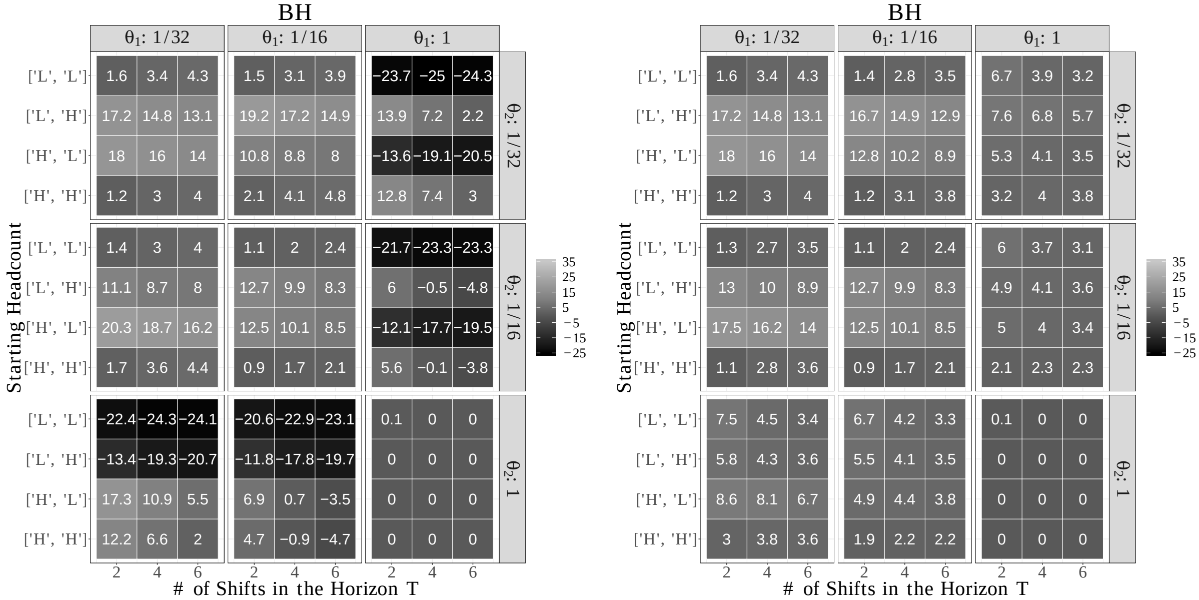

Figure EC.7 Average transient cost reduction (left) and total queue-length reduction (right) over different initial conditions compared to dedicated staffing under the BH heuristic (right) for varying horizon lengths and abandonment rates. Cost parameters are $h=(1,1)$ and $\gamma=(12,12)$ and system parameters are $\mu=(0.5,0.5)$, $r=(1/2,1/2)$, $\rho=0.96$, and $\tau=12$.

policies we also plot the performance under the GDR policy obtained by solving Problem 2 with time-varying dynamics. Due to the increased computational burden introduced by the time-varying dynamics, we solve the problem with $N=4$ at the beginning of each stage. We observe that, similar to the stationary case, the cost reduction under the partially flexible design is significant when the queues do not reach steady-state early into the shifts.

To further demonstrate the performance of the heuristics, in Tables EC.2 and EC.3 we report the detailed output of our simulation experiments for two examples. Table EC.2 corresponds to a *balanced* system with equal offered load for all classes, i.e., $r_i=1/4, i\in\mathcal{I}$, and $\mu_i=0.5, i\in\mathcal{I}$, with $h=(4,3,2,1)$, $\tau=12$, $\rho=0.92$. Table EC.3 corresponds to a system with $r_1=r_2=1/8$, $r_3=r_4=3/8$, and $\mu_1=\mu_2=0.25$, $\mu_3=\mu_4=0.5$, with $h=(4,3,2,1)$, $\tau=12$, $\rho=0.92$. That is, the offered load for the lower-priority classes is higher but the service times are higher on average for the higher-priority classes. For both examples, we also report the percentage of expected total cost and queue length reduction over that of dedicated staffing for all policies. (Note that notation CMUT(a) indicates that the reported results are associated with the best performing safety factor a.) We observe that compared to the CMUT policy, the BH usually achieves a lower total cost reduction as it does not incorporate the cost parameters. However, it still achieves a significant average queue length reduction compared to dedicated staffing. Further, note that in the balanced case the expected queue length under the BH policy is equal for all classes. That is, cost reduction

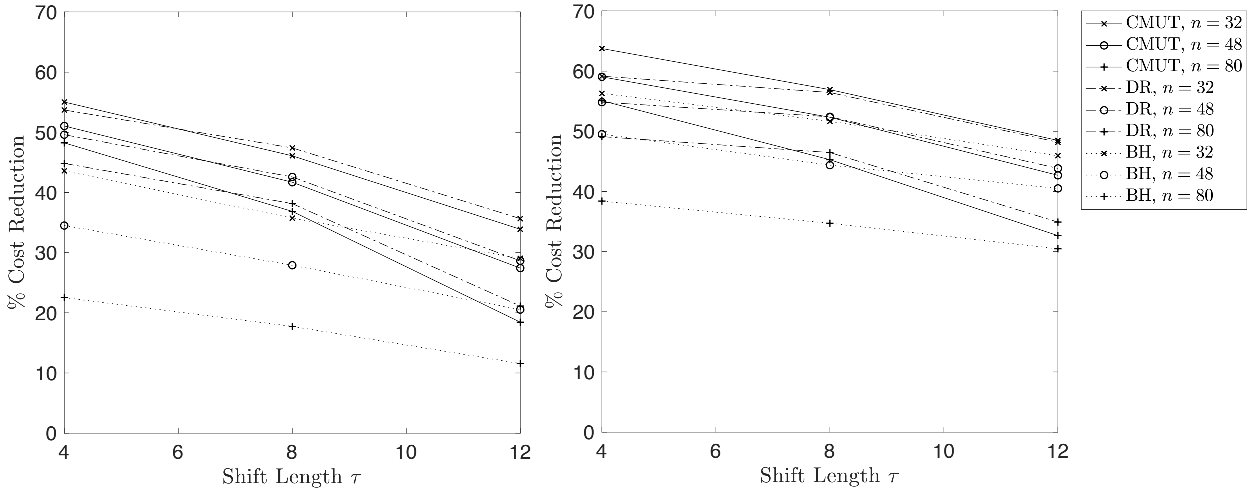

Figure EC.8 Performance of the heuristic policies in terms of estimated expected long-run average cost under time-varying arrivals and system parameters $\mu_i = 0.5$ and $r_i = 1/4$ for $i \in \{1,2,3,4\}$, $\tau = 12$, $\rho = 0.92$ (left) and $\rho = 0.96$ (right).

	$n=32$				$n=48$			
Perf. meas.	BH	CMUT(2)	$c\mu$	Ded.	BH	CMUT(2)	$c\mu$	Ded.
$\mathbb{E}[Q_1]$	$8.24 \pm .06$	$7.06 \pm .04$	$0.24 \pm .00$	$11.53 \pm .19$	$10.07 \pm .08$	$8.51 \pm .07$	$0.24 \pm .00$	$12.70 \pm .15$
$\mathbb{E}[Q_2]$	$8.17 \pm .06$	$7.04 \pm .04$	$0.58 \pm .00$	$11.57 \pm .17$	$10.06 \pm .10$	$8.44 \pm .06$	$0.58 \pm .00$	$12.60 \pm .14$
$\mathbb{E}[Q_3]$	$8.11 \pm .07$	$7.06 \pm .03$	$3.00 \pm .02$	$11.45 \pm .17$	$10.03 \pm .10$	$8.49 \pm .06$	$3.57 \pm .02$	$12.64 \pm .17$
$\mathbb{E}[Q_4]$	$8.07 \pm .05$	$12.71 \pm .15$	$16.47 \pm .08$	$11.49 \pm .15$	$9.98 \pm .08$	$15.45 \pm .20$	$22.84 \pm .09$	$12.57 \pm .14$
$\sum_i \mathbb{E}[Q_i]$	$32.59 \pm .20$	$33.87 \pm .16$	$20.29 \pm .09$	$46.04 \pm .29$	$40.14 \pm .35$	$40.89 \pm .30$	$27.23 \pm .10$	$50.52 \pm .31$
$\sum_i h_i \mathbb{E}[Q_i]$	$81.75 \pm .51$	$76.19 \pm .25$	$25.19 \pm .11$	$115.23 \pm .86$	$100.52 \pm .87$	$91.79 \pm .55$	$32.67 \pm .12$	$126.48 \pm .86$
% queue red.	29%	26%	56%	-	21%	19%	46%	-
% cost red.	29%	34%	78%	-	21%	27%	74%	-

Table EC.2 Detailed performance of the heuristic policies in terms of their estimated expected long-run average improvement compared to optimal dedicated staffing for system parameters $r_i = 1/4$, $\mu_i = 0.5$, $i \in \{1,2,3,4\}$, $h = (4,3,2,1)$, $\tau = 12$, and $\rho = 0.92$. (The numbers after ± correspond to %95 confidence intervals.)

is achieved by uniformly reducing the queue length for all classes. Under CMUT, similar to the $c\mu$ policy, the queue length for the first 3 classes are significantly reduced, while the queue length for the lowest-priority class is higher compared to dedicated staffing.

References

Bauschke HH, Lucet Y, Phan HM (2016) On the convexity of piecewise-defined functions. *ESAIM: COCV* 22(3):728-742.

Benveniste LM, Scheinkman JA (1979) On the differentiability of the value function in dynamic models of economics. *Econometrica: Journal of the Econometric Society* 727-732.

Dai JG (1995) On positive harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *The Annals of Applied Probability* 5(1):49-77.

Dai J, Tezcan T (2011) State space collapse in many-server diffusion limits of parallel server systems. *Mathematics of Operations Research* 36(2):271-320.

	$n = 32$				$n = 48$			
Perf. meas.	BH	CMUT(1)	$c\mu$	Ded.	BH	CMUT(1)	$c\mu$	Ded.
$\mathbb{E}[Q_1]$	$2.65 \pm .04$	$2.95 \pm .03$	$0.24 \pm .00$	$10.05 \pm .30$	$2.90 \pm .05$	$3.25 \pm .03$	$0.24 \pm .00$	$2.19 \pm .03$
$\mathbb{E}[Q_2]$	$2.60 \pm .05$	$2.88 \pm .04$	$0.38 \pm .00$	$9.83 \pm .28$	$2.89 \pm .05$	$3.24 \pm .03$	$0.40 \pm .00$	$9.84 \pm .25$
$\mathbb{E}[Q_3]$	$12.68 \pm .21$	$8.04 \pm .06$	$0.58 \pm .00$	$12.76 \pm .16$	$15.98 \pm .11$	$10.38 \pm .06$	$0.58 \pm .00$	$14.74 \pm .15$
$\mathbb{E}[Q_4]$	$12.69 \pm .19$	$19.52 \pm .59$	$18.55 \pm .12$	$12.67 \pm .17$	$16.02 \pm .16$	$21.02 \pm .45$	$24.33 \pm .12$	$40.42 \pm .84$
$\sum_i \mathbb{E}[Q_i]$	$30.61 \pm .45$	$33.40 \pm .65$	$19.75 \pm .12$	$45.30 \pm .42$	$37.79 \pm .32$	$37.89 \pm .52$	$25.55 \pm .12$	$67.18 \pm .97$
$\sum_i h_i \mathbb{E}[Q_i]$	$56.43 \pm .83$	$56.04 \pm .75$	$21.82 \pm .13$	107.86 ± 1.42	$68.25 \pm .60$	$64.50 \pm .64$	$27.65 \pm .13$	108.15 ± 1.32
% queue red.	32%	26%	56%	-	37%	44%	62%	-
% cost red.	48%	48%	80%	-	44%	40%	74%	-

Table EC.3 Detailed performance of the heuristic policies in terms of their estimated expected long-run average improvement compared to optimal dedicated staffing for system parameters $r_1 = r_2 = 1/8$, $r_3 = r_4 = 3/8$, $\mu_1 = \mu_2 = 0.25$, $\mu_3 = \mu_4 = 0.5$, $h = (4, 3, 2, 1)$, $\tau = 12$, and $\rho = 0.92$. (The numbers after ± correspond to %95 confidence intervals.)

Dattorro J (2005) *Convex Optimization and Euclidean Distance Geometry* (Meboo Publishing).

Maglaras C (2000) Discrete-review policies for scheduling stochastic networks: Trajectory tracking and fluid-scale asymptotic optimality. *The Annals of Applied Probability* 10(3):897-929.

Ok EA (2007) *Real analysis with economic applications* (Princeton University Press).

Stokey N, Lucas R (1989) *Recursive methods in economic dynamics* (Harvard University Press).